İSTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF SCIENCE AND TECHNOLOGY

CHURN MODELING IN TELECOMMUNICATIONS SECTOR

M.Sc. Thesis by
Müge ÖZMEN

Department : Management Engineering
Programme: Management Engineering

JUNE 2006

İSTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF SCIENCE AND TECHNOLOGY

CHURN MODELING IN TELECOMMUNICATIONS SECTOR

M.Sc. Thesis by
Müge ÖZMEN
507971024

Date of submission : 2 May 2006

Date of defence examination: 13 June 2006

Supervisor (Chairman): Prof. Dr. Demet BAYRAKTAR

Members of the Examining Committee Prof.Dr. Sıtkı GÖZLÜ

Doç. Dr. Tufan Vehbi KOÇ

JUNE 2006

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

TELEKOM SEKTÖRÜNDE AYRILACAK MÜŞTERİLERİN TAHMİNİ

YÜKSEK LİSANS TEZİ
Müge ÖZMEN
507971024

Tezin Enstitüye Verildiği Tarih :	**2 Mayıs 2006**
Tezin Savunulduğu Tarih :	**13 Haziran 2006**

Tez Danışmanı :	**Prof. Dr. Demet BAYRAKTAR**
Diğer Jüri Üyeleri	**Prof. Dr. Demet BAYRAKTAR**
	Prof.Dr. Sıtkı GÖZLÜ
	Doç. Dr. Tufan Vehbi KOÇ

HAZİRAN 2006

ACKNOWLEDGEMENTS

I gratefully acknowledge my advisor Prof. Dr. Demet Bayraktar who has given the invaluable comments, support and guidance throughout this thesis.

I would like to thank Talin Avakyan, Işık Gürer, Özlem Alan, İnci Bölükbaşı who gave valuable feedback and support during different phases of my study.

Finally, I am grateful to my parents for their endless support and encouragement through out my whole education life.

05.2006

Müge Özmen

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **ii**
ABBREVIATIONS **v**
LIST OF TABLES **vi**
LIST OF FIGURES **vii**
LIST OF SYMBOLS **viii**
SUMMARY **ix**
ÖZET **x**

1. INTRODUCTION **1**
1.1 Telecommunications Sector 4

2. DATA MINING **8**
2.1 Data Mining Definiton 8
2.2 History of Data Mining 9
2.3 Data Mining Approach 11
2.4 Application Areas 13
2.5 Data Mining Techniques 15
2.5.1 Supervised (Predictive) Modeling Techniques 15
2.5.2 Unsupervised (Descriptive) Modeling Techniques 17

3. DATA MINING METHODOLOGY **18**
3.1 Problem Definition 19
3.2 Data Preprocessing 20
3.2.1 Data Acquisition 20
3.2.2 Data Cleaning 22
3.2.3 Data Transformation & Transposition 23
3.3 Modeling 24
3.3.1 Preparation of the Model Data Set 24
3.3.2 Model Building and Testing 26
3.3.3 Decision Trees 26
3.3.4 Logistic Regression 33
3.4 Measure of Performance 35
3.5 Scoring 38

4. CHURN MODELING **40**
4.1 Problem Definition 40
4.2 Data Preprocessing 45
4.3 Modeling 46
4.3.1 Preparation of the Churn Model Data Set 48
4.3.2 Prepaid Churn Modeling 56
4.4 Measure of Performance 65

4.5 Scoring 70

5. CONCLUSION 72

REFERENCES 77

APPENDIX A: RESULTS OF CORRELATION ANALYSIS 80

APPENDIX B: RESULTING DECISION TREE WITH GINI INDEX 81

APPENDIX C: RULES OF EACH NODE IN DECISION TREE 83

APPENDIX D: SCORING CODE OF RESULTING REGRESSION MODEL86

CURRICULUM VITAE 93

ABBREVIATIONS

SEG	: SAS Enterprise Guide
SEG	: SAS Enterprise Miner
AI	: Artificial Intelligence
RDBMS	: Relational Database Management Systems
OLAP	: On Line Analytical Processing
IBM	: International Business Machines
KDD	: Knowledge Discovery in Databases
DM	: Data Mining
CRM	: Customer Relationship Management
CART	: Classification and Regression Trees
CHAID	: Chi-Squared Automatic Interaction Detection
AID	: Automatic Interaction Detection
ID3	: Iterative Dichotomizer
DSS	: Decision Support System
DWH	: Data Warehouse
USD	: United States Dolar
3G	: Third Generation
HSDPA	: High Speed Data Access
WiMAX	: Worldwide Interoperability for Microwave Access
GSM	: Global System for Mobile Telecommunications
SIM	: Subscriber Identity Module

LIST OF TABLES

Table 2.1 Evolution of data mining 11
Table 3.1 A data mining data mart example.. 21
Table 3.2 Confusion Matrix 38
Table 4.1 List of variables for churn prediction modeling................................... 42
Table 4.2 Frequency and percentage of churners in target population................. 47
Table 4.3 Rejected Variables.. 49
Table 4.4 Contract cancellation reason 50
Table 4.5 Contract status.. 50
Table 4.6 Distribution of Num_Complaint_LM and G_Num_Complaint_LM.... 51
Table 4.7 Input Variables 54
Table 4.8 Critical Factors Affecting Prepaid Churn... 60
Table 4.9 Results of Stepwise Regression.. 64
Table 4.10 Confusion Matrix of Stepwise Regression Model 68
Table 4.11 Confusion Matrix of Gini Tree Model 69
Table 4.12 Confusion Matrix of a Validation Data Set.. 70
Table 5.1 Average of different OnNet phone numbers that Customer dialed ..….73
Table 5.2 Ratio of Churners Grouped by Value Segment.................................... 74
Table 5.3 Comparison of random and model selection.. 75
Table A 1 Results of correlation analysis …….…………………………………..80

LIST OF FIGURES

Figure 1.1 : Subscriber bases around the globe 5
Figure 1.2 : Regional mobile density 5
Figure 1.3 : Leading mobile markets 6
Figure 1.4 : Turkish mobile communication market share 7
Figure 2.1 : Standard and data mining approaches on information detection....... 12
Figure 3.1 : The generic data mining method 19
Figure 3.2 : Data mining modeling process 25
Figure 3.3 : Pruning of a tree 29
Figure 3.4 : ID3 algorithm by Quinlan 31
Figure 3.5 : Ordinary Linear Regression 35
Figure 3.6 : Logistic Regression 35
Figure 3.7 : Noncumulative and cumulative response curves 36
Figure 3.8 : Cumulative captured response curve and lift curve 37
Figure 3.9 : Captured response curves for two different models 37
Figure 4.1 : Time frames in Churn Modeling 45
Figure 4.2 : Physical view of data mining data mart 45
Figure 4.3 : Modeling and scoring time frame.. 47
Figure 4.4 : Prepaid churn prediction model... 48
Figure 4.5 : Captured response curve of Entropy, Gini and Chi-Sqaure tree. 57
Figure 4.6 : Comparison Lift Chart for Selecting Number of Branches of Tree .. 59
Figure 4.7 : Gini Rule describing overfitting 61
Figure 4.8 : Gini Rule describing the customers who are more likely to churn ... 61
Figure 4.9 : Captured response curve of regression models 64
Figure 4.10 : Response curve of the models 66
Figure 4.11 : Captured response curve of the models.. 67
Figure 4.12 : Lift curve of the models 68

LIST OF SYMBOLS

P_i^2	: Probability of the class (i) being choosen twice
logit(pi)	: logit transformation of the probability of the event
β_0	: Intercept of the regression line,
β_1	: Slope of the regression line,
p_i	: Probability of ith event for a binary variable,
X_1	: Predictor variable.
r	: Number of target levels
B	: Number of branches

CHURN MODELING IN TELECOMMUNICATIONS SECTOR

SUMMARY

Data mining is a process applied for the exploration of the data in order to find valuable hidden information. Churn is the word mostly used in the telecommunications industry and generally the action of the customer to leave the company for some reason, so a churn model predicts which customers are likely to leave your company in the near future. Churn modeling is one of the application areas of data mining widely used for customer retention in all industries. This allows companies to increase customer loyalty by implementing effective retention strategies.

This thesis focuses on the wireless telecommunications sector in Turkey. If a prepaid customer does not refill his/her card within 6 months after his/her last refill, the wireless company subject to this study cancels the contract of the prepaid customer. A churn prediction model is developed to produce a score for each prepaid individual customer who is likely leave the company in 6 months due to this involuntary reason.

Logistic regression and decision trees are used to predict the churners. All the algorithms that are employed in this study are illustrated in detail. Number of complaint calls done by the customer to Call Center for last month, being a member of loyalty program, total number of refills of the customer for last 6 months and number of different phone numbers that customer dialed last month in the network of this wireless company have been found as most significant indicators of prepaid churn.

After determining which customers are likely to churn, this study ends with assessing churners value to the organization and provides recommendations on how to use churn scores in order to show the importance of proactive customer relationship management.

TELEKOM SEKTÖRÜNDE AYRILACAK MÜŞTERİLERİN TAHMİNİ

ÖZET

Veri madenciliği, verinin incelenerek, gizli ve değerli bilginin keşfedilmesini sağlar. Ayrılacak müşterilerin tahmini tüm sektörlerde çok sık kullanılan veri madenciliği uygulama konularından biridir. Rakip firmaya geçmeyi planlayan müşterilerin tahmini ise şirkete bu müşterilerin bağlılığını arttırmayı hedefleyen kampanyalar düzenleme fırsatını ve değerli müşterilerin elde tutulmasını sağlar. Böylece şirket müşteri ayrılmadan stratejik kararlar ve önlemler alabilir.

Bu tezde yapılan çalışma Türk iletişim sektörüne yöneliktir. Bu teze konu olan mobil iletişim şirketi, eğer kontörlü bir hat son kontör yüklemesinden sonra 6 ay içinde tekrar kontör yüklemezse, bu müşterinin kontratını iptal eder. Bu çalışma, sözkonusu telekom şirketinden ayrılacak müşterilerin tahmininin veri madenciliği teknikleri ile modellenmesi ve bu sayede her müşteriye bir puan atanmasını kapsar. Bu puan, müşterinin 6 ay sonra şirketi terk etme olasılığını gösterir.

Her veri madenciliği uygulaması arkasında çalışan algoritmalar vardır. Karar ağaçları ve regresyon modellemesi ayrılacak müşterilerin tahmini sırasında kullanılmış ve bu algoritmalar tez içinde ayrıntılı olarak sunulmuştur. Kontörlü hatta sahip müşterilerin geçen ay çağrı merkezine yaptıkları şikayet sayısı, bir bağlılık programına üye olmaları, son 6 aydaki kontör yükleme sayıları ve geçen ay bu mobil şirketin şebekesinde aradıkları farklı kişi sayısı müşterilerin şirketten ayrılma tahmininde önemli birer gösterge olarak ortaya çıkmıştır.

Bu çalışmada hangi müşterilerin gideceği bulunduktan sonra, gitme ihtimali yüksek kontörlü müşterilerin elde tutulması ile şirkete sağladıkları katkı değerlendirilmiştir. Müşteri ilişkileri yönetiminde erken önlem alınabilmesi için, gitme ihtimalini gösteren olasılık değerlerinin nasıl kullanılabileceği konusunda önerilerde bulunulmuştur.

1. INTRODUCTION

Proactive customer management is becoming essential for all sectors due to competition incited by globalization and maturing markets. Companies no longer want to treat their customer base as a collection of revenue generating units. On the contrary they want to have close and personal relationship with each of them. Customer relationship management (CRM) starts with in-depth analysis of the customer behavior, their habits, desires and needs [1].

Nearly all organizations planning to implement a customer relationship management program will need a decision support system (DSS). A data-driven DSS analyzes large pools of data found in organizational systems, and supports decision making by allowing users to extract valuable information [2, 3]. It is the data warehouse (DWH) that becomes the hearth of a DSS in which all the data from operational databases are collected. Online analytical processing (OLAP), data mining and statistical tools are used to analyze the data stored in DWH and come up with deductive conclusions. The results of the analysis and profiling activity are used extensively in planning promotional campaigns and treatment strategies.

The telecommunications sector has quickly evolved from offering local and long-distance telephone services to providing many other comprehensive communication services including voice, fax, cellular phone, images, e-mail and other data traffic such as web data transmission. The telecommunications market is rapidly expanding and highly competitive due to regulation changes in telecommunications sector in many countries and the development of new computer and communication technologies. This creates a great demand for data mining in order to help understand the business involved, identify telecommunications patterns, make better use of resources and improve the quality of service [4].

The market in the telecoms industry is maturing today and recognizes the importance of proactive customer relationship management because as the market gets saturated, growth comes from these three areas:

- Cross-selling and up-selling: maximizing the profit of existing customers
- Retention and up-selling: keeping profitable customers and getting rid of (or upgrading) unprofitable ones
- Poaching: stealing new customers from competitors

Acquiring new customers is more expensive than retaining existing customer base in telecommunications sector due to handset subsidization, welcome letters or calls and some other costs. Loyal customers are willing to buy more from the vendor they like and trust. Therefore, telecommunications companies realize that it is more profitable to retain the current customers and churn prediction is one of the important application areas of data mining to keep valuable customers.

Churn is the word used in the telecommunications industry and generally the action of the customer to leave the company for some reason, so a churn model predicts which customers are likely to leave your company in the near future. In other words, churn prediction is a process of identifying, predicting and understanding which customers are likely to leave and switch to other competitors in the future. A churn prediction application, as an end-to-end solution, accesses the data on customers to derive accurate predictions of who is likely to churn, assesses their value to the organization and provides an insight into what factors are influential in making their decisions [5].

After determining which customers are likely to churn, an effective retention strategy can be implemented by targeting a particular Marketing effort on these customers such as giving customers discounts on air time. Churn analysis is utilized by not only Marketing but also customer services, sales and finance. These departments need to determine what causes to churn, how much financial impact it has on the company, and how sales and customer service area may be able to prevent churn.

A real-world example for churn management is briefly illustrated in this section to gain better insight to churn management. Bouygues Telecom, from France, launched its DCS 1800 mobile telephone network in May 1996 and achieved a remarkable success. It has more than 300.000 customers, and revenue per customer has exceeded expectations. New telecommunications operators enter the market with aggressive acquisition policies to achieve a positive return on investment in shortest time possible. The result for the end customer is richer product variety, lower prices, more services. The downside for the Bouygues Telecom is a dramatic increase in the customer churn. To overcome this problem, they established a customer retention and development department responsible to take activities to minimize churn. The keyword to achieve this is the data mining for churn prediction. Their customer retention strategy is to identify customers having highest probability to churn by applying data mining study, and to take Marketing actions proactively to prevent those customers from churn. They build a churn model using the product Churn/CPS of the SLP company and rate their customers with their churn score. Churn/CPS lets the Bouygues Telecom to segment and score the customers, build models to predict churn [6].

Customers can leave the company for many different reasons. According to the churn reasons, we can identify the different kinds of churn. We can categorize churn by who initiates the action — the company or customer.

We call it voluntary churn if the customer first initiates the action. In this case, it can be categorized further down, based on the various churn reasons, such as contract expiration, handset change, service quality, competition, technology change, regulation change, and so forth.

Involuntary churn is another one if the company initiates the action. In this case, the company can decide to terminate their service with the customer for some reasons, for example, not paying the bill or not refilling prepaid card for several months.

There are two basic groups of customers for wireless telecommunications industry. Postpaid customers are the ones who pay at the end of every billing period. Prepaid customers pay before making any call.

In this study, the important objective is to implement churn prediction models to keep the profitable prepaid individual customers for a wireless telecommunications company. Since it is not recommended to predict all different kinds of churn together, involuntary churn for prepaid customers is aimed to be predicted in this study. If a prepaid customer does not refill his/her card within 6 months after his/her last refill, this wireless company cancels the contract of the prepaid customer.

The primary goal of churn model developed for this wireless company is to generate a list of prepaid contracts that are likely to be cancelled in 6 months. The customers holding these contracts can then be targeted with special offers in a timely manner that are designed to prevent them from involuntary churn. Otherwise it can result with losing customers to other competitors or contract activation costs in case the customer would like to get a new prepaid line. In addition to these, detecting the causes of churn that lie within the influence of the company makes it possible to focus on and eliminate those causal factors such as poor refilling procedures and technical quality issues.

The statistical data record for churn analysis in wireless communications is typically not the customer but the contract. In other words, propensity of cancellation is calculated on a per contract rather then per customer basis. The main reason is that various important predictor variables, such as the length of time since a contract was signed, are associated with contracts rather than customers. Furthermore, although a customer may hold several contracts, with each one contributing to revenue, the company wants to prevent them all from being cancelled wherever possible. However, since mailings and other follow-up actions are targeted at customers rather than contracts, there must be some post-analysis processing to summarize predictions for each customer. For example, corporate customers are usually treated separately from individual customers. Therefore this thesis focuses on individual customers.

1.1 Telecommunications Sector

With more than 2 billion subscribers worldwide at the end of 2005, mobile telephony has become the most popular telecommunications access mode, well ahead of landline calling and its base of 1,2 billion fixed main lines [7].

Mobile services now constitute operators’ prime source of revenues: earnings for 2005 are estimated at 573 billion USD, or half of the globe’s telecommunications services market [7].

In the space of a decade, the mobile sector has moved through the ranks to occupy centre stage in the telecommunications world, despite which the coming months will undoubtedly offer up a number of challenges:

- At the technological level, with the transition to 3G and the evolution to HSPDA, and still other fixed wireless networks like WiMAX, which are potentially both competitors for and complementary with cellular networks.

- At the industrial level, with a subscriber base that could grow to 3 billion users around the globe by 2010, and a handset market that is expected to remain healthy for several years to come [7].

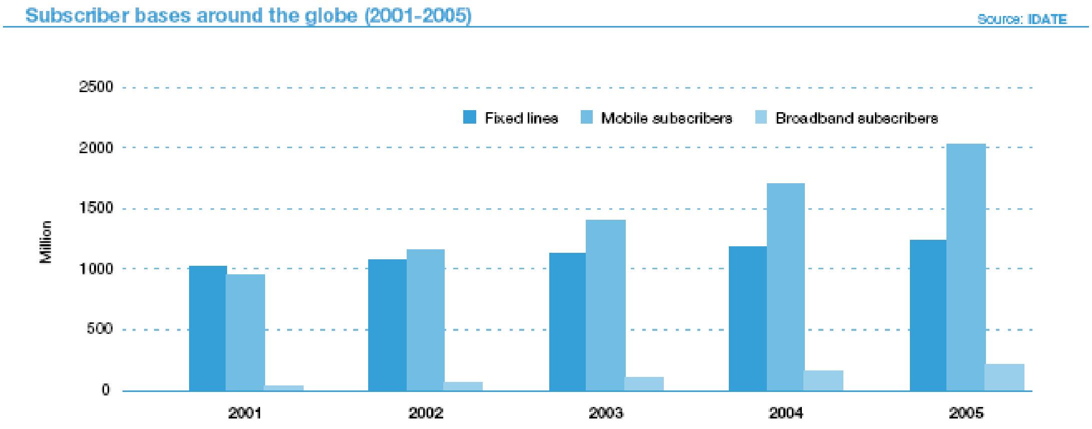


Figure 1.1: Subscriber bases around the globe

At the end of 2004, there were 1,7 billion mobile subscribers throughout the world as seen in Figure 1.1 [7]. The net increase in subscriber bases in 2004 (300 million) is to be found mostly in the developing countries. 32% of these additional subscribers came in fact from developing countries in Asia, 16% from Latin America and 11% from countries in the Africa/Middle East region as seen in Figure 1.2 [7].

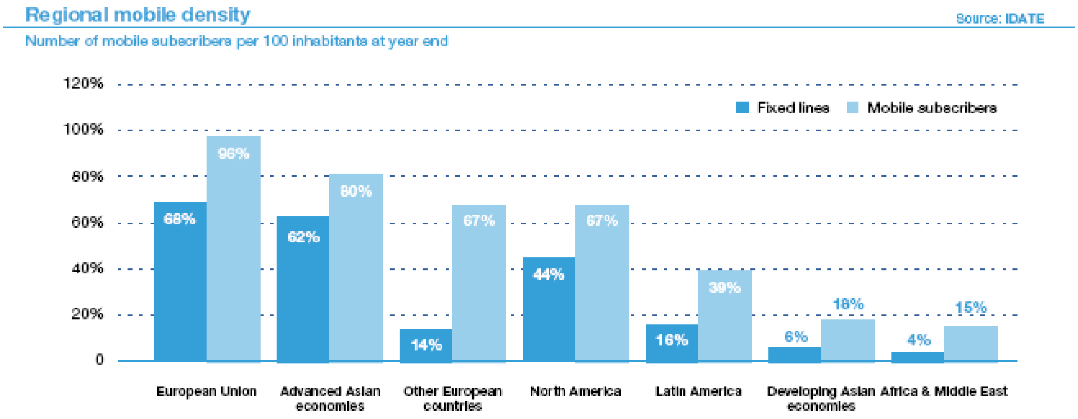
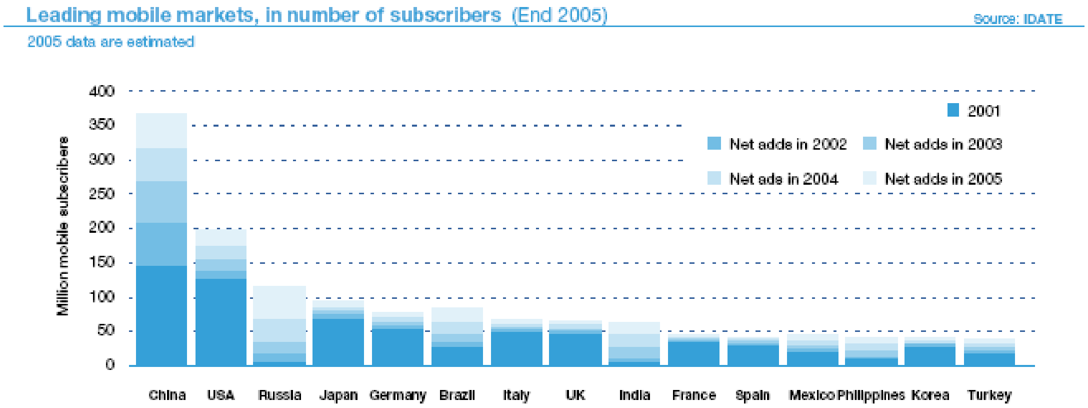


Figure 1.2: Regional mobile density


Leading mobile markets, in number of subscribers (End 2005)

Source: IDATE

2005 data are estimated

Figure 1.3: Leading mobile markets

The developing economies are thus gradually taking on increasing importance in the world market and accounted for 45% of the total number of subscribers around the world at the end of 2004 and Turkey is one of them as seen in Figure 1.3 [7].

Mobile services have been available in Turkey since 1986, when Turk Telecom (TT) launched its analog network. However, at that time, the incumbent's presence was negligible as it had about 150,000 subscribers and already reached its full capacity. The sector had its true launch in 1994, when Turkcell and Telsim started their operations with a revenue sharing agreement with TT. One year later, the mobile subscribers had reached two per cent of the total telephone subscribers, putting Turkey five years behind the world average in mobile penetration. By the end of 1998, the penetration more than doubled to reach five per cent with 3,4 million subscribers. Subscriber growth gathered pace after 1998 when each of the two operators received a GSM900 license [8].

Extensive Marketing and the launch of prepaid cards were the main catalysts that led to a rapid rise in penetration that marked 12 per cent at the end of the year 1999. In the second and third quarters of the year 2000, more than half of the additions were prepaid subscribers [8].

Two new operators joined Turkcell and Telsim in 2001. Aria which is owned by İşbank and Telecom Italia Mobile and Aycell subsidiary of Turkish Telecom was awarded a 25 year GSM1800 license.

Competition in the four-operator market was centered upon service quality, customer segmentation, subsidization and distribution. Telsim’s corporate image was still weak despite aggressive Marketing efforts that started in 2000. Aycell had an advantage of starting out with the most geographically diverse distribution network in place. Especially in the rural parts of the country, it would be a very valuable asset. On the other hand, the privatization of Turk Telecom continued to complicate matters for Aycell [18].

Aria and Aycell merged under Avea in 2004 to create important operational and financial synergies. The privatization process of the 55% of the shares of Turk Telecom was completed in November 2005. Accordingly, Oger Telecom through its 55% stake at Turk Telekom has joint control of Avea, along with Telecom Italia Mobile. Telsim was sold to British GSM operator Vodafone in December 2005. Based on these developments in Turkish mobile communication sector, Figure 1.4 shows the market share of the companies in this sector [36-38].

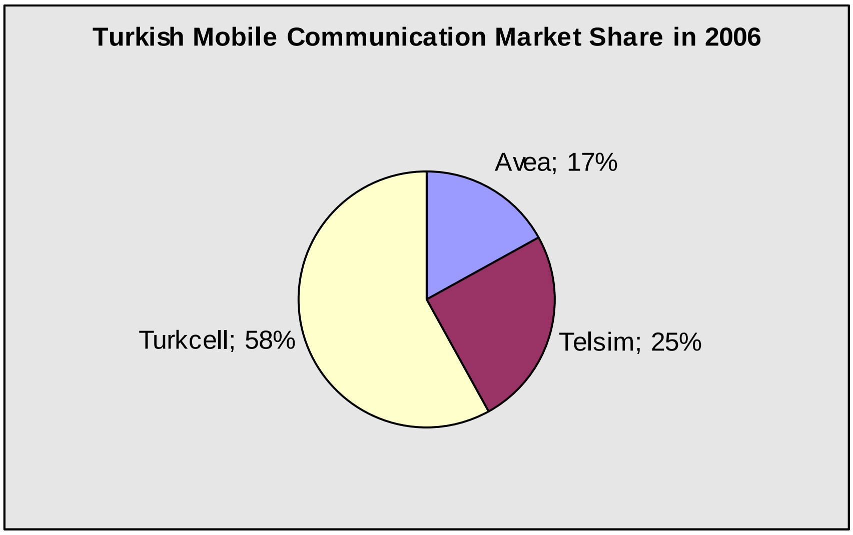


Figure 1.4: Turkish mobile communication market share

After all these recent developments in telecommunications sector in Turkey, competition is getting tougher each day and churn management is becoming vital for the competitors. This thesis develops a churn model for a wireless company to gain a competitive advantage in such a dynamic sector. Churn prediction for especially the existing GSM service providers is essential to fight with the aggressive Marketing strategies of new entrants.

2. DATA MINING

In this chapter, an overview of data mining is presented as an initial step. Afterwards, the methodology of data mining is explained in detail, starting with the definition of the problem which is the most important stage. The decisions about how the data mining will proceed is also considered in this step. Subsequently, data preprocessing steps including data acquisition, cleaning and transformation are handled. The data mart of mining process is obtained at the end of this step by transposing the tables into just one table.

Data mart is manipulated at the initial step of the data mining modeling. A subset of the data mart is prepared as a model set which will be used by the modeling techniques. Throughout the modeling stage, different candidate models are developed with different techniques and algorithms. The techniques that are employed in this study are also clarified in this section. Consequently, different methods for measuring the performance of those models are considered. The model with the best performance is chosen as the final model.

2.1 Data Mining Definiton

Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules [5]. Data mining uses technologies such as neural networks, decision trees or standard statistical techniques to search large volumes of data. In doing so, data mining builds models for patterns that accurately predict customer behavior.

Data mining is a discovery process, in that one can uncover information that would typically not found without data mining. Therefore the valuable information discovered at the end of data mining is what is implied by the phrase “hidden gold”.

Data mining has no fixed presentation of data and allows the user to create inquiries based on the information required.

Data mining techniques emerge as a new area with its strong foundation in applied statistics and artificial intelligence. This evolution began with storing the business data on computers, continued with improvements in data access methods, and nowadays emerging technologies allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigates to prospective and proactive information delivery.

Data mining is a collection of tools and techniques used for inductive rather than deductive analyses. Using sophisticated data mining tools, analysts explore detailed data and business transactions to uncover meaningful insights, relationships, trends, or patterns within the business activity or history in order to predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining is used to identify hypothesis; traditional queries are used to test hypothesis [9].

2.2 History of Data Mining

Recently data mining has been the subject of many articles in business and software magazines. However, just a few years ago, few people had even heard of the term data mining. Though data mining is the evolution of a field with a long history, the term itself was only introduced recently, in the 1990s.

Data mining roots are traced back along three family lines. The longest of these three lines is classical statistics. Without statistics, there would be no data mining, as statistics are the foundation of most technologies on which data mining is built. Classical statistics embrace concepts such as regression analysis, standard distribution, standard deviation, standard variance, discriminate analysis, cluster analysis, and confidence intervals, all of which are used to study data and data relationships. These are the very building blocks with which more advanced statistical analyses are underpinned. Certainly, within the heart of today's data mining tools and techniques, classical statistical analysis plays a significant role.

Data mining's second longest family line is artificial intelligence, or AI. This discipline, which is built upon heuristics as opposed to statistics, attempts to apply human-thought-like processing to statistical problems. Because this approach requires vast computer processing power, it was not practical until the early 1980s, when computers began to offer useful power at reasonable prices. AI found a few applications at the very high end scientific/government markets, but the required supercomputers of the era priced AI out of the reach of virtually everyone else. The notable exceptions were certain AI concepts which were adopted by some high-end commercial products, such as query optimization modules for Relational Database Management Systems (RDBMS).

The third family line of data mining is machine learning, which is more accurately described as the union of statistics and AI. While AI was not a commercial success, its techniques were largely co-opted by machine learning. Machine learning, able to take advantage of the ever-improving price/performance ratios offered by computers of the 80s and 90s, found more applications because the entry price was lower than AI. Machine learning could be considered an evolution of AI, because it blends AI heuristics with advanced statistical analysis. Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced AI heuristics and algorithms to achieve its goals.

Data mining, in many ways, is fundamentally the adaptation of machine learning techniques to business applications. Data mining is best described as the union of historical and recent developments in statistics, AI, and machine learning. These techniques are then used together to study data and find previously-hidden trends or patterns within. Data mining is finding increasing acceptance in science and business areas which need to analyze large amounts of data to discover trends which they could not otherwise find [10,11].

Evolution of business data to business information starts with data collection in 1960's. Data collection provides answers to retrospective questions. In other words, it addresses the questions related with past. In 1980's, with the relational databases, data access methods improved dramatically. In 1990's, data warehousing and decision support systems are founded with multidimensional databases, OLAP. Today, data mining is a field which utilizes advanced algorithms, multi processor computers and massive databases. The basic difference between data mining and previous fields is that data mining aims to answer prospective questions about future. The Table 2.1 describes evolutionary steps of data mining [12].

Table 2.1: Evolution of data mining

Year	Evolutionary Step	Enabling Technology
1960's	Data collection and database creation	computers, tapes, disks
1970's	Relational data model	faster and cheaper computers
1980's	RDBMS, advanced data models	faster and cheaper computers with more storage, On-line analytical processing (OLAP),multidimensional databases, data warehouses
1990's	Data warehousing and data mining	faster and cheaper computers with more storage, advanced computer algorithms

2.3 Data Mining Approach

Organizations are accumulating vast quantities of data in databases, with the recent trend to implement a data warehouse architecture increasing the quality and accessibility of data. This is all being done at great cost, but the information is only valuable if used effectively.

Users have been using query tools, OLAP servers, Business Intelligence tools, Enterprise Information Systems and a wide range of other packaged software to examine their data. However, the more numerate analysts have recognized that there are hidden patterns, relationships and rules in their data which cannot be found by using these traditional methods.

The answer is to use specialist 'data mining' software which harnesses advanced mathematical algorithms to examine large volumes of detailed data. Data mining is the process of extracting valid, previously unknown and ultimately comprehensible information from large databases and using it to make critical business decisions. The software is able to sift large volumes of data to find nuggets of information which yield gold in the form of competitive advantage.

Data mining can be carried out on any data file, from a spreadsheet to a data warehouse. Transaction processing systems can be mined to generate benefits which can help to justify implementing data warehouse architecture.

Data mining is very different from querying, where the user knows what is in the database and knows what information to ask for. 'Data mining with a query tool is like mining coal with a spoon,' says Tim Negris, IBM Software's Vice President for Sales and Marketing. 'Yes, you can do it, but human beings are doing what could be done by a computer. Those who ignore history repeat it, those who query repeat it [13].

Data mining is about discovering new things about business from the data that have been collected. Using standard statistical techniques to explore database does not discover new things. In reality what is being done is making a hypothesis about the business issue that is addressed and then attempting to prove or disprove the hypothesis by looking for data to support or contradict the hypothesis.

Data mining uses an alternative approach beginning with the premise that no patterns of data are known by the user. In this case, user does not have to develop a hypothesis, and just simply ask, what is new, interesting and valuable in the data. In this case, data mining algorithm tells about all the previously unknown and ultimately comprehensible information that users had. Data mining therefore provides answers without users having to ask specific questions.

The difference between the two approaches is summarized in Figure 2.1 [14].

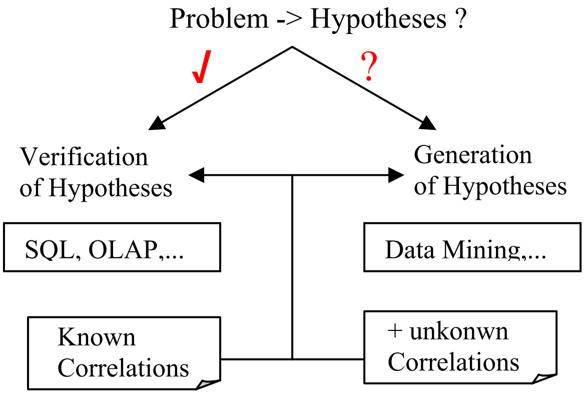


Figure 2.1: Standard and data mining approaches on information detection

Data mining has two different goals: verification of a user’s hypothesis, and discovery, the finding of new patterns in data. Discovery (sometimes referred to as knowledge discovery in databases, or KDD) includes prediction, (regression and classification) and description (summarization, visualization, and detection of changes and deviations). Some KDD tools are generic; others are domain specific. Domain specific tools represent an important trend, moving knowledge discovering technology directly into the hands of business users. Among the elements that make this possible are putting the problem in the business user’s terms, providing support for specific key business analyses, representing results in a form geared to the business problem being solved, and providing support for an exploratory process [15].

2.4 Application Areas

Data mining, extracting meaningful patterns and rules from large quantities of data, is useful in any field where there are large quantities of information and something worth learning. It does this by using sophisticated techniques such as artificial intelligence to build a model of the real world based on data collected from a variety of sources.

Data mining can be used for the following purposes:

- Research (e.g. pharmaceuticals industry…)
- Process improvement (e.g. manufacturing industry…)
- Marketing (e.g. insurance…)
- Customer Relationship Management (e.g. banking…)

Having fulfilled the purposes above, organizations understand customer behavior, predict buying patterns, eliminate inappropriate surgical procedures, detect fraud, and other applications. It has helped companies to provide more meaningful services to customers, increase revenue, reduce expenses.

Customer churn and loyalty are important issues for most of the companies. Data mining can predict which customers are likely to leave the company and go to a competitor. Using this information, Marketing strategies can be developed for customer retention as a part of Customer Relationship Management.

Fraud detection is widely used in credit card services and telecommunications. Data mining uses historical data to build models of fraudulent behavior and use techniques to identify which transactions are most likely to be fraudulent in future.

Below are some examples of application areas of data mining:

Retail

- Identifying customers for a spesific offering [16]

Pharmaceutical

- Discovering new uses for existing drugs

Healthcare - Insurance

- Utilization analysis of hospitals
- Identifying behavior pattern of risky customers
- Improving preventive care
- Predicting which customers will buy new policies

Banking

- Identifying loyal customers
- Predicting for credit card customer attrition
- Detecting patterns of fraudulent credit card usage
- Credit Scoring [17]

Telecommunications

- Identifying products and services that maximize life time value of customers
- Establishing Marketing campaigns to improve market share
- Improving cross selling efforts
- Customer Segmentation [18]

Internet Applications

- Determining most probable customers who will perform e-commerce [19]
- Identifying the user interest patterns [20]

Education

- Finding important factors for student success [21]

Government

- Understanding criminals behaviour and identifying suspicious ones [22]

2.5 Data Mining Techniques

Data mining techniques can be classified into two groups: **supervised techniques** and **unsupervised techniques**. If the user knows what he is looking for, this is a style of supervised technique. Supervised approach is very common in the business world and predictive modeling is an example of this. For instance, if the churn rate for the next month is to be predicted, then purpose of the mining process is known, and the knowledge about customers if they remained loyal or not is required. In other words, for supervised learning, independent variables are fed into the model and the dependent variable is predicted. The prediction is compared with the actual dependent value to assess the validity of the model.

There is no such process for unsupervised technique. Data speaks for itself. Unsupervised technique finds patterns and profiles in the data, and leaves the interpretation to the user. Clustering technique is an example of unsupervised learning.

In other words, unsupervised modeling is used to recognize relationship in the data, and supervised modeling is used to explain those relationships once they have been found [23].

2.5.1 Supervised (Predictive) Modeling Techniques

Predictive data mining is applied to a range of techniques that find relationships between a specific variable, called the target variable and the other variables in the data.

If the data element under investigation is discrete meaning that it has a small number of fixed values, the task is called classification. On the other hand if the data element is continuous, meaning that it can take a large number of values and exhibits a common unit of measure, the task is called regression.

Classification is a learning method frequently adopted in the fields of data mining, statistics, machine learning, genetic algorithm and neural networks [4]. The classification problem is a two-step process, where the first is to build a classification model by analyzing the training sample set described by attributes and the second is to use this model to classify the future sample for which the class label is not known.

For example, we can use the classification model learned from the existing customers' data to predict what services a new customer would like [24].

Classification predicts class or group membership. With classification the predicted output, the class, is categorical. A categorical variable has only a few possible values, such as yes-no, high-middle-low, etc. Regression predicts a specific value. Regression is used in cases where the predicted output can take on many possible values, and the output is therefore continuous.

Techniques that dominate the commercially available classification and regression tools today are, decision trees, neural networks, naïve bayes, K–nearest neighbor.

Decision Trees

Decision trees are a way of representing a series of rules that lead to a class or value. The graphic output is similar in structure to a tree. A decision tree consists of the decision node, branches, and leaves. The first component is the decision node, which specifies a test to be carried out. Each branch of a node leads either to another decision node or stopping point, called a leaf node. By navigating the decision tree, one can assign a value or class to a case by deciding which branch to take, starting at the top node and moving to each subsequent node until a leaf node is reached. Each node uses the data from the case to choose the appropriate branch.

Neural Networks

Neural networks are more complicated than other techniques. Based on an early model of human brain function, it mimics the brain's ability to learn from its mistakes. It is often referred to as a "black box" technology and involves very careful data cleansing, selection, preparation, and preprocessing. A neural network starts with an input layer, where each node corresponds to a predictor variable. These input nodes are connected to a number of nodes in a hidden layer. Each input node is connected to every node in the hidden layer. The nodes in the hidden layer may be connected to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables. Neural networks are often used in financial markets for prediction and forecasting.

Naïve Bayes

Naïve bayes analyzes the relationship between each independent variable (the predictor) and the dependent variable (the predictee) to derive a conditional probability for each relationship. When a new case is analyzed, a prediction is made by combining the effects of the independent variables on the dependent variable. Naïve bays require only one pass through the data to generate a classification model. This makes it a very efficient data mining technique. However it does not handle continuous data, limiting inputs only to categorical data.

K-Nearest Neighbor

K-nearest neighbor is a technique that classifies each record in a data set based on a combination of the classes of the K records most similar to is in a historical data set. It has no distinct training (model building) phase because the training data is actually the model.

2.5.2 Unsupervised (Descriptive) Modeling Techniques

Descriptive data mining is applied to a range of techniques which find patterns inside the data without any prior knowledge of what patterns exist. There are two common methods for unsupervised modeling, association and clustering.

Association

Association is used to determine which things go together. A typical application that can be built using an association function is market basket analysis. It finds affinity groupings that discover what items are usually purchased with others predicting the frequency with which certain items are purchased at the same time.

Clustering

Clustering is task of segmenting a diverse group into a number of similar subgroups or clusters. In clustering, there are no predefined classes and no examples. The records are grouped together on the basis of self-similarity. It is up to the miner to determine what meaning, if any, to attach to the resulting clusters. Clustering is often used to prepare data for another step in analysis. Some of the common algorithms used to perform clustering include Kohonen feature maps and K-means [5, 15, 25].

3. DATA MINING METHODOLOGY

A data mining application should be applied according to a flow. The flow is based on data mining logic. The generic data mining method comprises of seven steps.

- Defining the problem in a precise statement
- Defining the data model and the data requirements
- Sourcing data from all available repositories and preparing the data (the data could be relational or in flat files, stored in a data warehouse, computed, created on-site or bought from another party. They should be selected and filtered from redundant information).
- Evaluating the data quality
- Choosing the mining function and defining the mining run
- Interpreting the results and detecting new information
- Deploying the results and the new knowledge into business

These steps are illustrated in Figure 3.1

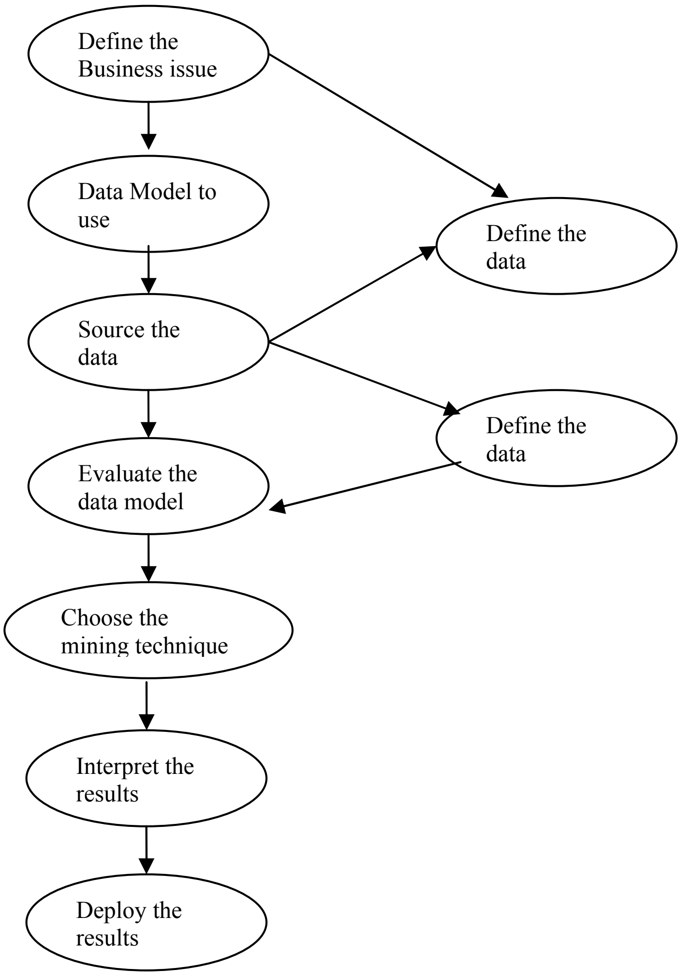


Figure 3.1: The generic data mining method

3.1 Problem Definition

Identification of the business problem is crucial for a successful data mining process. Defining the problem is the trickiest step since it involves a subjective approach. Only the business experts working on those specific fields know what the real problem is, and what the solutions are. However, the purpose of data mining is to let the data speak. Although business experts provide experience and intuition, this should be verified by examining data.

After talking to the business experts, data miners comprehend the problem and decide whether the data mining effort is necessary. Data mining process does not produce a solution for all real-world problems. Furthermore, business experts know what should be included in the data. For instance, only the business experts know which variables affect the churn prediction the most. At the end of this step, the problem will be identified. Consequently data mining variable list will be defined.

In this initial step, the period of time and the target population are specified. Time period is the interval of time from which historical data records are taken. Furthermore, the target population has to be specified since data will be prepared for the population of interest in the following steps. For example, business people may require to do churn prediction for only high-valued customers. Then the data related to high-valued customers will only be collected.

Identification of the business problem is the core step of the data mining process. All information necessary during the data mining process is defined at the problem definition step, and a mistake made at this step would cause the whole process to be unsuccessful.

3.2 Data Preprocessing

In this section, the acquisition of the data essential for data mining is introduced first. Afterwards, the way to clean data from invalid, missing values and outliers, and derivation of new variables are explained in detail.

3.2.1 Data Acquisition

The first step in data preparation is the data acquisition. More specifically, the relevant data set is identified, accessed and retrieved from various sources; converted and then consolidated in this step. Mostly, a data warehouse will speed up the data acquisition step. Data set has to be in such a format that all the data should be in a single table, and each row should correspond to an instance related to business problem definition.

All data mining algorithms use a standard format of the data which is called data mining data mart. A data mart is a subset of the data resource, usually oriented to a specific purpose or major data subject. Table 3.1 shows an example of a typical data mining table. Rows of this table indicate the units of action such as individual customers. For instance, each row corresponds to customers having valid phone numbers for a wireless communications company. The resulting table represents the target population during the data mining process. The columns, also called variables, of Table 3.1 represent the properties of selected units of action such as demographic information about customers.

Variable types can be defined by several ways such as value type, time period, derivation, and dependency. Nominal variables are the value type variables that cannot be ordered. They have a limited number of possible values. Examples for nominal variables would be the city where the bill is posted and type of customers. Ordinal variables can be ordered, such as age and flag variables. Nominal and ordinal variables are kinds of class variables since they have class types ordered or not. Continuous variables, also called interval variables, can take arbitrary values in a given range. Examples are paid invoice and frequency of transactions.

Table 3.1: A data mining data mart example

Id Field	**Customer Information**				**Transaction Data**			**Target**
	Age	Region	Job	M.Stat	Paid in $	Refill Frequency (day)	Number of SMS	
10001	19	MAR	Tailor	Single	10	60	1	TRUE
10002	25	EGE	Teacher	Married	10	50	5	TRUE
10102	40	MAR	Doctor	Married	50	10	10	FALSE
…	…	…	…	…	…	…	…	…
10100	55	KAR	Engineer	Married	20	30	7	FALSE

Historical variables represent data that occur at regular time intervals. When the time period moves forward, the value of the historical variables for the new month replaces the ones for the oldest month. Customer billing data is the most well known example of the historical variables. However, static variables have fixed values during the time period of interest. It may change over the period but those changes are usually not recorded. Area code and address of a customer are some examples.

The most beneficial variables are derived from existing data using some statistical measures such as summation, mean and variance. Total amount of invoice can be an example of a derived variable calculated by summing up all monthly invoices.

On the other hand, for supervised data mining applications, there is a variable different from the others. Target variables are predicted at the end of the data mining modeling. It is appropriate to describe target variables as dependent and the others as independent. In the case of churn modeling, target is a binary variable indicating whether the customer has churned or not. For unsupervised data mining studies like customer segmentation, there is no need for target variables since there is no prediction.

3.2.2 Data Cleaning

Data sets are not always clean and accurate. A column containing a list of city names may have the values “New York”, “NewYork” or “new york”, although all three denote the same city. This is referred as a consistency problem.

Another data cleaning problem is the misspelled or wrongly typed words. Data warehouses usually solve the problem of data cleaning since data coming from the operational databases are consolidated and cleaned during the loading process. Even if data are obtained from a data warehouse, it is possible that data set is not clean. In this case, data formatting must be carried out.

Some data points may have values that are quite rare or far out-of bounds from others in the column. Those values are called outliers because they are outside the expected values of the data. If the outlying value is extreme, it could seriously alter the accuracy of a model that is built. Outliers are handled by one of the two methods depending on the modeling algorithm. First method is the exclusion of the outlying value from the analysis. Second method is the imputation of an outlier that is replacing it with the mean, minimum or with another value.

Missing values represent another kind of dirty data. There are several reasons why missing data occur. Customer demographic data usually have a high percentage of missing values. Since this type of data is usually categorical, it is possible to either discard the missing field or replace it with a value. If a customer answers all of the questions in a survey except one of them, this indicates more than a missing value. For instance, customer may not give his phone number on purpose since s/he wants to be in fraudulent activity in the near future. This kind of empty data is valuable during the data mining process, and they must be considered as useful information. They are usually coded with “U” or “Unknown”.

Treatment of missing values for interval variables are much more difficult since assignment of a value for a numerical missing field changes the statistics. For transactional kind of data, missing value usually means nonexistent data. For instance new customers will not have data for previous months, or if a customer has no credit cards, then no data related to credit card transaction will be witnessed. All missing values for this type of fields are replaced with zero. However, for other numerical fields such as age, assigning a value for missing values changes the statistical distribution of the field. Missing values of this type is excluded from the analysis or imputed somehow. Some algorithms like decision trees can manage missing values directly without any imputations but regression cannot. There are several ways to impute missing values in numerical data fields such as assigning the average, mean or modal value, and building the classification model to impute the missing values.

3.2.3 Data Transformation & Transposition

After cleaning the data from outliers and missing values, the next phase is data transformation. This step concerns the derivation of a new variable and altering its distribution. For statistical analysis, it is expected for the input data to have a linear relationship with the target variable and error terms to be normally distributed with common variance. However, machine-learning methods do not constrain data to be normal or linear. Even methods like decision trees can manage outliers. Furthermore, nonnormality and heteroscedasticity can be handled through transformations such as logarithmic, power or trigonometric transformations. Another transformation method is the discretization of numeric data using quantiles which makes it easy to interpret data.

Derivation of new variables is essential for data mining projects. Ratios, averages, trend and variances are all derived variables from existing ones, and they perform significant role during the modeling process.

Data mining algorithms work on a single table. The necessary data describing an individual customer must be in a single row. Data sets could be stored in several tables, and those tables must be merged into one table. It is also possible to handle the transposition of the data at the first steps of the data management, and then deal with data cleaning, missing value imputation and data transformation.

3.3 Modeling

In this section, the way to develop a data mining model is clarified. First of all, the data set for modeling is prepared and then the model building begins. This section ends with describing the approaches to assess the performance of the models developed.

3.3.1 Preparation of the Model Data Set

The model data set describes the data that is used as input to the modeling technique. A predictive model is as good as the data which has been used to build the model. Before data mining modeling, data partitioning is applied as shown in Figure 3.2 and the data set is split into three data sets: training set, validation set and test set. All of these three data sets have to be mutually exclusive and contain no common records.

There is an important point to consider during the data splitting. If the data mining style is unsupervised such as customer segmentation, then it is not necessary to split the data. There is no target variable to test the efficiency of the models. For supervised data mining styles, different data sets are used with different objectives. Training data set is used for preliminary model fitting. The analyst attempts to find the best model weights using this data set. Validation data set is used for monitoring and tuning the model. Monitoring process involves selecting among models of different types. Tuning process optimizes the selected model on validation data, for instance pruning in decision trees. Test data set is used to obtain a final result of the model. Cases in the test set must be treated as new data.

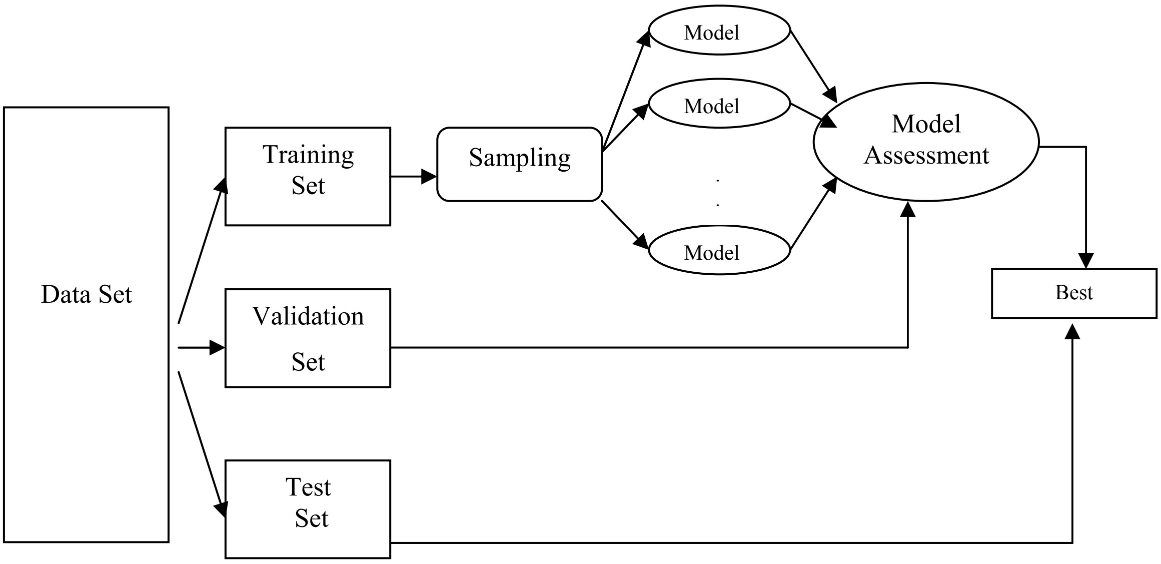


Figure 3.2.: Data mining modeling process

There is another critical concern during the preparation of the model set. If the data mining model is predicting a rare case, the percentage of this case must be clearly identified. Oversampling is the method of taking more of the rare events and fewer of the common events. If there are two events to occur, 20-30 per cent of the rare event often produces satisfactory results [5]. Oversampling increases the proportion of the less frequent outcome. Oversampling is called as stratified sampling in statistics. All the records of the given outcome are called a stratum of the data. A stratified data set is the one in which the modeler sets the ratio of responders to non-responders to a desired value and an optimal stratification ratio is achieved by trial and error.

The last step of preparing the model set is the variable selection. At the end of data preprocessing, there could be hundreds of variables, which are too many to implement the model and they could be superfluous by presenting the same or very similar information as others, but increasing the run time. Dependent or highly correlated variables could be found with statistical tests like bivariate statistics, linear and polynomial regression. Dependent variables should be reduced by selecting one variable for all others or by composing a new variable for all correlated ones by factor or component analysis.

Not all variables remaining after the statistical check are nominated as input; only variables with a clear interpretation and variables that make sense for the end user should be selected. A proven data model simplifies this step. The selection of variables in that stage can indeed only be undertaken with practical experience in the respective business or research area.

3.3.2 Model Building and Testing

The fundamental step of the data mining process is building a model for a given business issue. Various different models are built on the training data set as demonstrated in Figure 3.2. Constructing a good model depends on both performance and complexity of the model. Neural networks may perform well but it is difficult to explain the results of a neural net for business use. When the number of layers is increased in a neural network or the leaves of a decision tree is enlarged, and then the complexity of the models will increase. The complexity level of a model can lead to underfitting or overfitting scenarios. A model with lowest complexity is not flexible enough and causes underfitting. An over complex model is too flexible and causes overfitting. The model developed using the training data is applied to the test data to score the customers. Since the actual value of the target variable is known in the test data set, the comparison between the predictive and the actual results is carried out. The method of model comparison is explained in the next section.

Data mining tools use machine learning or statistical modeling techniques such as decision trees, neural networks, k-means clustering, belief networks, and regression analysis. Many different algorithms are used for implementing these techniques. This section ends describing the techniques and algorithms that are specifically employed during this study.

3.3.3 Decision Trees

Decision trees are the analytical tools used for discovering the rules in the data. Tree-like structures are employed in computer science for a long time but it hasn't been a preferred process of knowledge discovery until 1984. In 1984, L. Breiman, J. Friedman, R. Olshen and C. Stone wrote a book called Classification and Regression Trees discussing a decision tree approach called CART. Later in 1993, J.R. Quinlan published the book "Programs for Machine Learning" and introduced an extension algorithm of ID3, C4.5 [12, 26].

A decision tree is a predictive model that can be viewed as a tree where each branch is an outcome of the test, each internal node is a test on a single attribute and each leaf is a class or class distribution. Since their structure and ability to generate rules easily, decision trees are favored techniques for building understandable models. Various decision tree algorithms exist like ID3/C4.5, CART and many others. Each produces trees that differ from one another in the number of splits allowed at each level of tree, how those splits are chosen when the tree is built, and how the tree growth is limited to prevent over fitting.

Decision trees are used in various applications of data mining for prediction, data preprocessing or exploration. Looking at the predictors and values that are chosen for each split, decision trees are used for exploration of the data set and business problem. For instance by only looking through the decision tree produced for mobile phone churn, it is possible to observe that if the sales channel is newspaper campaigns, then the customer churn is really high. Decision trees are also used for preprocessing data for other prediction algorithms. They are used on the first pass of data mining run to create a subset of possibly useful predictors that can be fed to other algorithms like neural networks. Also decision trees are increasingly being used for prediction.

There are also some problems where decision trees will not do as well. Some very simple problems in which there are just a few numbers of the predictors can be solved much more easily by regression analysis.

There is recent research in the machine learning and statistics communities on algorithms for decision tree classifiers. Among decision tree algorithms, C4.5 and CART have the best combinations of error rate and speed [27].

Decision tree algorithms are a form of supervised learning. The first step in the algorithm is the process of making the tree grow. Algorithm seeks to create a tree that works as perfectly as possible on all the available data. The goal of decision trees is to have homogenous leaves with respect to the prediction value. Decision trees are built through a process known as recursive-partitioning. It is an iterative process of splitting the data into partitions until some stopping points [26, 28].

Step 1: Independent and dependent variables are chosen from a data source. According to the goal of the data mining, the user chooses a dependent, in other words, a target variable.

Step 2: A variable among independent variables is deemed to be the most predictive for the dependent variable and is used to split the data. An obvious question at this point would be how a decision tree will pick one predictor among others and make the split. The algorithm chooses the split that partitions the data into parts that are purer than the original. In other words, decision tree algorithms choose the best split which decreases the disorder of the data set more than the other splits. The best split reduces the disorder of the whole data by creating more ordered smaller partitions.

Step 3: Then this splitting is applied to each new partition. Each new partition will now be the input of the algorithm recursively until the stopping criterion is achieved.

Step 4: Most of the decision tree algorithms stop growing the tree when each new partition is completely organized into just one value for the target variable or when the partition contains algorithmically defined minimum number of records. This is because of the statistical reasons. Few records are not enough to make predictions based on historical data.

After the tree has been grown enough to a certain size depending on the algorithm, there is still more work to do. The algorithm then should check if the model overfits the data. To overcome this problem, decision tree algorithms are using validation approach and trying many different simpler versions of the tree on a held-aside validation set. Pruning is the process of improving the performance of decision tree by removing leaves and branches. A pruned tree is in fact a subset of the full decision tree. Tree building algorithms make their best split at the beginning. Each partition is smaller then the whole population and undergoes to the same splitting criteria. This process goes on until the stopping criteria. As the partitions get smaller, they are being less representative of the population and overfit the data. There are some pruning techniques referred as bonsai techniques for avoiding overfitting. Bonsai techniques [5] try to stunt the growth of the tree before it gets too deep. This is also called top down pruning. For instance, setting a limit for the minimum number of records that must be in a node is a type of bonsai techniques.

Pruning methods let the decision tree to grow quite deep and then to prune off the branches that fail to generalize the data. This is called bottom up pruning. One common approach is to find the error rate associated with the subtrees of the initial tree. But these error rates should not be calculated with the same data set. Error rate decreases as the tree gets more complex. So a complexity term is added to the error rate to discourage greater complexity. An addition of a branch is allowed only when its improvement in tree performance is large enough to overcome the extra complexity. When the data is large enough to build some test data sets, the performance of initial tree and subtrees are measured on separate test data sets. Figure 3.3 shows how pruning works with training and test data sets [5].

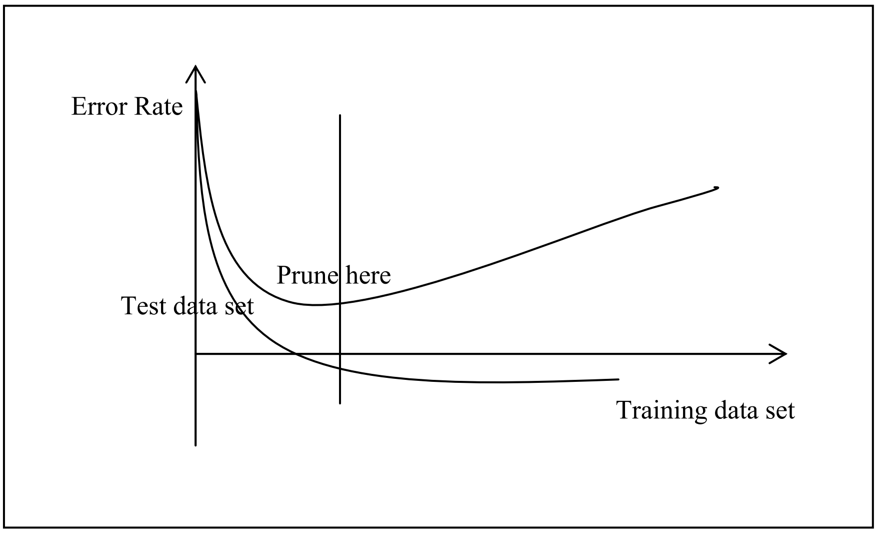


Figure 3.3: Pruning of a tree

While all decision-tree algorithms have the similar type of process, they employ different mathematical algorithms to determine the splitting criterion. They use different methods to find the best split that decrease the disorder of the data set. ID3 by Quinlan and CART by Brieman are two decision tree algorithms that will be explained next.

ID3/C4.5/C5.0 are algorithms introduced by J. Ross Quinlan for inducing decision trees [16]. The basic ideas behind ID3 are that:

- In the decision tree, each node corresponds to a non-categorical attribute, in other words a predictor, and each arc to a possible value of that attribute. A leaf of the tree specifies the expected value of the categorical attribute, independent variable, for the records described by the path from the root to that leaf.

- In the decision tree at each node, a non-categorical attribute which is the most informative is chosen among the attributes not yet considered in the path from the root.

- Entropy is used to measure how informative is a node. ID3 chooses a non-categorical variable on the basis of the information gain that this split provides. Gain represents the difference between the information needed to make a prediction correctly before and after the split. In other words, if the information required is much lower after the split is made, then it means that the split has decreased the disorder of the initial single data segment.

Information gain is the difference between the entropy of the original segment and the accumulated entropy of the resulting split segments. Entropy is a well-defined measure of disorder or information found in the data [29].

If there are n equally probable possible messages, then the probability p of each is 1/n and the information conveyed by the message is "-log (p) = log (n)". That is, if there are 16 messages, then log(16) = 4 and 4 bits are needed to identify each message [30].

In general, if the following probability distribution is given, P = (p_1, p_2... p_n), then the information found in this distribution*,* also called entropy of P *is:*

$$Info(P) = -(p_1 \log(p_1) + p_2 \log(p_2) + ... + p_n \log p_n) \qquad \textbf{(3.1)}$$

The ID3 algorithm is used to build a decision tree, given a set of R which is composed of non-categorical attributes C_1, C_2...$C_{k,}$ the categorical attribute C, and a training set T of records. At the beginning, only the root is present. At each node the following divide and conquer algorithm is executed, trying to choose the best split, with no backtracking allowed as seen Figure 3.4.

```
Function ID3 (R: a set of non-categorical attributes,
              C: the categorical attribute,
              T: a training set) returns a decision tree;
begin
        If T is empty, return a single node with value Failure;
        If T consists of records all with the same value for
                the categorical attribute, return a single node with that value;
        If R is empty, return a single node with as value the most frequent of
                the values of the categorical attribute that are found in
                records of T;
        Let D be the attribute with largest Gain(D,T) among attributes in R;
        Let $\{d_j| j=1,2, .., m\}$ be the values of attribute D;
        Let $\{T_j| j=1,2, .., m\}$ be the subsets of T consisting  respectively of
                records with value $d_j$ for attribute D;
        Return a tree with root labeled D and arcs labeled  $d_1, d_2, .., d_m$ going
                respectively to the trees
        ID3(R-{D}, C, $T_1$),
        ID3(R-{D}, C, $T_2$),
         ...,
        ID3(R-{D}, C, $T_m$);
end ID3;
```

Figure 3.4: ID3 algorithm by Quinlan

High cardinality predictors in ID3 affect the accuracy of the resulting model. This is because of the many small segments that will be formed with little data in them. A high cardinality predictor is the one that has many possible values to perform splitting. A field with a customer name or a zip code can be taken as examples. An experienced data mining specialist will discard those variables. For instance, the number of records in data set is 20 and there are 20 different customer names. If the splitting criterion is customer name for churn prediction, then there will be 20 different small segments with only one record in each of them. The segments are fully homogenous. Entropy of those resulting 20 segments is zero meaning no disorder anymore, and there will be no other splits that will be better than this. So the model will choose customer name as the best split, but this model will never work well for the data other than this historical data.

To deal correctly with those high cardinality predictors, ID3 improved the gain theory and introduced gain ratio. J.R. Quinlan suggests using the following ratio instead of using only gain:

$$GainRatio(D,T) = \frac{Gain(D,T)}{SplitInfo(D,T)} \qquad \textbf{(3.2)}$$

where SplitInfo(D,T) is the information due to the split of T on the basis of the value of the noncategorical attribute D. Thus SplitInfo(D,T) is

$$SplitInfo(D,T) = -(p_1 \log(p_1) + p_2 \log(p_2) + ... + p_n \log p_n) \qquad \textbf{(3.3)}$$

where D has n classes.

C4.5 is an enhancement of ID3 algorithm in several subjects. Non-categorical variables with missing values can still be used. The algorithm can deal with training sets that have records with unknown attribute values by evaluating the gain, or the gain ratio, for an attribute by considering only the records where that attribute is defined. Using produced decision tree, records that have unknown attribute values can be classified by estimating the probability of the various possible results [31].

Leo Brieman, Jerome Friedman, Richard Olshen and Charles Stone (1984) developed CART (Classification and Regression Trees) [26]. In building a CART tree, best predictor is picked according to how well it splits apart the records with different predictions. As explained before, the major difference between decision tree construction algorithms is choosing the best split. The measure is the reduction in diversity for choosing the best split. Information gain is the method used by C4.5 algorithm and Gini index is another diversity metric used by CART algorithm. The Gini value for a given segment is calculated to be one minus the sum of squared probabilities for each prediction. In other words, Gini index is the probability that the second event chosen belongs to a different class than the first. Therefore, Gini index will be maximum when the proportions of each prediction value are equivalent, as well as highest entropy, and it will be the minimum when the segment is homogenous. The limiting value is 0.5 for binary events or 1/n when there is n number of categories. This value is reached when each class has exactly the same number of members. The total reduction in diversity is the diversity at root minus the weighted average of the diversity of the segments.

Probability of the class (i) being choosen twice is P_i^2 . The diversity index is simply one minus the sum of all P_i^2 . When there are only two classes the formula is found [5]:

$$Gini\ (T) = 2\ P_1(1-P_1) \qquad \mathbf{(3.4)}$$

$$Gini(X,T)= {}^{n}\textstyle\sum_{i=1}\ (|T_i|/\ |T|)\ Gini(T_i) \qquad \mathbf{(3.5)}$$

$$Gain(X,T)= Gini(T) - Gini(X,T) \qquad \mathbf{(3.6)}$$

The Pearson chi – squared test can be used to judge the worth of the split. It tests whether the class proportions are the same in each child node. The test statistics measures the difference between the observed cell counts and what would be expected if the branches and target classes were independent.

The statistical significance of the test is not monotonically related to the size of the chi-squared test statistics. The degrees of freedom of the test is (r-1)(B-1) where r is number of target levels and B is number of branches. The expected value of a chi-square test statistics with v degrees of freedom equals v. Consequently, tree with more branches, will naturally have larger chi-squared statistics [40].

3.3.4 Logistic Regression

Regression analysis enables you to characterize the relationship between a response variable and one or more predictor variables. In linear regression, the response variable is continuous. In logistic regression, the response variable is class type [32].

Logistic regression model uses the predictor variables, which can be class type or continuous, to predict the probability of specific outcomes. Since the probabilities are being modeled, a linear regression model would not be appropriate. The relationship between the probability of the outcome and a predictor variable is usually nonlinear. Because of this, a logistic regression model applies a transformation to the probabilities. For a binary outcome variable, logistic regression model with one predictor variable have the form:

$$log\,it(p_i) = ln(\frac{p_i}{1 - p_i}) = \beta_0 + \beta_1 X_1 \qquad \textbf{(3.7)}$$

where

logit(p)

$logit(p_i)$=logit transformation of the probability of the event,

β_0 = intercept of the regression line,

β_1 = slope of the regression line,

p_i = probability of ith event for a binary variable,

X_1= predictor variable.

After estimating the parameters of regression model in Equation 3.7, the real probabilities are calculated using the following formula:

$$p_i = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}} \qquad \textbf{(3.8)}$$

where

β_0 = intercept of the regression line,

β_1 = slope of the regression line,

p_i = probability of ith event for a binary variable,

X_1= predictor variable.

Figure 3.5, below, shows the linear regression of the observed probabilities, Y, on the independent variable X. The problem with ordinary linear regression is falling outside legitimate and meaningful range of 0.0 to 1.0, inclusive when you extend the regression line a few units upward or downward along the X axis. Logistic regression, as shown in Figure 3.6, fits the relationship between X and Y with a special S-shaped curve that is mathematically constrained to remain within the range of 0.0 to 1.0 on the Y axis.

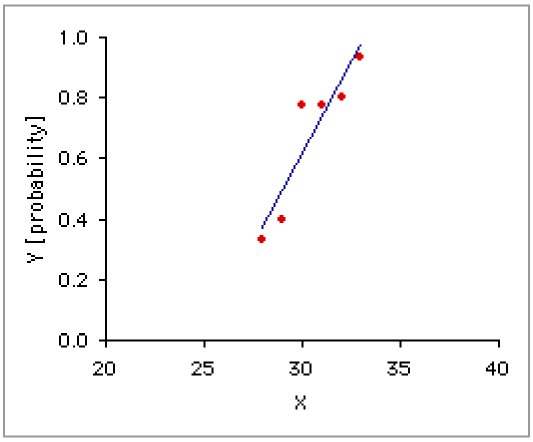


Figure 3.5: Ordinary Linear Regression

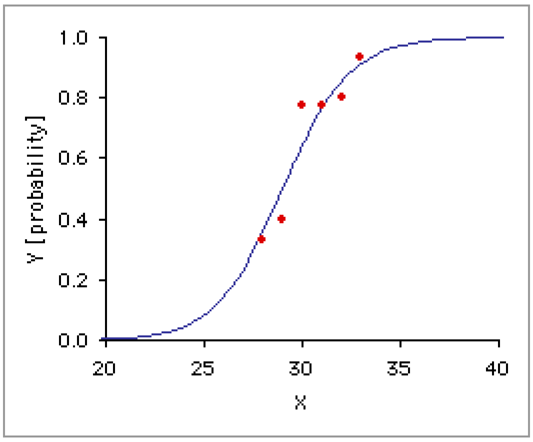


Figure 3.6: Logistic Regression

3.4 Measure of Performance

At the end of model building phase, there are several models produced by different algorithms and ready for the assessment. The performance of each algorithm is measured by a response curve, also called gains chart. To make a response curve, customers are ranked by the predicted probability of the response in descending order and divided into deciles. Within each decile, actual percentage of the responders is calculated using the validation data set. Afterwards, the deciles are plotted on the horizontal axis and the actual percentage of responders in each decile is put on vertical axis. Cumulative response curve has the cumulative percentage of the respondents on the vertical axis.

If the performance of the model is good, the proportion of actual responders will be relatively high in the first decile. Figure 3.7 illustrates both the noncumulative and cumulative response curves. The line passing through both curves is the baseline that reveals the percentage of responders if a random sample of the customers is taken in each decile. In Figure 3.7, the ratio of actual responders over the population in the first decile is 70 per cent. The ratio of actual responders is 20 per cent if a random sample of the same size as in first decile is taken.

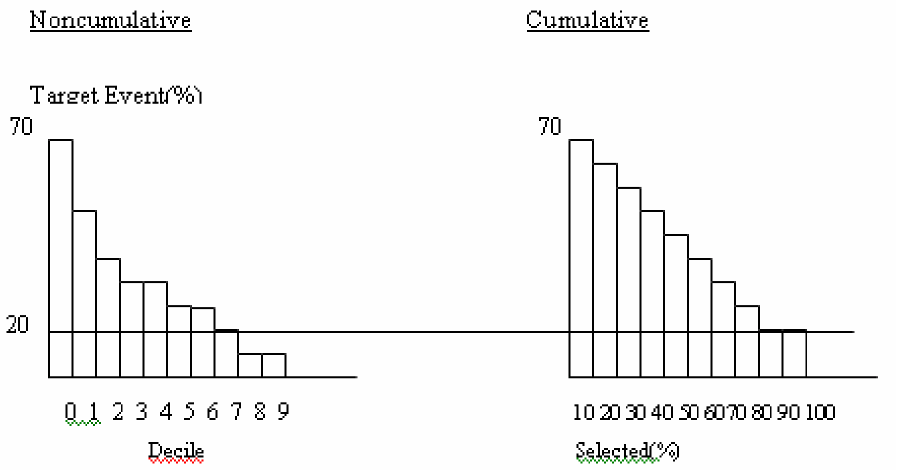


Figure 3.7: Noncumulative and cumulative response curves

There are two more curves used to depict the performance of the models. Lift curves represent similar information as response curves but on a different scale. This curve is obtained by dividing the response rate in each decile by the population response rate. The lift chart plots relative improvement over baseline.

$$LiftValue = \frac{Response \quad rate \quad in \; each \quad decile}{Response \quad rate \quad in \;\; population} \qquad \textbf{(3.9)}$$

Response curve provides information about the percentage of responders in each decile. In captured response curves, the percentage of the responders in a decile over the total number of responders is investigated. Figure 3.8 includes two curves; one of them stands for the lift curve and the other one for the captured response curve. The captured response for first decile is 35 per cent which means if 10 per cent of the customers are reached, 35 per cent of responders are obtained.

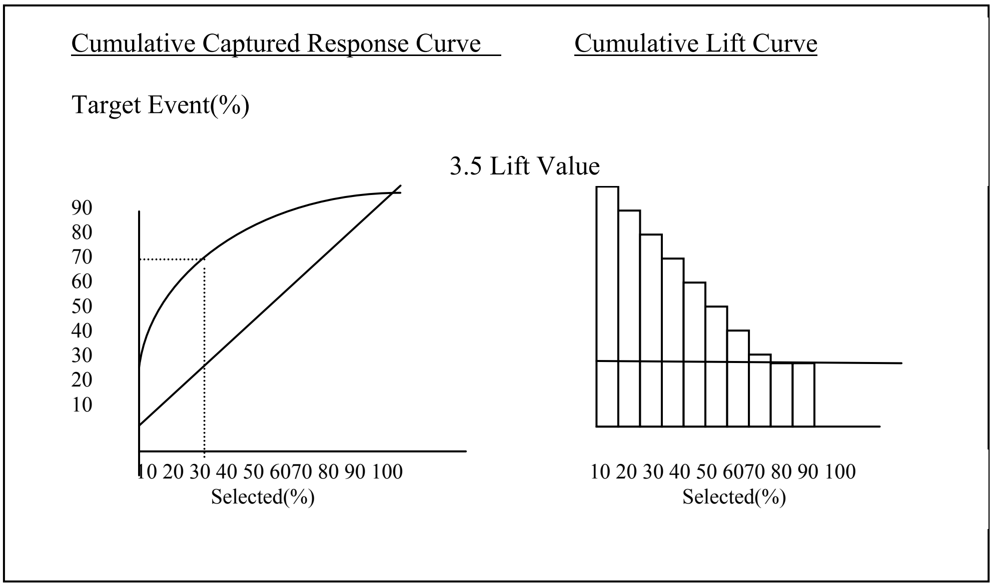


Figure 3.8: Cumulative captured response curve and lift curve

When the lift curve is close to the baseline, it means that model is not good. The more is the lift value the better is the model. The model two represented in Figure 3.9 is better than the model one. If 30 per cent of the customers are contacted, 85 per cent of the actual responders are obtained in model 2, and 55 per cent of the responders in model one.

To select the best model, different criteria are considered also other than the curves described. A model can perform better in all response curves, but its complexity may overcome this advantage. The more complex a model is, the more difficult it is to understand it. The model must be stable and behave similar on the different data sets.

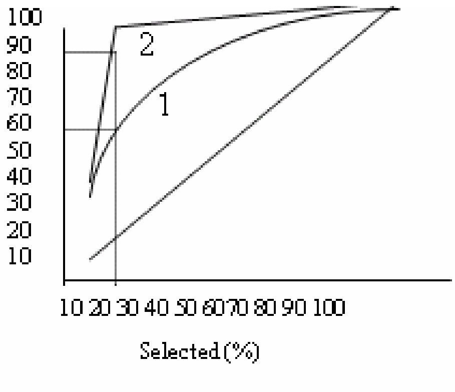


Figure 3.9: Captured response curves for two different models

A two class classification model can produce two classifications leading to four results. Taking “1” to indicate class membership and “0” to indicate class nonmembership, Table 3.2 shows the possibilities. The model can classify as 1 when the actual result is 1, or it can classify as 1 when the actual result is 0. Similarly it can indicate 0 when the actual result is 0, or indicate 0 when the actual result is 1. This table forms the basis of what is called a confusion matrix because it allows an easy indication of where the model is confused (classifies 1 for 0 and 0 for 1) and where it isn't confused.

Table 3.2: Confusion Matrix

Model	**Class1**	**Class0**	**Total**
Is 1	189	35	224
Is 0	27	173	200
Total	216	208	424

Table 3.2 summarizes the performance of the model. The column labeled "Class 1" contains a count of the number of times the model predicted 1 as the class. Similarly column "Class 0" contains count of the number of times the model predicted 0. The row headed "Is 1" contains counts for the number of instances that actually are 1, and similarly row "Is 0" contains count for the number of instances that actually are 0. At the bottom in row "Total" is the column total, and to the right is the row total.

The interpretation is very straightforward. The 189 at the intersection of "Class 1" and "Is 1" indicates that 189 of the instances that the model predicted to be in Class 1 actually were in that class. The cell at the intersection of "Class 0" and "Is 1" shows that 35 instances were predicted to be 0, but actually were 1. Similarly, the "Is 0" row shows the appropriate counts for instances that actually are 0. Column totals add up to the predicted counts of 0's and 1's, row totals sum the totals for the number of actual instances in each class [33].

According to the above confusion matrix, it is calculated that, 35 plus 27 instances are misclassified. The total number of instances that create error is 62. The error rate of the model is easily calculated as 62 over 424 which is the total number of instances that are put into the model. As a result the error rate of the model equals to 0,15.

3.5 Scoring

Modeling phase is concluded when the best model is chosen. Afterwards, this model is used to score new cases. Scoring is the generation of predicted values for a data set that does not contain a target variable. This is an ongoing process until the model in use is expired because of degradation in performance.

4. CHURN MODELING

This chapter presents the modeling and analysis for predicting the customers who are likely to leave a wireless company in near future, which is often called churn prediction in the telecoms industry, using the data mining methodology explained in previous section.

The market in Turkish telecommunications industry is maturing today and recognizes the importance of proactive customer relationship management, focusing on existing customer care in order to keep valuable customers and to make them more profitable to the company. At this point, churn modeling will give valuable business insights to setup effective Marketing strategies to prevent customers from leaving the company because as the telecoms market becomes more saturated, acquiring the new customer is getting more expensive than retaining the existing customer base.

The process of churn prediction described in this chapter is based on the predictive modeling in data mining method and includes the prediction of churn probability for each customer.

Data mining projects start with choosing a data mining software tool according to the business needs. SAS tools (Enterprise Miner and Enterprise Guide) are used as the data mining software in this study. Enterprise Miner performs functions such as clustering, associations, generating decision trees applying regression and building neural networks. Enterprise Guide supports the total data mining process for creating and analyzing a data mart.

4.1 Problem Definition

Churn prediction is strongly related with the customer retention process in the company. The customer retention process involves three issues:

- Identify which customers are going to leave.

- Determine which customer you want to keep among them.
- Develop retention policy (campaign) to prevent the desirable customers from leaving.

As a foundation of the retention process, churn prediction is a very meaningful component in the company. However, the challenges of the churn prediction is to predict the future customer behavior and take action with the customers based on that prediction.Therefore it is important to build a successful churn prediction model that fits properly into the customer retention process of the company.

The most important steps in maximizing churn prediction capability are filtering of churn types and clear definition of churn.

Churn is generally the action of the customer to leave the company for some reason. According to the churn reasons, we can categorize churn by who initiates the action- the company or customer. Churn due to service quality, competition, moving out of service area and so forth is voluntary churn because action is initiated by the customer. Churn due to not paying the bill, not refilling counters periodically and so forth is involuntary churn because the company decides to terminate customer's service. Due to high involuntary churn rate in this wireless company, this study intends to predict involuntary churn. Predicting involuntary churn is very valuable, since the information can be used to reduce losses.

There are two basic groups of customers for a wireless telecommunications company. Postpaid customers are the ones who pay at the end of every billing period. Prepaid customers pay before making any call. In this study, prepaid customers are considered since proportion of prepaid customers in this wireless company is higher than postpaid customers.

Prepaid customers are also classified into business and individual customers in this wireless company. The scope of this study is specified as the prepaid individual customers since the retention process is different than business customers. For postpaid and business customers, different data mining analysis have to be realized.

If a prepaid customer does not refill within a hundred eighty days after his/her last refill, his/her contract is cancelled by this wireless company. Therefore the company wants to know which prepaid customers will be subject to this involuntary churn at the end of six months.

Every customer must have a SIM card to be able to use a GSM handset, also called equipment. SIM cards have to be purchased with handsets. In data, each SIM card is associated with one handset.

Contract is an agreement between the company and its customers. One contract means one SIM card. Each customer can have more than one contract. The company wants to prevent each of them from churning. Furthermore, data is associated on the contract base rather than the customer base for various important predictor variables like the length of time since a contract was signed. Therefore, contracts will be referenced as customers from now on in the study.

Status of a contract can be active or deactive. Active contracts are the contracts that are in use. Deactive contracts are the cancelled ones voluntarily or involuntarily. Active customers are considered in churn modeling.

Identification of Churn Indicators

After defining churn problem clearly, identification of the data required to address the business issue is the next step.

Here are the types of data that has been analyzed to find out indicators of prepaid churn in the data warehouse of this wireless company:

- Customer information data
- Payment data
- Refill data
- Loyalty data
- Call data
- Customer indices derived from transaction data
- Churn Indicator(Target)

As a result of data analysis, the variables in Table 4.1 have been addressed as indicators for prepaid churn.

Table 4.1: List of variables for churn prediction modeling

	VARIABLE NAME	**DESCRIPTION**
	Customer Info	
1	Contract_ID	Contract Identifier

2	ContractStartDate	Contract Start Date
3	ContractStatusID	Status Identifier of the Contract
4	ContractEndDate	Contract Cancellation Date
5	ContractEndReasID	Contract Cancellation Reason Identifier
6	ContractDuration	Total Months of the Contract since Contract Start Date
7	Num_ActivePhones	Total Number of Active Phones that Customer has
8	Age	Customer Age
9	SegmentID	Customer's Value Segment Identifier
10	Current_Tariff_ID	Customer's Current Tariff Identifier
11	Num_Complaint_LM	Number of Complaint calls done by the Customer to Call Center for last month
	Payment	
12	Last_PaidAmt	Total Paid Amount in USD last month by the Customer
13	Previous_PaidAmt	Total Paid Amount in USD previous month by the Customer
14	VASPaidAmt_LM	Total Paid Amount in USD for Value Added Services(VAS) last month by the Customer
15	VASPaidAmt_PM	Total Paid Amount in USD for Value Added Services(VAS) previous month by the Customer
16	PaidAmtPerMinute_LM	Paid Amount in USD per Minute by the customer last month
17	PaidAmtPerMinute_PM	Paid Amount in USD per Minute by the customer previous month
18	PaidAmtPerMinute_2M	Paid Amount in USD per Minute by the customer 2 months ago
	Refill Behaviour	
19	Avg_RefillDay	Average Refill Day of the Customer for last 6 months
20	Tot_NumofRefill	Total Number of Refills of the Customer for last 6 months
21	Tot_AmtofRefill	Total Amount of Refills of the Customer for last 6 months
22	LastRefillAmt	Customer's Last Refill Amount
23	LastRefillDay_EOM	Last Refill Day compared to end of last month
24	LastRefillDay_CTT	Last Refill Day compared to Today
25	Balance_EOM	Customer's Counter Balance as end of month
	Loyalty	
26	Loyalty_Flag	If the customer is member of Loyalty Program then 1, otherwise 0
27	BonusCntLeft	Customer's Bonus Counter left
28	BonusDayLeft	Number of days left for Bonus Counter Load for the Customer
29	BonusDayPassed	Number of days passed since last Bonus Counter Load for the Customer
	Call Behaviour	
30	OutCallMin_LM	Total Minutes of Outgoing Calls done by the Customer last month
31	IncCallMin_LM	Total Minutes of Incoming Calls to the Customer last month
32	OutSMSNum_LM	Total Number of SMS sent by the Customer last month
33	IncSMSNum_LM	Total Number of SMS received by the Customer last month
34	OutCallMin_PM	Total Minutes of Outgoing Calls done by the Customer previous month
35	IncCallMin_PM	Total Minutes of Incoming Calls to the Customer previous month
36	OutSMSNum_PM	Total Number of SMS sent by the Customer previous month
37	IncSMSNum_PM	Total Number of SMS received by the Customer previous month
38	LastCallDate	Customer's Last Call's Date
	Derived Indices	
39	DiffCmp1Dial_LM	Number of different phone numbers of Competitor1 that Customer dialed last month
40	DiffCmp1Rcvd_LM	Number of different phone numbers that Customer received from Competitor1 last month
41	DiffCmp2Dial_LM	Number of different phone numbers of Competitor2 that Customer dialed last month
42	DiffCmp2Rcvd_LM	Number of different phone numbers that Customer received from Competitor2 last month
43	DiffOnNetDial_LM	Number of different OnNet phone numbers that Customer dialed last month
44	DiffOnNetRcvd_LM	Number of different phone numbers that Customer received from OnNet last month

45	DiffCmp1Dial_PM	Number of different phone numbers of Competitor1 that Customer dialed previous month
46	DiffCmp1Rcvd_PM	Number of different phone numbers that Customer received from Competitor1 previous month
47	DiffCmp2Dial_PM	Number of different phone numbers of Competitor2 that Customer dialed previous month
48	DiffCmp2Rcvd_PM	Number of different phone numbers that Customer received from Competitor2 previous month
49	DiffOnNetDial_PM	Number of different OnNet phone numbers that Customer dialed previous month
50	DiffOnNetRcvd_PM	Number of different phone numbers that Customer received from OnNet previous month
	Target	
51	Churn_Flag	If the customer churned after 6 months then 1, otherwise 0

Determination of Time Window

When sourcing all the data defined in Table 4.1, it is necessary to define the following three items to decide which time frame of customer data and churn information are going to be used in the model.

- Data window: Time frame for input variables that is used for constructing model
- Forecasting window: Time frame for the prediction and used when sourcing the target prediction variable (churn indicator). The churn prediction model is often referred to as “WHO and WHEN” model which means that it tries to answer the questions: who is going to leave the company and when are they going to leave. The forecasting window is the “WHEN” part of churn prediction modeling. In the phase of building model, the forecasting window is the time frame to examine whether the customers left the company or not.
- Time lag: Interval between data window and forecasting window.

In this study, one month for some variables and six months for some other variables as a data window and six months as a forecasting window, as shown in Figure 4.1 is used. Since forecasting window is long enough to setup and execute the proper retention actions for Marketing, there is not a need to define time lag.

In the model building phase, six months of historical data from August to January for customers who are active as of the end of January is used with churn information, whether or not these customers left the company in July. This model can be applied to customers who are active as of the end of February to predict probable churners in August.

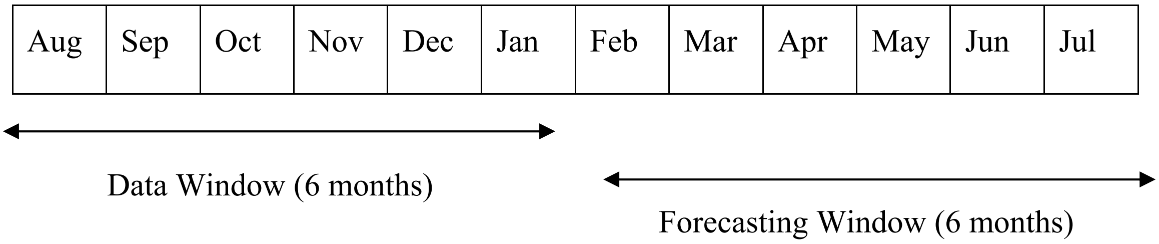


Figure 4.1: Time frames in Churn Modeling

The smaller forecasting window is much better in terms of model performance. Forecasting six months ahead is challenging but this wireless company cancels prepaid contracts if customers don't refill within six months after their last refill.

Target Population

Data used for prepaid churn modeling is modified and coded due to confidentiality. There are 51,337 random contracts in the churn modeling data set.

4.2 Data Preprocessing

The primary source of data for this thesis was a data warehousing system. Data in the DWH is loaded from operational databases such as refill, call details. Final data mining data mart is established using the data in DWH as seen in the Figure 4.2.

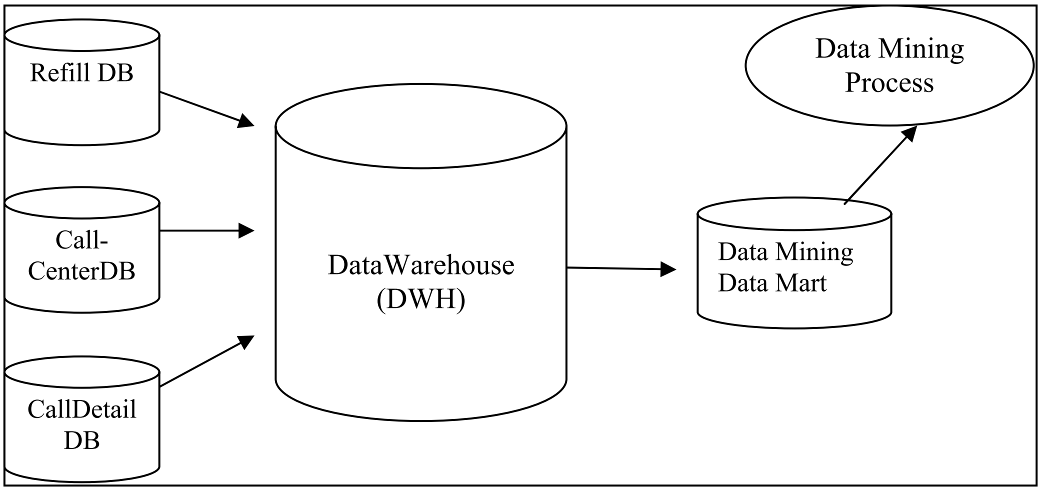


Figure 4.2: Physical view of data mining data mart

Preparing data mining data mart is not as simple as shown in the Figure 4.2. Several steps of data management are done to get the final table. First step is to load necessary data from DWH in normalized form. Normalized tables mean that they are based on an identifier; in this case the primary key is the Contract_ID since the units of action are the contracts. Then transposition is applied to the loaded data tables since some of them are not based on Contract_ID only. Data cleaning is carried out for missing values. Afterwards, all transposed tables are joined to get the final data mining data mart for churn modeling. After this step, data visualization is done to discover correlated variables and they are eliminated. The data mining data mart is ready for churn modeling at the end of data manipulation steps. Data exploration is also done using SEM, and will be explained during modeling sections.

Integer variables are associated with calls, so when there is a missing value for those variables, it means that there is no call. Thus missing values for those variables are replaced with zero. Missing value imputation for interval variables associated with calls is done at the data management part. Data cleaning will also be carried out for other class variables such as nominal and ordinal in modeling part using SEM.

The data mart for the underlying application is a combined table established by joining all tables prepared from DWH based on Contract_ID.

4.3 Modeling

Churn analysis is based on a predictive modeling approach aimed to discover the likelihood of the customers who will leave the company involuntarily. This churn model will predict the churn propensity of the customers whose contracts are still active by the end of month. Modeling data includes customers that are active by the end of January 2005. To eliminate the localized events such as slight fluctuations in economy, refill behavior of customers are observed for six months.

Unlike unsupervised modeling, there is a target variable for churn prediction which indicates if the customers have churned or not after 6 months. In this section, churn prediction model is developed using mostly six months of data starting with August 2004 to the January 2005. The information about a customer has churned or not is taken from July 2005 which is called the target month. This modeling process is handled in February 2006. This month can be considered as the current month since the data prepared for modeling is from the months before February 2006.

Churn target is a binary variable in the data set indicating either a customer has churned or not in July 2005. If the customer has churned, the value of target variable is one for that customer and zero otherwise. The aim of the model is to predict the churners, because of that value of one for churn target is called target event.

Table 4.2 shows some actual statistics about the target variable such as the frequency and the percentage of its each value for modeling data.

Table 4.2: Frequency and percentage of churners in target population.

If the customer churned after 6 months then 1, otherwise 0				
Churn_Flag	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	49820	97.05	49820	97.05
1	1517	2.95	51337	100

The customers can be scored for any target month using the model developed. For instance, it is possible to guess who will churn in August 2005, using the data from September 2004 to February 2005 as input to the model developed. Scoring will occur in March 2005 and churners in August will be predicted.

Necessary Marketing effort such as campaigns and advertisements will be developed between April and June to avoid the churners. The modeling and scoring actions are depicted in Figure 4.3.

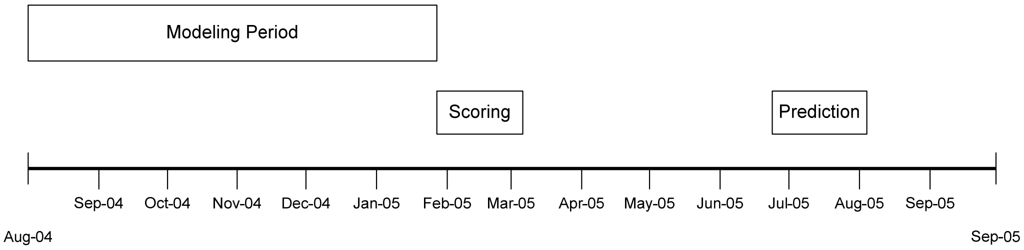


Figure 4.3: Modeling and scoring time frame

Churn prediction model is developed using SEM. In this section, preparation of the model data set for churn modeling will be explained first. Afterwards, the models developed will be analyzed.

The graphical user interface summarizing the employed models is shown in Figure 4.4. The model leading to the best prediction is selected as the final model.

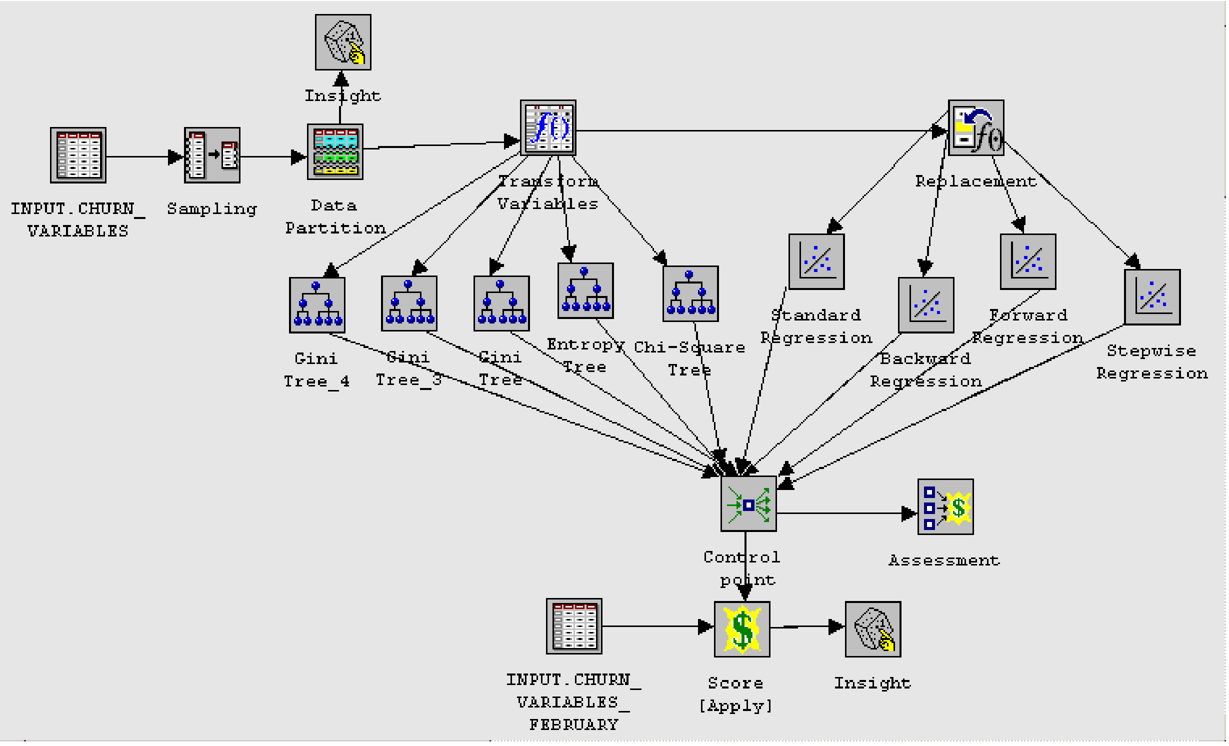


Figure 4.4: Prepaid churn prediction model

4.3.1 Preparation of the Churn Model Data Set

The input data source node specifies a data source and lists the details about the variables. At this node, the dataset INPUT.CHURN_VARIABLES prepared at the end of data management steps is selected as input. All the variables in Table 4.1 in Section 4.1 is explored statistically. According to the distributions of the variables:

- Some of the variables are labeled as redundant and omitted from the model.
- Some of the variables are transformed in to new variables.
- Some of the variables are discretized.

Additionally, the correlations of the variables are analyzed and the ones that are highly correlated are not used in the model. From a business point of view, it simply means that these correlated variables represent the same implication and provide redundant information. From a statistical point of view, using two variables which are highly correlated will artificially emphasize the prediction according to these variables. It is recommended to select one of the variables to represent a group of correlated variable. Correlations between variables are discovered using the Correlations node of SEG and results can be seen in Appendix A.

Rejected Variables

All the rejected variables are shown in the Table 4.3.

Table 4.3: Rejected Variables

Name	Model Role	Measurement	Type	Variable Label
BONUSDAYLEFT	rejected	interval	num	Number of days left for Bonus Counter Load for the Customer
CONTRACTENDDATE	rejected	interval	date	Contract Cancellation Date
CONTRACTENDREASON	rejected	unary	char	Contract Cancellation Reason
CONTRACTSTARTDATE	rejected	interval	date	Contract Start Date
CONTRACTSTATUS	rejected	unary	char	Contract Status
DIFFONNETRCVD_PM	rejected	interval	num	Number of different phone numbers thatCustomer received from OnNet previous month
INCSMSNUM_LM	rejected	interval	num	Total Number of SMS received by the Customer last month
LASTCALLDATE	rejected	interval	date	Customer's Last Call's Date
NUM_COMPLAINT_LM	rejected	ordinal	num	Number of Complaint calls done by the Customer to Call Center for last month
PREVIOUS_PAIDAMT	rejected	interval	num	Total Paid Amount in USD previous month by the Customer
SEGMENTID	rejected	ordinal	num	Customer's Value Segment Identifier
TOT_AMTOFREFILL	rejected	interval	num	Total Amount of Refills of the Customer for last 6 months

Below are the rejection reasons of the variables:

BONUSDAYLEFT: It is in opposite correlation with BonusDayPassed. As BonusDayLeft (Number of days left for Bonus Counter Load for the Customer) descreases, BonusDayPassed (Number of days passed since last Bonus Counter Load) increases.

CONTRACTENDDATE: If a date, time, or date-time format is assigned to a variable, it is eliminated from the model data set since it is just informative variable. ContractEndDate is an example of this. It is just used for analysis purpose.

CONTRACTENDREASON: If there is only one value for a variable, then the measurement level is set to unary and it is eliminated from modeling since it is not a distinguishing data. As shown below, only the customers that churn have contract end reason since the rest of the customers was still active at the end of six months.

Table 4.4: Contract cancellation reason

Contract Cancellation Reason				
ContractEndReason	**Frequency**	**Percent**	**Cumulative**	**Cumulative**
			Frequency	**Percent**
	49820	.	.	.
Involuntary	1517	100	1517	100

CONTRACTSTARTDATE: It is rejected with the same reason as ContractEndDate.

CONTRACTSTATUS: It is rejected with the same reason as ContractEndReason. Since the population in model data set has customers that is active by the end of January, there is just one value for contract status as shown below. This variable is chosen for control purpose.

Table 4.5: Contract status

Contract Status				
ContractStatus	**Frequency**	**Percent**	**Cumulative**	**Cumulative**
			Frequency	**Percent**
Active	51337	100	51337	100

DIFFONNETRCVD_PM: It is highly correlated with DiffOnNetRcvd_LM. Number of different phone numbers that customer received from OnNet for last month is preferred to previous month since it can reflect customer call behaviour currently more.

INCSMSNUM_LM: It is found highly correlated with OutSMSNum_LM. The more SMS customers send, the more they receive.

LASTCALLDATE: It is rejected with the same reason as ContractEndDate

NUM_COMPLAINT_LM: It does not have a proper distribution as seen in the table below and to handle the outliers it is discretized. A new variable named G_NumComplaint_LM is created with grouping missing values under zero, grouping 0 values under 1 and values greater than 1 under group 2. The distribution of the new variable can also be seen in the table below. Only the discretized one(G_NumComplaint_LM) is used in the model in order not to increase the effect of this variable.

Table 4.6: Distribution of Num_Complaint_LM and G_Num_Complaint_LM

Number of Complaint calls done by the Customer to Call Center for last month				
Num_Complaint_LM	**Frequency**	**Percent**	**Cumulative**	**Cumulative**
			Frequency	**Percent**
.	48	.	.	.
0	49746	96.99	49746	96.99
1	357	0.7	50103	97.69
2	243	0.47	50346	98.16
3	320	0.62	50666	98.79
4	289	0.56	50955	99.35
5	287	0.56	51242	99.91
6	41	0.08	51283	99.99
7	6	0.01	51289	100
Grouped values of Num_Complaint_LM variable				
G_Num_Complaint_LM	**Frequency**	**Percent**	**Cumulative**	**Cumulative**
			Frequency	**Percent**
0	48	0.09	48	0.09
1	49746	96.9	49794	96.99
2	1543	3.01	51337	100

PREVIOUS_PAIDAMT: It is found highly correlated with OutCallMin_PM. The more customers call, the more they pay. If two customers buy the same refill card, they pay the same amount but the talking minutes differ based on calls to different operators. Therefore OutCallMin_PM variable is found more distinguishing than Previous_PaidAmt.

SEGMENTID: It is correlated with ContractDuration because the longer contract duration, the more valueable customer is and he/she could be in high segment. Since ContractDuration represents more objective and solid information, it is preferred to SegmentID.

TOT_AMTOFREFILL: It is highly correlated with Tot_NumofRefill. It has many possible values to perform splitting. Therefore Tot_NumofRefill is preferred to be used in modeling.

After data exploration analysis is carried out to discover rejected and correlated variables, sampling is done over 51,337 customers. Since prepaid involuntary churn rate is less than %3, the data set for prediction modeling is also supposed to be constructed with small frequencies of churners like 3%. The mining technique will then very quickly be able to create a good model (with 97% accuracy) by labeling all cases as negative (not churners). This will however not contribute any new information.

As a general rule, oversampling should be done if the outcome we are trying to predict occurs in less than 10 percent of the entire data set. Oversampling involves creating a data set where the relative amount of occurrences is higher than in the original data set. Using a technique that samples by pulling cases out at random will typically be optimal. The resulting data set will be a random stratified sample of the original data set [34].

As a result of trials for best oversampling, stratified sampling method with equal size is employed to achieve a churn rate of 50 per cent to be able to develop a good churn model and a total of 3,034 (1517 churners + 1517 non churners) customers out of 51,337 are used for modeling.

The data is then split into three data sets: 60 percent as a training set, 30 percent as a validation set and the remaining 10 per cent as a test set. Actually, many ratios for data splitting are attempted and the one above yields the best solution. For example, when the data is split into 80 per cent as training data and 20 per cent as validation data, it is resulted with an overfit model.

After data partitioning, some variables are derived in transform variables node. Main competitor of this wireless company is Competitor1. If number of different phone numbers that customer dial or receive from Competitor1 increases, this shows that calling circle of the customer is moving towards to Competitor1 and this could be an indication of churning to Competitor1 because cost of calling Competitor1 is higher than calling numbers in the network of this wireless company. Therefore trend of different phone numbers that customer dial or receive from Competior1 could be a better indicator than just number of different phone numbers. Trend_ DiffCmp1Rcvd and Trend_ DiffCmp1Dial variables indicate if the customer is increasing or decreasing number of different phone numbers that (s)he dials or receives from Competior1 in the last 2 months. This variable is less than one if the usage is increasing and equal to one if it is in a decreasing trend.

$$Trend\_DiffCmp1Rcvd = \frac{DiffCmp1Rcvd\_PM}{\max(DiffCmp1Rcvd\_PM, DiffCmp1Rcvd\_LM)} \qquad \textbf{(4.1)}$$

$$Trend\_OutCallMin = \frac{OutCallMin\_PM}{\max(OutCallMin\_PM, OutCallMin\_LM)} \qquad \textbf{(4.2)}$$

A similar trend variable Trend_DiffOnnetDial is derived to see the impact on churn if the customer is increasing or descreasing his/her number of different phone numbers that (s)he calls in the network of this wireless company in the last 2 months.

$$Trend\_DiffOnnetDial = \frac{DiffOnnetDial\_PM}{\max(DiffOnnetDial\_PM, DiffOnnetDial\_LM)} \qquad \textbf{(4.3)}$$

If the customer is decreasing his outgoing calls in the last 2 months, the customer could be a potential churner. In order to understand this, the following trend variable is created.

$$Trend\_DiffOnnetDial = \frac{DiffOnnetDial\_PM}{\max(DiffOnnetDial\_PM, DiffOnnetDial\_LM)} \qquad \textbf{(4.4)}$$

Another variable called Trend_PaidAmtPerMin is derived to indicate the overall trend of the paid amount per minute in the last three months and calculated as the slope of the line:

$$y = B_0 + B_1 i \qquad \textbf{(4.5)}$$

$$B_1 = \frac{n\sum_{i=1}^{n} iy_i - \sum_{i=1}^{n} i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} i^2 - \left(\sum_{i=1}^{n} i\right)^2} \qquad \textbf{(4.6)}$$

where

i= 1,2,3,4,…,n
n= Number of months
y_i= Total paid amount per minute in month i

If the Equation 4.6 is solved for three months, the following formula is obtained for the variable Trend_PaidAmtPerMin.

$$Trend\_PaidAmtPerMin = \frac{3\,(PaidAmtPerMinute\_LM - PaidAmtPerMinute\_2M)}{16} \qquad \textbf{(4.7)}$$

As a result of these steps, 47 input variables that are used for modeling are tabulated in the Table 4.7.

Table 4.7: Input Variables

Name	Model Role	Measurement	Type	Variable Label
CONTRACT_ID	id	interval	num	Contract Identifier
AGE	input	interval	num	Customer Age
AVG_REFILLDAY	input	interval	num	Average Refill Day of the Customer for last 6 months
BALANCE_EOM	input	interval	num	Customer's Counter Balance as end of month
BONUSCNTLEFT	input	interval	num	Customer's Bonus Counter left
BONUSDAYPASSED	input	interval	num	Number of days passed since last Bonus Counter Load for the Customer
CONTRACTDURATION	input	interval	num	Total Months of the Contract since Contract Start Date
CURRENT_TARIFF_ID	input	Ordinal	num	Customer's Current Tariff Identifier
DIFFCMP1DIAL_LM	input	interval	num	Number of different phone numbers of Competitor1 that Customer dialed last month
DIFFCMP1DIAL_PM	input	interval	num	Number of different phone numbers of Competitor1 that Customer dialed previous month
DIFFCMP1RCVD_LM	input	interval	num	Number of different phone numbers that Customer received from Competitor1 last month
DIFFCMP1RCVD_PM	input	interval	num	Number of different phone numbers that Customer received from Competitor1 previous month
DIFFCMP2DIAL_LM	input	ordinal	num	Number of different phone numbers of Competitor2 that Customer dialed last month

DIFFCMP2DIAL_PM	input	ordinal	num	Number of different phone numbers of Competitor2 that Customer dialed previous month
DIFFCMP2RCVD_LM	input	ordinal	num	Number of different phone numbers that Customer received from Competitor2 last month
DIFFCMP2RCVD_PM	input	ordinal	num	Number of different phone numbers that Customer received from Competitor2 previous month
DIFFONNETDIAL_LM	input	interval	num	Number of different OnNet phone numbers that Customer dialed last month
DIFFONNETDIAL_PM	input	interval	num	Number of different OnNet phone numbers that Customer dialed previous month
DIFFONNETRCVD_LM	input	interval	num	Number of different phone numbers that Customer received from OnNet last month
G_NUM_COMPLAINT_LM	input	ordinal	num	Grouped values of Num_Complaint_LM variable
INCCALLMIN_LM	input	interval	num	Total Minutes of Incoming Calls to the Customer last month
INCCALLMIN_PM	input	interval	num	Total Minutes of Incoming Calls to the Customer previous month
INCSMSNUM_PM	input	interval	num	Total Number of SMS received by the Customer previous month
LASTREFILLAMT	input	ordinal	num	Customer's Last Refill Amount
LASTREFILLDAY_CTT	input	interval	num	Last Refill Day compared to Today
LASTREFILLDAY_EOM	input	interval	num	Last Refill Day compared to end of last month
LAST_PAIDAMT	input	interval	num	Total Paid Amount in USD last month by the Customer
LOYALTY_FLAG	input	binary	num	If the customer is member of Loyalty Program then 1, otherwise 0
NUM_ACTIVEPHONES	input	interval	num	Total Number of Active Phones that Customer has
OUTCALLMIN_LM	input	interval	num	Total Minutes of Outgoing Calls done by the Customer last month
OUTCALLMIN_PM	input	interval	num	Total Minutes of Outgoing Calls done by the Customer previous month
OUTSMSNUM_LM	input	interval	num	Total Number of SMS sent by the Customer last month
OUTSMSNUM_PM	input	interval	num	Total Number of SMS sent by the Customer previous month
PAIDAMTPERMINUTE_2M	input	interval	num	Paid Amount in USD per Minute by the customer 2 months ago
PAIDAMTPERMINUTE_LM	input	interval	num	Paid Amount in USD per Minute by the customer last month
PAIDAMTPERMINUTE_PM	input	interval	num	Paid Amount in USD per Minute by the customer previous month
TOT_NUMOFREFILL	input	interval	num	Total Number of Refills of the Customer for last 6 months
VASPAIDAMT_LM	input	interval	num	Total Paid Amount in USD for VAS services last month by the Customer
VASPAIDAMT_PM	input	interval	num	Total Paid Amount in USD for VAS services previous month by the Customer
TREND_OUTCALLMIN	input	interval	num	Trend of Outgoing Calls
TREND_DIFFONNETDIAL	input	interval	num	Trend of Different OnNet number dialed
TREND_DIFFCMP1RCVD	input	interval	num	Trend of Diff numbers received from CMP1

TREND_DIFFCMP1DIAL	input	interval	num	Trend of Different number of CMP1 dialle
TREND_PAIDAMTPERMIN	input	interval	num	Trend for PaidAmtPerMin for last 3 month
CHURN_FLAG	target	binary	num	If the customer churned after 6 months then 1, otherwise 0

4.3.2 Prepaid Churn Modeling

After data derivation, the data set is ready for modeling. Seven different models are developed. The first three are decision tree models based on Chi-Square, Entropy and Gini Index respectively. The last four models are developed using forward, backward, stepwise and standard regression methods.

Decision Tree Models

Three decision tree models are developed using three different algorithms with Tree node of SEM. Decision trees require one target variable, and at least one predictor variable. The input variables used in the decision tree is the one produced at the previous section as seen in Table 4.7.

An advantage of the decision tree technique over other modeling techniques, like regression, is that it produces a model that represents interpretable rules. Another advantage of the tree is the treatment of missing data. The search for a splitting rule uses the missing values of an input. Because of this property, there is no need to replace missing values.

Before running a decision tree, there are some modeling issues that must be specified related to the splitting criterion and pruning. First, a decision tree algorithm must be selected. **Entropy Reduction, Gini Index** and **Chi-Square** algorithms are used for each of the decision tree models developed. In Figure 4.4, there are three decision tree nodes labeled with its applied tree algorithm. The only difference between them is the algorithm used to construct the tree. Pruning of the trees is made in the same way for all the tree models. The details about those decision tree algorithms are explained in section 3.3.3.

The pruning type applied to the trees is from top-down and bottom-up approach of the pruning methodology. Stopping rules for pruning are set as:

- Minimum number of observations in a leaf is 5.
- Observations required for a split search is 18.

- Maximum Depth of Tree is 6.

Also there is one more parameter, but it is not a stopping rule.

- Maximum Number of Branch from a Node is 2.

Number of branch defines, how many child nodes, a predecessor node can be split into.

Those stopping rules allow top-down pruning that means the pruning is carried out during the construction of the tree. Bottom up pruning works by assessing the performance of the subtrees which are created by cutting of the branches.

The model assessment measure is used to select the best subtree based on the results obtained from the validation data. The model assessment measure applied for the subtrees is the total leaf impurity that calculates the total impurity in each leaf of the subtrees and the one with least total leaf impurity is chosen as the best one. Total leaf impurity is the summation of all Gini Index in each leaf.

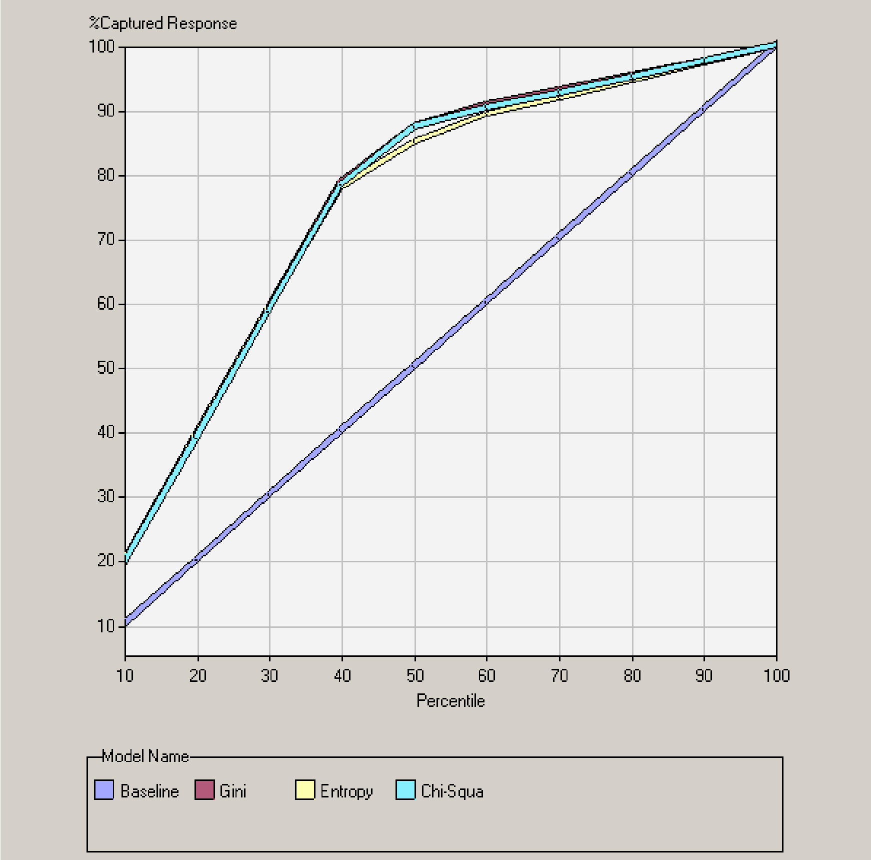


Figure 4.5: Captured response curve of Entropy, Gini and Chi-Sqaure tree.

The results of Chi-Sqaure, Entropy and Gini based decision trees are shown in the captured response curve in Figure 4.5. The detailed explanation about how to use this curve is presented in section 3.4. This curve is interpreted as if the 50 per cent of the ranked customers are contacted; it is likely to get 87 per cent of the churners with Gini Index and Chi-Square, 84 per cent of churners with Entropy Reduction. Although Gini Index and Chi-Square models have the same results for 50 per cent, when 10 per cent of the ranked customers are contacted, 19.73 per cent of churners are captured with Gini Index and 19.32 per cent with Chi-Square. Even though there is a slight difference, Gini based tree performs better than the Chi-Square based tree.

After deciding the best splitting algorithm as Gini Index, the following one is to decide, the number of branches of the tree, because all the tree models developed above are binary trees. Three different Gini based tree models are developed by setting maximum number of branches from a node to 2, 3, and 4 as seen in Figure 4.4.

The model named “Gini Tree” means, Gini Reduction is chosen as splitting criteria and the number of branches is 2. The models named “Gini Tree_3” and “Gini Tree_4” mean, Gini Reduction is chosen as splitting criteria and the number of branches is 3 and 4 respectively. In Figure 4.6 the comparison lift chart of the three models are seen.

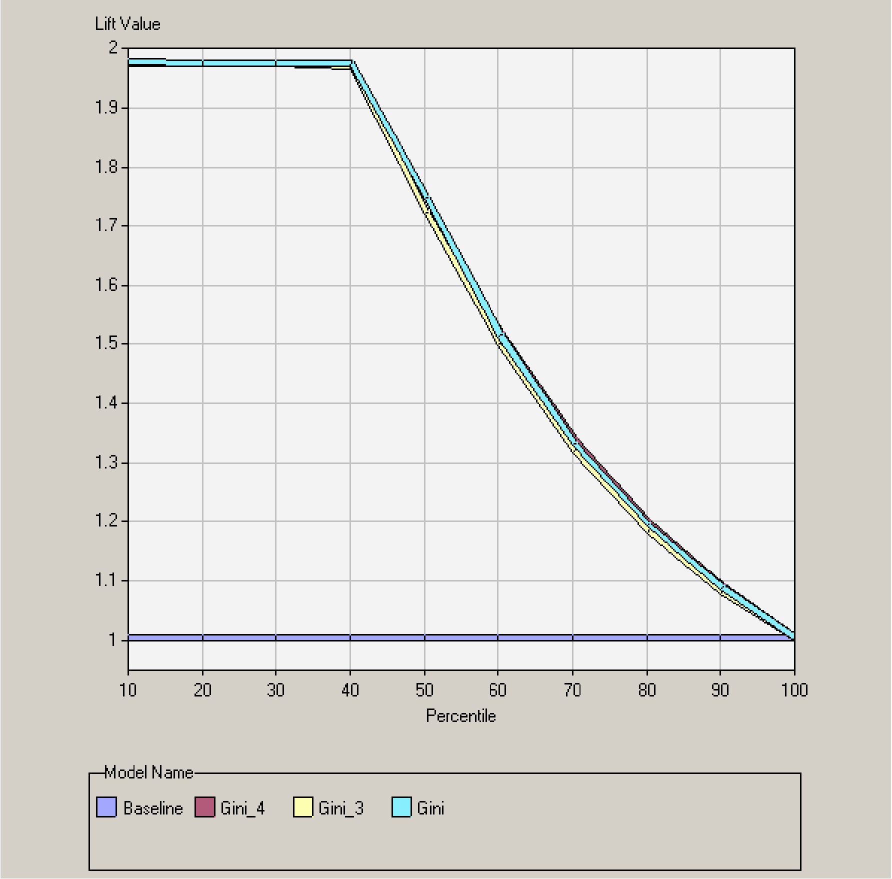


Figure 4.6: The Comparison Lift Chart for Selecting Number of Branches of Tree

Lift values of the three models are very close to each other. "Gini Tree" model with 2 branches has slightly better lift values for each percentile than Gini Tree_3 and Gini Tree_4 models. Lift charts shows that Gini Tree model is 1.97 times better than having no model(random selection) for the first 10 per centile. In section 3.4, the performance measurement for lift charts are explained.

When choosing the best model, interpretability of the model should also be taken into consideration. If Gini Tree_3 or Gini Tree_4 had slightly better results compared to Gini Tree with 2 branches, it would be a still better decision to choose Gini Tree with 2 branches as the best model since it is less complex to map the rules of the tree to the business.

The decision tree model can be seen in Appendix B. According to the Gini tree, the variables that forms the nodes are the critical factors that affect prepaid churn. The result of the pruned Decision Tree modeling has 12 different leaves and each leaf has its own rules. As a result of decision tree model, the critical factors affecting prepaid churn can be found in Table 4.8.

Table 4.8: Critical Factors Affecting Prepaid Churn

Name	Importance	Rules	Variable Label
G_NUM_COMPLAINT_LM	1	2	Grouped values of Num_Complaint_LM variable
INCCALLMIN_LM	0.2418	1	Total Minutes of Incoming Calls to the Customer last month
OUTCALLMIN_LM	0.1761	2	Total Minutes of Outgoing Calls done by the Customer last month
LOYALTY_FLAG	0.131	1	If the customer is member of Loyalty Program then 1, otherwise 0
BALANCE_EOM	0.1236	1	Customer's Counter Balance as end of month
INCCALLMIN_PM	0.1113	1	Total Minutes of Incoming Calls to the Customer previous month
CONTRACTDURATION	0.1061	1	Total Months of the Contract since Contract Start Date
OUTSMSNUM_LM	0.0812	1	Total Number of SMS sent by the Customer last month
LAST_PAIDAMT	0.0747	1	Total Paid Amount in USD last month by the Customer

In Table 4.8 the column named “Importance” discloses the level of importance of the variable for the model. The Importance value is calculated by the model. The column named “Rules” shows how many times the attribute appear in the tree. If the value is greater than 1 then this means that variable is used at different levels of the tree.

Some of the rules are very distinguishing in case of target event. Rules of all leaves can be seen in Appendix C. The number of customers defined by each rule is illustrated as N. If the churn rate is high but the number of customers is very small like node 30 in Figure 4.7, then this rule can be disregarded since it will most probably lead overfitting.

```
IF  Grouped values of Num_Complaint_LM variable EQUALS 0

AND If the customer is member of Loyalty Program then 1, otherwise 0 EQUALS 1

AND 1.0499997139 <= Total Minutes of Outgoing Calls done by the Customer last month

AND 1.4083328247 <= Total Minutes of Incoming Calls to the Customer last month

THEN

 NODE   :     30

 N      :      5

 1      :  100.0%

 0      :    0.0%
```

Figure 4.7: Gini Rule describing overfitting

The rule in Figure 4.8 describes the customers who are more likely to churn than the overall population. The churn rate in target population is nearly 3 per cent but the churn rate in the following rules is more than 90 per cent. After investigating the rule, it is not difficult to conclude that number of a few complaint calls done by the customer to Call Center for last month effect churn rate. Same variable is also observed at the end of regression analysis as the most effective indicator for prepaid churn.

```
IF  Grouped values of Num_Complaint_LM variable EQUALS 2
THEN
 NODE   :      3
 N      :    727
 1      :  98.9%
 0      :   1.1%
```

Figure 4.8: Gini Rule describing the customers who are more likely to churn

Regression Models

Regression node of SEM is used to develop four different logistic regression models with standard, forward, backward and stepwise selection methods.

Regression uses only full cases in the model. This means that any case, or observation, that has a missing value will be excluded from consideration when building the model. This could result in high loss of data and other reasons for imputing missing values include the following:

- Decision trees handle missing values directly, whereas regression does not. It is more appropriate to compare models built on the same set of observations.
- If the missing values are in some way related to each other or to the target variable, the models created without those observations may be biased.
- If missing values are not imputed during the modeling process, observations with missing values can not be scored with the score code built from the models [35].

Therefore missing values should be imputed prior to running a regression model. In the replacement node before regression models, missing interval variables such as Age are replaced with the mean of that variable and missing class variables like DIFFCMP2DIAL_LM are replaced with the most frequent value of that variable.

Since the churn target variable is binary, logistic regression is chosen to predict the churn target. A brief introduction of logistic regression is illustrated in Chapter 3.

In forward regression node, the best one-variable model is selected first. Then the best two variables among those that contain the first selected variable is chosen. This process continues until it reaches the point where no additional variables have a probability value(p-value) less than the specified entry p-value. P-values are also known as significance levels [35].

In backward regression node, it starts with the full model. Next, the variable that is least significant is removed from the model. This process continues until all of the remaining variables a probability value(p-value) less than the specified stay p-value [35].

In stepwise regression node, after each variable is entered into the model, it looks at all the variables already included in the model and deletes any variable that is not significant at the specified level. The process ends when none of the variables outside the model has a p-value less than the speficied entry value and every variable in the model is sigfinicant at the specified stay value [35].

P-value is set to 0.05 for all the models above.

In standard regression node, all canditate effects are included in the final model.

During modeling, different group of variables is tried to find the most effective ones. One of the trials was comparing model performances with and without trend variables. When trend variables are chosen as input, the variables used to form trend variables are not selected as input in order not to overemphasize those variables. For instance, OUTCALLMIN_LM and OUTCALLMIN_PM is not selected as input for modeling when TREND_OUTCALLMIN is selected as input. As a result of this trial, it is found that regression models without trend variables performed better than the models with trend variables.

Figure 4.9 shows the results of four regression models. It is apparent that they perform quite similar. Through the end of the curve, backward regression model performs slightly well than the other models. To be able to choose best model, we need to think about business needs and goals. Business always want to use limited resources in a company more efficiently. When we analyze the captured response curve with this perspective, stepwise regression is a better model than others. When 10 per cent of the ranked customers are contacted; it is likely to get 19.97 per cent of all the churners with Stepwise and 19.75 with Forward regression model. When 50 per cent of the ranked customers are contacted; it is likely to get 88.79 per cent of the churners with Stepwise and 87.47 per cent of churners with Forward regression model. Consequently, contacting less customers for a campaign with a sigificant response rate helps using company resources more efficicently.

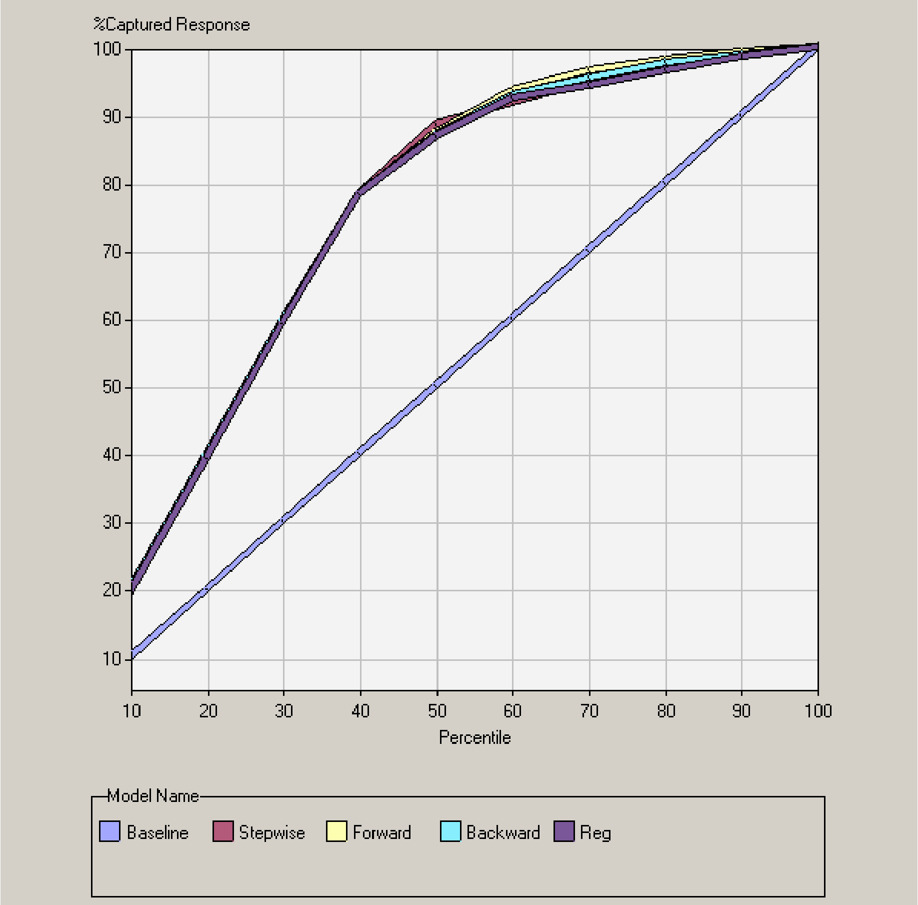


Figure 4.9: Captured response curve of regression models

Table 4.9 shows the maximum likelihood estimates of variables affecting prepaid churn as a result of stepwise regression. T-score is equal to the parameter estimate divided by its standard error.

Table 4.9: Results of Stepwise Regression

Name	**Parameter Estimate**	**Effect T-scores**	**Variable Label**
Intercept:Churn_Flag=1	5.5762892	0.2945765	If the customer churned after 6 months then 1, otherwise 0
G_Num_Complaint_LM 0	5.9640535	0.1575385	Grouped values of Num_Complaint_LM variable with value 0
Loyalty_Flag 0	0.8896592	7.250192	If the customer is member of Loyalty Program then 1, otherwise 0
Tot_NumofRefill	0.085449	6.2176129	Total Number of Refills of the Customer for last 6 months
DiffOnNetDial_LM	-0.1239867	-7.1154541	Number of different OnNet phone numbers that Customer dialed last month

G_Num_Complaint_LM 1	-6.0022226	-0.3170865	Grouped values of Num_Complaint_LM variable with value 1

Binary dummy variables are also created for class type predictors such as G_Num_Complaint_LM 0 and G_Num_Complaint_LM 1. The predictor total points is not an effective factor and can be put out of the regression analysis. The regression model can be formulated as:

$$\begin{aligned} \log it(p) = {} & 5.96405353 G\_Num\_Compla\mathrm{int}\_LM0 + 0.88965917 Loyalty\_Flag0 \\ & + 0.08544902 Tot\_Numof\mathrm{Re}\, fill - 0.12398668 DiffOnNetDial\_LM \\ & - 6.00222262 G\_Num\_Compla\mathrm{int}\_LM1 \end{aligned} \tag{4.8}$$

P is the probability to churn. The function logit is carried out to make the relationship between probabilities and the predictor variables are nonlinear. Transformation, logit, is applied to obtain probabilities from the logistic regression. The details about the logistic regression is explained in Chapter 3.

As it is seen in the results, the customers that are not member of Loyalty Program are more likely to churn than the members. Number of complaint calls to Call Center was an expected factor for prepaid churn and it came out as a distinguishing variable for indicating churn. The more a customer complains to Call Center, the more likely (s)he churns. Another important variable is number of different phone numbers that a customer dialed last month in the network of this wireless company. If this number decreases, it means that the customer is calling competiors numbers more or the customer is having financial difficulties. The last significant variable is total number of refills of the customer for last 6 months which shows the refill behaviour of the customer. If the number of refills is declining by time for a customer, this might indicate that customer has dual simcard and using competitor's line more than this wireless company's gsm line and he might leave for the competitor soon. History is the best predictor of the future and this variable proves it.

4.4 Measure of Performance

There are three decision tree models developed, one with Gini index, one with Chi-Square and the other with Entropy reduction. It is observed that the one with Gini index performs better than others. Gini based tree with 2 branches is selected as the final decision tree. Four regression models; stepwise, forward, backward and standard are developed. Among these models, stepwise regression is chosen as the best regression model.

It is time to illustrate which one of those models is the best and will be selected to score the prepaid customers churn propensity for next months. Gini tree and Stepwise regression model are the candidates of the final model. For this purpose, the response, captured response and lift curves are utilized to compare both models.

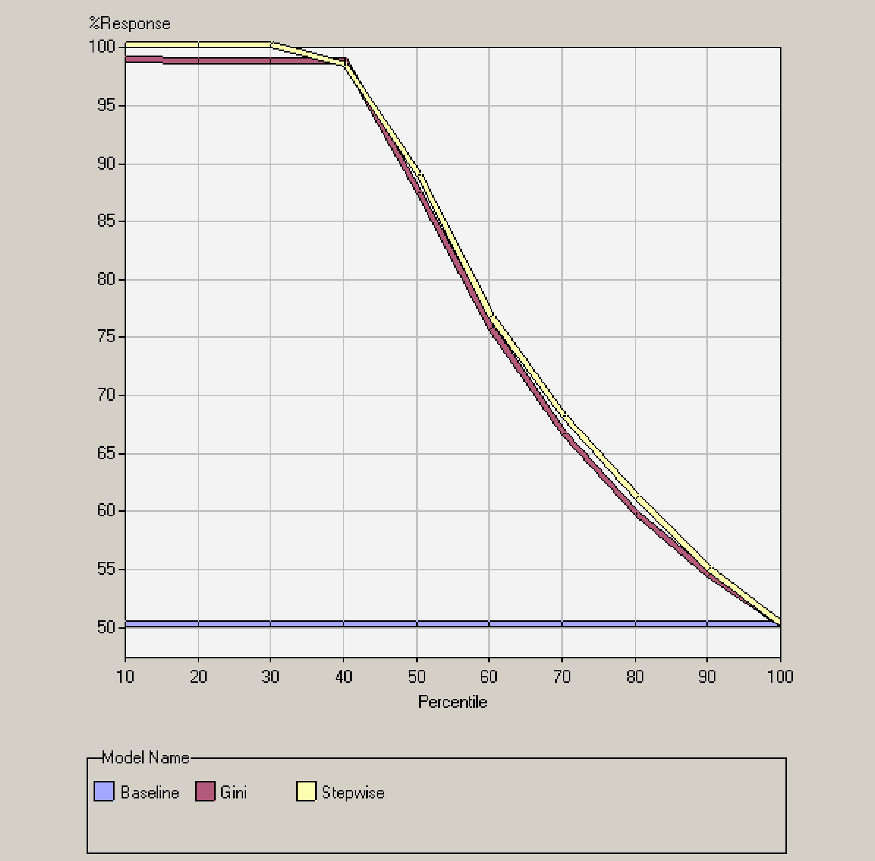


Figure 4.10: Response curve of the models

In Figure 4.10, response curves of the both models are presented and it shows that stepwise regression model performs better than Gini tree. This curve indicates that if 10 per cent of the highly scored customers are got in touch then 100 per cent of them are the churners according to stepwise regression model, 98 per cent of them are the churners according to the Gini tree.

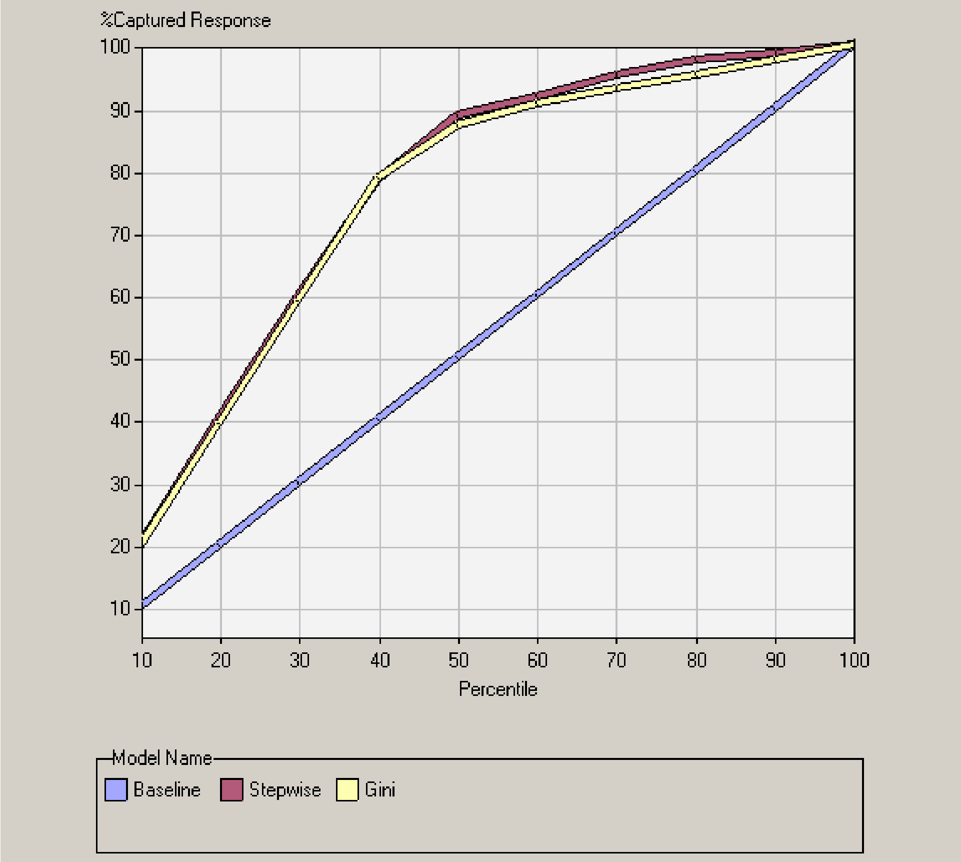


Figure 4.11: Captured response curve of the models

In Figure 4.11, the results of both models are assessed using another graphical method, captured response curves. When 10 per cent of the customers are contacted, it is possible to catch 19.97 per cent of all the churners according to the stepwise regression model, 19.73 per cent according to Gini tree and difference between regression and tree model gets bigger towards 50 per cent. When 50 per cent of the customers are contacted, it is possible to catch 88.79 per cent of all the churners according to the stepwise regression and 87.32 per cent according to Gini tree model.

The last graphical method, lift curves, is illustrated in Figure 4.12 and used to compare the models. As explained in Chapter 3, lift curve represents the ratio of churners in each decile to the churners in whole population. This curve indicates that stepwise regression model performs 1.99 times and Gini tree performs 1.97 times better than the random up to nearly percentile 40.

All of the measureLift curve also proves as the other curves that regression model two performs better than the others and selected as the final model.

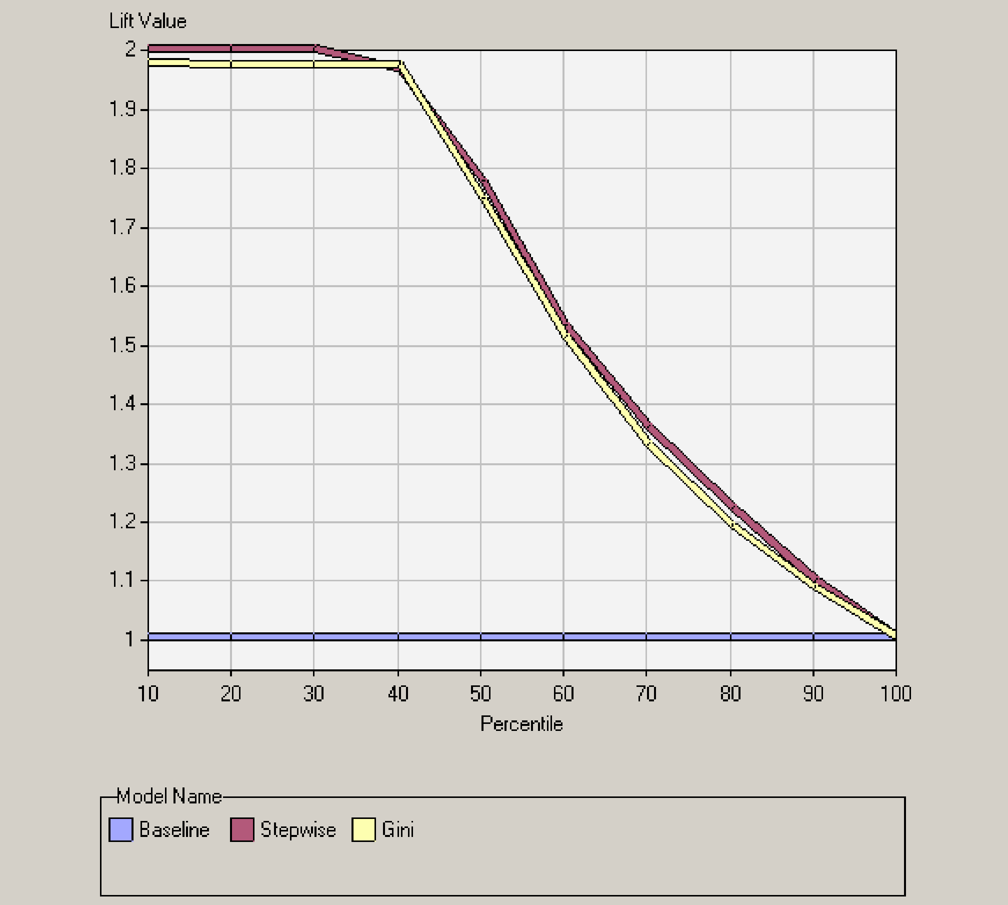


Figure 4.12: Lift curve of the models

The last assesment tool to see the quality of predictive models is the confusion matrix. Table 4.10 is the cross tabulation of the actual and predicted classifications.

Table 4.10: Confusion Matrix of Stepwise Regression Model

Real		Estimated		
		0	1	Total
0	Frequency	874	36	910
	Percent	48.05	1.98	50.03
	Row Percent	**96.04**	3.96	
	Column Percent	85.85	4.49	
1	Frequency	144	765	909
	Percent	7.92	42.06	49.97
	Row Percent	15.84	**84.16**	
	Column Percent	14.15	95.51	
	Total	1018	801	1819
		55.96	44.04	100

Totally 1819 records are used for training the stepwise regression model as a result of sampling and data partitioning. Here is the interpreation of the model results:

- 765 customers (out of 1819) were correctly predicted as churners.
- 42.06% of all the customers were correctly predicted as churners.
- Of all the actual churners (909), 84.16% were predicted as churners.
- Of all the predicted churners (801), 95.51% were actual churners.

The model predicted 1018 customers as non-churnes but 144 of them were churners which means wrongly predicted and the model predicted 801 customers as churners but 36 of them were not churners. Based on this, the total error ratio of the model is 9% [(144 +36) / 1819 = 0,09]. That means the results of the model are right by 91%.

Table 4.11 is the interpreation of the decision tree model based on the confusion matrix.

Table 4.11: Confusion Matrix of Gini Tree Model

Real		**Estimated**		
		0	**1**	
0	Frequency	880	30	910
	Percent	48	2	50.03
	Row Percent	**97**	3	
	Column Percent	89	4	
1	Frequency	105	804	909
	Percent	6	44	49.97
	Row Percent	12	**88**	
	Column Percent	11	96	
		985	834	1819
		54	46	100

- 804 customers (out of 1819) were correctly predicted as churners.
- 44% of all the customers were correctly predicted as churners.
- Of all the actual churners (909), 88% were predicted as churners.
- Of all the predicted churners (834), 96% were actual churners.

The model predicted 985 customers as non-churnes but 105 of them were churners which means wrongly predicted and the model predicted 834 customers as churners but 30 of them were not churners. Based on this, the total error ratio of the model is 7% [(144 +36) / 1819 = 0,09]. That means the results of the model are right by 93%.

When we evaluate the results above, stepwise regression is a better model than gini tree model based on the three out of four measure of performance criteria so stepwise regression is selected as the final model.

4.5 Scoring

Modeling task is not completed once a model is determined. The model must be practically applied to new cases and this process is called scoring. The Score node in Figure 4.4 combines data manipulation and modeling steps into a score code that can be executed to score new data sets.

Modeling was done on January 2005 data and scoring code of regression model in Appendix D is used to score February 2005 data set. This new data set has the same variables as in January 2005 and contains August 2005 churners. A churn propability score between 0 and 1 (1 being the highest churn probability) is assigned to each prepaid customer. By this way, the regression model is validated with respect to time by comparing the performance of it in the separate unused data set. The result is in Table 4.12.

Table 4.12: Confusion Matrix of a Validation Data Set

Real		Estimated		
		0	1	
0	Frequency	9090	610	9700
	Percent	90.90	6.10	97
	Row Percent	**93.71**	6.29	
	Column Percent	99.29	72.19	
1	Frequency	65	235	300
	Percent	0.65	2.35	3
	Row Percent	21.67	**78.33**	
	Column Percent	0.71	27.81	
		9155	845	10000
		91.55	8.45	100

There are 10000 customers in the February data set and churn rate is 3% which is very similar to actual churn rate(2.95%) of January 2005.

The total correct classified observation rate is 86.02% [(93.71 + 78.33) / 2] and the performance of the modeling data was 90.1%. Thus, the performance loss (90.1 – 86.02/90.1) is nearly 5%. The model can be accepted as quite stable. We should note that since equal stratified sampling was used, the cut-off score was 0.5 for all models.

5. CONCLUSION

In this study, a challenging objective is aimed to be implemetend which is prediction of the prepaid customers who are likely to leave a Turkish wireless telecommunications company. Since prepaid customers do not receive a bill, the wireless companies do not have updated reliable information such as demographic, geographical or financial information about their prepaid customers as much as their postpaid customers. Wireless telecommunications companies try to understand their prepaid customers' behaviour mostly based on their transaction data which can be call or refill records.

According to literature, churn prediction was investigated for postpaid individual customers in telecommunications. In that study, different factors such as bill city, handset had been found as important indicators of churn since postpaid customers have completely different characteristics and behavior from prepaid customers [18].

If a prepaid customer does not refill his/her card within 6 months after his/her last refill, wireless company subject to this study cancels the contract of the prepaid customer. The company wants to know which prepaid contracts are likely to be cancelled in 6 months so that effective retention strategies can be implemented in advance to keep these customers.

In order to propose a solution for this problem, the methodology of data mining is explained first and then a churn prediction model is developed to produce a score for each prepaid individual customer who is likely leave the company in 6 months due to involuntary reasons by using steps in data mining methodology. Supervised modeling techniques, decision trees and logistic regression, are used and model developed through logistic regression has been chosen as the best model.

Four main variables have been found that indicate prepaid churn at the end of modeling. The most important of them is the number of complaint calls to call center. In order take pro-active action at call center, churn scores can be integrated to call center systems so that agents can have more information about customers including churn score when they are dealing with customers. In addition to call center systems, churn scores can be deployed to any contact channel for customers such as dealers.

Other important variable of churn is being a member of a loyalty program. Customers that are not member of loyalty program have high probability of churn. Therefore these customers can be offered to join loyalty program when they contact call center or any other contact channel explaining the monetary benefits of loyalty programs to customer because prepaid customers are more price sensitive than postpaid customers. Besides this reactive action, the best way offering to join loyalty program is through sending SMS for prepaid customers.

Another indicator for churn is number of different OnNet phone numbers that customer dialed last month. As seen in Table 5.1, average of different phone numbers that churners dialled last month is nearly half of non-churners.

Table 5.1: Average of different OnNet phone numbers that Customer dialed last month

Number of different OnNet phone numbers that Customer dialed last month					
Churn_Flag	Mean	Minimum	Maximum	Number of Records	Number of Missing Records
0	10.134	0	92	49423	397
1	6.22388	0	66	1407	110

There could be many reasons for decreasing the number of different phone numbers that customer dialled in the network of this wireless company. The customer might be having financial difficulties or the customer might be a visitor in Turkey for summer and decreasing his number of calls towards the end of his stay. While offering free minutes to customer having financial difficulties would prevent him from churning, it will not stop a visitor from churning. It would be just revenue loss for the wireless company to offer free minutes to customers that will churn in any condition. Therefore Marketing should analyze other characteristics of these customers further to understand why prepaid customer is calling different OnNet numbers less in last month and treat the customer with right offer.

The last significant variable that affects churn is the total number of refills of the customer for last 6 months. Since this variable reflects customer refill behaviour, other variables can be assessed with this variable to understand the churners value to the organization and get more insight about factors that are influential in making their decisions. If the number of different OnNet phone numbers that customer dialled last month has a sharp decline but there is not a sharp decrease in number of refills for last six months, then customer might have been abroad for last month and called less people. As stated above, business had better assess churn indicators with other important descriptive information such as contract age or value segment to establish a proper retention offer set for each different group of churners.

Table 5.2: Ratio of Churners Grouped by Value Segment

Customer's Value Segment Identifier						
	Churn_Flag=0		Churn_Flag=1			
SegmentID	Frequency	Percent	Frequency	Percent	Total Frequency	Ratio of Churners
1	12	0.12	1	0.33	13	0.08
2	200	2.06	10	3.33	210	0.05
3	293	3.02	11	3.67	304	0.04
4	333	3.43	14	4.67	347	0.04
5	2558	26.37	66	22	2624	0.03
6	4332	44.66	125	41.67	4457	0.03
7	1972	20.33	73	24.33	2045	0.04

When we assessed the churn distribution by value segment for February data, 41.67% percent of total churners are in segment 6 which is a low-value segment for this wireless company in Table 5.2. Looking at the ratio of churners to total number of customers in each segment points out that segment 1 has a higher churn rate than other segments. Segment 1 has high-value prepaid customers. Hence, this wireless company had better focus on churners in Segment 1 with special offers in a timely manner when a list of prepaid churners has been given to Marketing as a result of churn model because these customers are more profitable than other segments.

Companies do not have unlimited resources. If there is a capacity for each Marketing campaign in the company, then Marketing would like to have high response rate with a low target customer base for the campaign. At this point, data mining offers results that would derive business benefits such as providing customers that have high propensity for response. In this thesis, list of probable prepaid churners are aimed to be provided to business so that proper retention campaigns can be executed to have a high response rate instead of running campaigns for churners selected randomly from the customer base which will result with a low response rate. Table 5.3 shows the benefit of having a prepaid churn model in the company:

Table 5.3: Comparison of random and model selection

	Random Selection	Selection based on Churn Model
Target Population	400,000	400,000
Number of Churners	1200	2400
Average Revenue ($8 per churner)	$9,600	$19,200

There are 4.000.000 prepaid customers of this wireless company and 12,000 churners every month. Churn rate is nearly 3% as it was in the modeling and verification data sets. When 10% of the randomly selected customers are contacted through a test campaign, it is likely to catch 1200 of all the churners while it is possible to capture 2400 churners when 10% of customers are chosen based on the prepaid churn model developed in this study. The average paid amount of last and previous month for churners is $8. Preventing a prepaid customer from churn results in $8 more revenue for this wireless company. If the company was saving $9600 in revenue with traditional targeting methods, now it can save $19,200 with data mining results.

Prepaid churn results can also be utilized by sales and finance departments in the company. Sales departments need to determine what causes to churn and change their sales strategies. Finance department needs to know how much financial impact it has on the company so that they can budget for annual operations accordingly.

If we summarize the contributions of this study, here are the main ones:

- Preventing revenue loss by planning promotional campaigns and retention strategies in advance to keep valuable customers with the company,
- Saving from subscriber acquisiton cost by keeping the existing customers,
- Utilising limited resources of the companies more efficiently,
- Providing competition advantage,

- Providing guidance to companies or individuals that would like to improve themselves about this topic.

While data mining offers benefits to business, performance of the data mining models degrades by time. As the company starts to execute the retention campaign based on the churn prediction model, some customers would respond and change their behavior from churn to staying. A retention campaign may affect a customer's behavior and also the market environment can be changed, as well as government regulation. A prediction model should be upgraded to adapt these changes and maintain performance.

Even though churn models play an increasingly important role in the telecommunications industry and the prepaid churn model developed in this thesis can be customized in other wireless companies, the development of churn prediction model is not recommended for companies that have low churn rate. Predictive models in such cases have limited applicability to the real life. However, churn rate was found to be three per cent in this study, which is a significant value to establish a churn prediction model.

Having identified a list of probable prepaid churners by using data mining techniques in this study, the future work might be to automate the customer retention model and to improve CRM system for the entire company. The churn score of a customer can be utilized during the derivation of a customer lifetime value for the same customer. The results of data mining models are usually related to each other and the combination of them represents a complementary solution in a CRM system.

REFERENCES

[1] **Flanagan and E.Safdie**, “Building a Succesful CRM Environment”, *Applied Technologies Group, Technology Guide Series*, 1998, http://www.techguide.com

[2] **Laudon, K.C and J.P. Laudon**, *Management Information Systems*, Prentice-Hall Inc, New Jersey, 2000.

[3] **Tanrıkorur, T.,** *Enterprise DSS Architecture: A Hybrid Approach*, http://www.dmreview.com, 1998.

[4] **Han, J., & Kamber, M**, 2001. Data mining: Concepts and techniques, Morgan Kaufmann, SanFrancisco

[5] **Berry, M. J. A. and Linoff, G.S.,** 2000. Mastering Data Mining: The Art and Science of Customer Relationship Management, John Wiley & Sons, New York.

[6] **Peters, D.,** *Churn/CPS Improves Profits for Bouygues Telecom,* http://www.dmreview.com , 1998.

[7] http://www.gsmworld.com/documents/external/mobile2006.pdf, 2006 Mobile Markets Report

[8] **Hükmenoğlu, H. and H. Alperat**, *GSM Sector Review*, EgeYatırım Equity Research, Turkey, 2001

[9] **Bashein, B.J., and Markus, M.L.,** 2000. Data Warehouses, More Than Just Mining, Financial Executives Resources Foundation, USA.

[10] http://www.data-mining-software.com/data_mining_history.htm, 2005.

[11] http://eot.apac.edu.au/modules.php?name=Content&pa=showpage&pid=5 2005.

[12] **Pedreschi, D.,** 2005. Data & WebMining, Pisa KDD Lab, ISTI-CNR & Univ. Pisa, http://www.di.unipi.it/~pedre/lucidiLU/S2-Class.pdf

[13] **Newing, R.,** 1996. Data Mining, *Management Accounting: Magazine for Chartered Management Accountants,* **74**, 34-38

[14] **Chye, K. and Gerry, K.**, 2002. Data mining and customer relations in the banking industry, *Singapore Management Review*, **24**, 1-27

[15] **Brachman, J., Khabaza, T., Kloesgen, W., and Simoudis, E.,** 1996. Mining Business Databases, Communications of the ACM. Vol **39**, no.11, 42-48

[16] **Altay, T.,** 2005. Knowledge Discovery in Databases and Data Mining Techniques: An Applied Study, *Master of Science Thesis*,

Marmarmara University Institute for Graduate Studies in Pure and Applied Sciences

[17] **Biçen, P.,** 2002. Data Mining: An Applied Study for Segmentation and Prediction, *Master Thesis*, Yıldız University Institute of Social Science

[18] **Bayram, E.,** 2001. Customer Segmentation and Churn Modeling In Wireless Communications, *Master of Science Thesis*, Boğaziçi University Institute for Graduate Studies in Science and Engineering.

[19] **Büyükakın, A.,** 2005. Data Mining with Fuzzy Logic, *Master of Science Thesis*, İstanbul Technical University Science Institute.

[20] **Ünal, O.,** 2003. Data mining applications on web usage analysis & user profiling, *Master Thesis*, İstanbul Technical University Institute of Social Science.

[21] **Bölükbaşı, İ.,** 2005. Defining Critical Factors Affecting Student Success: A Data Mining Approach, *Master of Science Thesis*, İstanbul Technical University Institute of Science and Technology.

[22] **Bozdogan, H.,** 2004. Statistical Data Mining and Knowledge Discovery, Chapman-Hall, Florida.

[23] **Chan, C. and Lewis, B.,** 2002. A Basic Premier on Data Mining, Information Systems Implementations, Morgan Kaufmann, San Diego.Management, 56-60

[24] **Chen, Y., Hsu , C., Chou,** S., 2003. Constructing a multi-valued and multi-labeled decision tree, *Expert Systems with Applications*, **25**, 199–209.

[25] **Güvenç, E.,** 2001. Student Performance Assessment In Higher Education Using Data Mining, *Master of Science Thesis*, Boğaziçi University Institute for Graduate Studies in Science and Engineering.

[26] **Berson, A. and Smith, S. J.,** 1997. Data Warehousing, Data Mining and OLAP, McGraw-Hill, New York

[27] **Lim, T., W. L. and Y. Shih,** "A Comparison of Prediction Accuracy, Complexity and Training Time of Thirty Three Old and New Classification Algorithms", *Journal of Machine Learning*, Vol.40, No.3, pp. 203-228, 1998.

[28] **Groth, R.,** *Data Mining: Building Competitive Advantage*, Prentice-Hall Inc, New York, 1999.

[29] **Mansuripur, M.,** *Introduction to Information Theory*, Prentice-Hall Inc, 1987.

[30] **Westphal, C. and Blaxton, T.,** *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*, John Wiley, New York, 1998.

[31] **Ruggieri, S.,** "Efficient C4.5", *To appear in IEEE Transactions on Knowledge and Data Engineering*, http://www.di.unipi.it/~ruggieri/ec45, 2001.

[32] **Johnson, R. and Wichern, D.,** *Applied Multivariate Statistical Analysis*, Prentice-Hall, 1988.

[33] **Pyle, D.,** 2003. Business Modeling and Data Mining, Morgan Kaufmann, San Francisco.

[34] **Baragoin, C., Andersen, C.M., Bayerl, S., Bent, G., Lee, J. and Schommer, C.,** 2001. Mining Your Own Business in Telecoms Using DB2 Intelligent Miner for Data, International Business Machines Corporation, California

[35] **Walsh, S., Potts, W., and Wielenga, D.,** 2002. Applying Data Mining Techniques Using Enterprise Miner, SAS Institute Inc., Cary.

[36] http://www.telsim.com.tr/hakkinda/telsim/tarihce.php

[37]http://www.avea.com.tr/sta/hakkinda/hakkinda/aveahakkinda.shtml?pagemenu=hakkinda.hakkinda

[38] http://www.turkcell.com.tr/index/0,1028,23400,00.html

[39] http://zlab.bu.edu/~zhangxl/Pearson_correlation.htm

[40] **Potts, W.,** 2001. Decision Tree Modeling, SAS Institute Inc., Neville.

APPENDIX A: RESULTS OF CORRELATION ANALYSIS

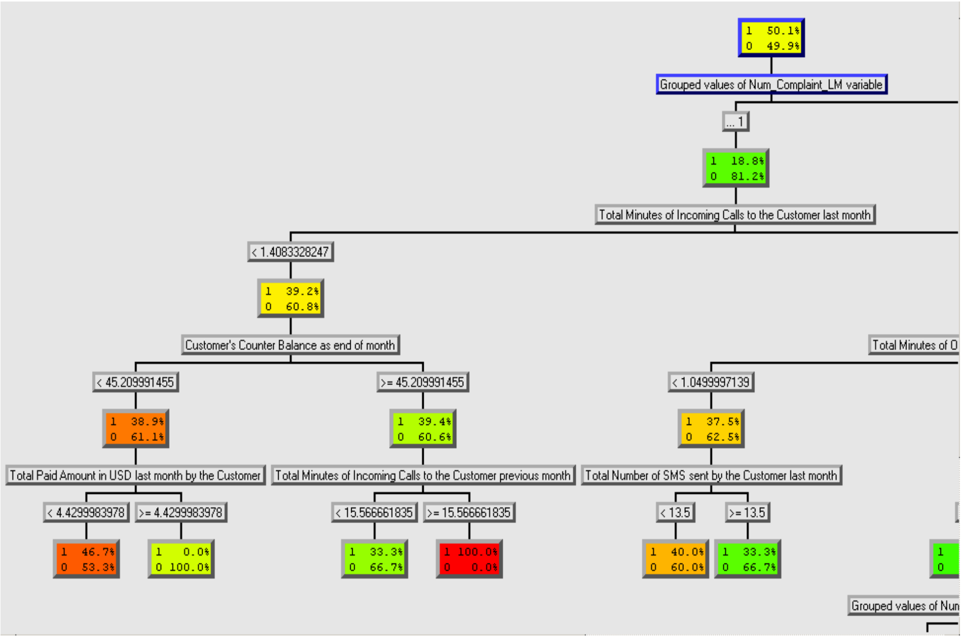
Bold values in Table A.1 shows the most correlated variables as a result of correlation analysis of all variables.

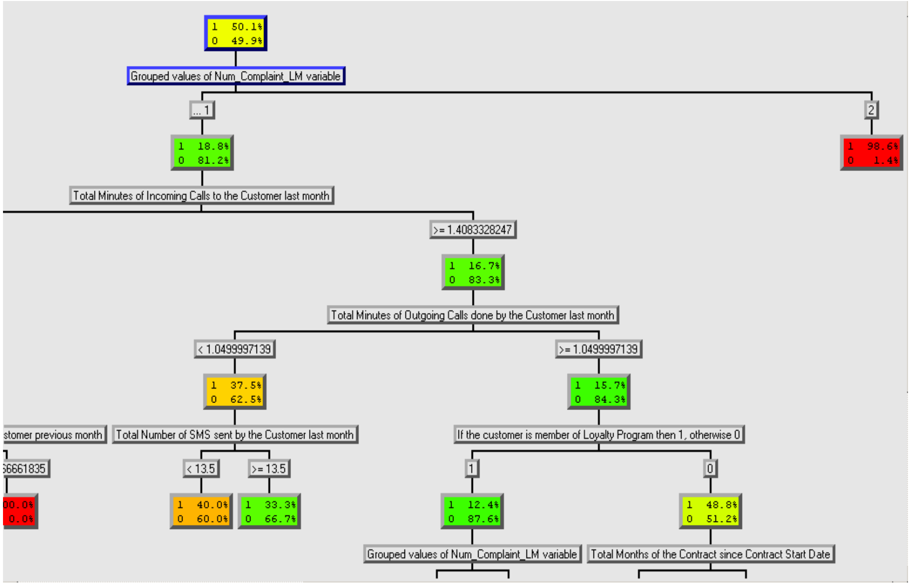
Pearson correlation coefficient measures the strength of a linear relationship between two variables. The correlation coefficient is always between -1 and +1. The closer the correlation is to +/-1, the closer to a perfect linear relationship [39].

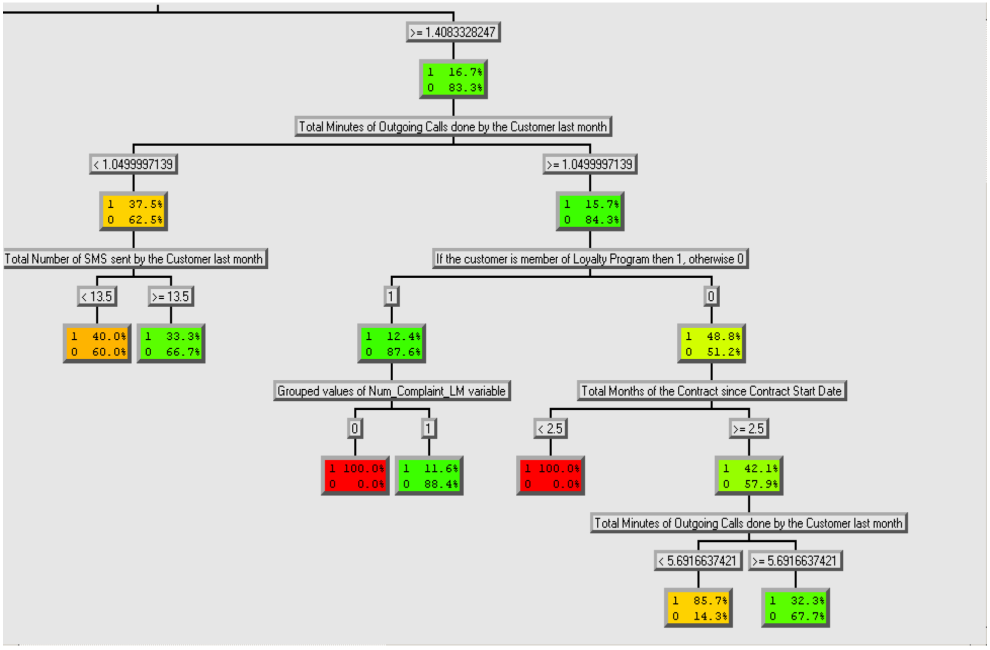
Table A 1: Results of correlation analysis

Pearson Correlation Coefficients					
	Previous_PaidAmt	**Tot_NumofRefill**	**Tot_AmtofRefill**	**BonusCntLeft**	**BonusDayPassed**
SegmentID	-0.33833	-0.40608	-0.45328	-0.37829	0.03689
Last_PaidAmt	0.51105	0.40137	0.5437	0.55506	-0.23597
Previous_PaidAmt	1	**0.65201**	**0.74588**	**0.76809**	-0.22385
Tot_NumofRefill	0.65201	1	**0.87396**	**0.71356**	-0.23304
Tot_AmtofRefill	0.74588	0.87396	1	0.83508	-0.25569
BonusCntLeft	0.76809	0.71356	**0.83508**	1	-0.29806
BonusDayLeft	0.40346	0.32547	0.37167	0.52013	**-0.64121**
DiffOnNetRcvd_PM	0.43186	0.38219	0.39211	0.39119	-0.20482
IncSMSNum_LM	0.20681	0.17849	0.16475	0.14513	-0.0797
OutCallMin_PM	**0.82529**	0.62025	0.72602	0.73234	-0.19251
	DiffOnNetRcvd_LM	**OutCallMin_LM**	**OutSMSNum_LM**	**OutCallMin_PM**	**ContractDuration**
SegmentID	-0.21378	-0.25992	-0.0889	-0.31895	**-0.66205**
Last_PaidAmt	0.42446	**0.77129**	0.33906	0.42676	0.07109
Previous_PaidAmt	0.30105	0.44781	0.19638	**0.82529**	0.04504
Tot_NumofRefill	0.26885	0.37677	0.18217	0.62025	0.10133
Tot_AmtofRefill	0.29857	0.53582	0.17877	**0.72602**	0.11907
BonusCntLeft	0.2716	0.54303	0.16398	0.73234	0.00601
BonusDayLeft	0.23459	0.33198	0.12383	0.34967	-0.08886
DiffOnNetRcvd_PM	**0.8075**	0.31665	0.13751	0.37692	0.13187
IncSMSNum_LM	0.23303	0.05766	**0.82184**	0.07138	0.04379
OutCallMin_PM	0.26141	0.54289	0.06856	1	0.04631

APPENDIX B: RESULTING DECISION TREE WITH GINI INDEX




1 50.1%
0 49.9%
Grouped values of Num_Complaint_LM variable
... 1
2
1 18.8%
0 81.2%
1 98.6%
0 1.4%
Total Minutes of Incoming Calls to the Customer last month
>= 1.4083328247
1 16.7%
0 83.3%
Total Minutes of Outgoing Calls done by the Customer last month
< 1.0499997139
>= 1.0499997139
1 37.5%
0 62.5%
1 15.7%
0 84.3%
stomer previous month
Total Number of SMS sent by the Customer last month
If the customer is member of Loyalty Program then 1, otherwise 0
66661835
< 13.5
>= 13.5
1
0
00.0%
0.0%
1 40.0%
0 60.0%
1 33.3%
0 66.7%
1 12.4%
0 87.6%
1 48.8%
0 51.2%
Grouped values of Num_Complaint_LM variable
Total Months of the Contract since Contract Start Date


>= 1.4083328247
1 16.7%
0 83.3%
Total Minutes of Outgoing Calls done by the Customer last month
< 1.0499997139
>= 1.0499997139
1 37.5%
0 62.5%
1 15.7%
0 84.3%
Total Number of SMS sent by the Customer last month
If the customer is member of Loyalty Program then 1, otherwise 0
< 13.5
>= 13.5
1
0
1 40.0%
0 60.0%
1 33.3%
0 66.7%
1 12.4%
0 87.6%
1 48.8%
0 51.2%
Grouped values of Num_Complaint_LM variable
Total Months of the Contract since Contract Start Date
0
1
< 2.5
>= 2.5
1 100.0%
0 0.0%
1 11.6%
0 88.4%
1 100.0%
0 0.0%
1 42.1%
0 57.9%
Total Minutes of Outgoing Calls done by the Customer last month
< 5.6916637421
>= 5.6916637421
1 85.7%
0 14.3%
1 32.3%
0 67.7%

APPENDIX C: RULES OF EACH NODE IN DECISION TREE

```
IF  Grouped values of Num_Complaint_LM variable EQUALS 2
THEN
 NODE   :     3
 N      :    727
 1      :  98.9%
 0      :   1.1%

IF  Total Paid Amount in USD last month by the Customer < 4.4299983978
AND Customer's Counter Balance as end of month < 45.209991455
AND Total Minutes of Incoming Calls to the Customer last month < 1.4083328247
AND Grouped values of Num_Complaint_LM variable IS ONE OF: 0 1
THEN
 NODE   :     14
 N      :     39
 1      :  87.2%
 0      :  12.8%

IF  4.4299983978 <= Total Paid Amount in USD last month by the Customer
AND Customer's Counter Balance as end of month < 45.209991455
AND Total Minutes of Incoming Calls to the Customer last month < 1.4083328247
AND Grouped values of Num_Complaint_LM variable IS ONE OF: 0 1
THEN
 NODE   :     15
 N      :      8
 1      :  37.5%
 0      :  62.5%

IF  Total Minutes of Incoming Calls to the Customer previous month
   < 15.566661835
AND 45.209991455 <= Customer's Counter Balance as end of month
AND Total Minutes of Incoming Calls to the Customer last month < 1.4083328247
AND Grouped values of Num_Complaint_LM variable IS ONE OF: 0 1
THEN
 NODE   :     16
 N      :     51
 1      :  27.5%
 0      :  72.5%

IF  15.566661835 <= Total Minutes of Incoming Calls to the Customer previous
       month
AND 45.209991455 <= Customer's Counter Balance as end of month
AND Total Minutes of Incoming Calls to the Customer last month < 1.4083328247
AND Grouped values of Num_Complaint_LM variable IS ONE OF: 0 1
THEN
 NODE   :     17
 N      :      8
 1      :  100.0%
 0      :    0.0%

IF  Total Number of SMS sent by the Customer last month <        13.5
AND Total Minutes of Outgoing Calls done by the Customer last month
   < 1.0499997139
AND 1.4083328247 <= Total Minutes of Incoming Calls to the Customer last month
AND Grouped values of Num_Complaint_LM variable IS ONE OF: 0 1
THEN
 NODE   :     18
 N      :     31
 1      :   67.7%
 0      :   32.3%

IF         13.5 <= Total Number of SMS sent by the Customer last month
AND Total Minutes of Outgoing Calls done by the Customer last month
   < 1.0499997139
AND 1.4083328247 <= Total Minutes of Incoming Calls to the Customer last month
AND Grouped values of Num_Complaint_LM variable IS ONE OF: 0 1
THEN
 NODE   :     19
 N      :      8
 1      :   12.5%
 0      :   87.5%

IF  Grouped values of Num_Complaint_LM variable EQUALS 0
AND If the customer is member of Loyalty Program then 1, otherwise 0 EQUALS
  1
AND 1.0499997139 <= Total Minutes of Outgoing Calls done by the Customer last
      month
AND 1.4083328247 <= Total Minutes of Incoming Calls to the Customer last month
THEN
 NODE   :     30
 N      :      5
 1      :  100.0%
 0      :    0.0%

IF  Grouped values of Num_Complaint_LM variable EQUALS 1
AND If the customer is member of Loyalty Program then 1, otherwise 0 EQUALS
  1
AND 1.0499997139 <= Total Minutes of Outgoing Calls done by the Customer last
      month
AND 1.4083328247 <= Total Minutes of Incoming Calls to the Customer last month
THEN
 NODE   :     31
 N      :    885
 1      :   9.2%
 0      :  90.8%

IF  Total Months of the Contract since Contract Start Date <         2.5
AND If the customer is member of Loyalty Program then 1, otherwise 0 EQUALS
  0
AND 1.0499997139 <= Total Minutes of Outgoing Calls done by the Customer last
      month
AND 1.4083328247 <= Total Minutes of Incoming Calls to the Customer last month
AND Grouped values of Num_Complaint_LM variable IS ONE OF: 0 1
THEN
 NODE   :     32
 N      :      8
 1      :  100.0%
 0      :    0.0%

IF  1.0499997139 <= Total Minutes of Outgoing Calls done by the Customer last
      month < 5.6916637421
AND          2.5 <= Total Months of the Contract since Contract Start Date
AND If the customer is member of Loyalty Program then 1, otherwise 0 EQUALS
  0
AND 1.4083328247 <= Total Minutes of Incoming Calls to the Customer last month
AND Grouped values of Num_Complaint_LM variable IS ONE OF: 0 1
THEN
 NODE   :     44
 N      :     16
 1      :  56.3%
 0      :  43.8%

IF  5.6916637421 <= Total Minutes of Outgoing Calls done by the Customer last
      month
AND          2.5 <= Total Months of the Contract since Contract Start Date
AND If the customer is member of Loyalty Program then 1, otherwise 0 EQUALS
  0
AND 1.4083328247 <= Total Minutes of Incoming Calls to the Customer last month
AND Grouped values of Num_Complaint_LM variable IS ONE OF: 0 1
THEN
 NODE   :     45
 N      :     33
 1      :  18.2%
 0      :  81.8%
```

APPENDIX D: SCORING CODE OF RESULTING REGRESSION MODEL

```
/*--------------------------------------------------------------*/
 /*  ENTERPRISE MINER: BEGIN SCORE CODE                          */
 /*--------------------------------------------------------------*/
  %macro DMNORLEN; 32 %mend DMNORLEN;

  %macro DMNORMCP(in,out);
  &out=substr(left(&in),1,min(%dmnorlen,length(left(&in))));
  &out=upcase(&out);
  %mend DMNORMCP;

  %macro DMNORMIP(in);
  &in=left(&in);
  &in=substr(&in,1,min(%dmnorlen,length(&in)));
  &in=upcase(&in);
  %mend DMNORMIP;


  DATA &_PREDICT ; SET &_SCORE ;

  *-------------------------------------------------------------*;
  * CODE_CLEAN *                                                *;
  * Code substitution: ARRAY RGDRF->A1111                       *;
  * Code substitution: ARRAY RGDRU->A4582                       *;
  * Code substitution: GOTO  RGDR1->G0265                       *;
  * Code substitution: ARRAY RGDEMA->A37484                     *;
  * Code substitution: GOTO  RGDEEX->G22894                     *;
  * Code substitution: ARRAY RGDEBE->A00419                     *;
  *-------------------------------------------------------------*;

  *-------------------------------------------------------------*;
  *  START_CHUNK 1461942370.5:T32LPA8U                          *;
  *-------------------------------------------------------------*;
  *  END_CHUNK 1461942370.5:T32LPA8U                            *;
  *-------------------------------------------------------------*;

  *-------------------------------------------------------------*;
  *  START_CHUNK 1461942371:T0V_GF0E                            *;
  *-------------------------------------------------------------*;
  *                                                              ;
  *  TOOL : Sampling                                             ;
  *  TYPE : SAMPLE                                               ;
  *  NODE : Sampling [T0V_GF0E]                                  ;
  *                                                              ;
  *-------------------------------------------------------------*;
  *-------------------------------------------------------------*;
  *  END_CHUNK 1461942371:T0V_GF0E                              *;
  *-------------------------------------------------------------*;

  *-------------------------------------------------------------*;
  *  START_CHUNK 1461942371.4:T15IYFEO                          *;
  *-------------------------------------------------------------*;
  *  END_CHUNK 1461942371.4:T15IYFEO                            *;
  *-------------------------------------------------------------*;

  *-------------------------------------------------------------*;
  *  START_CHUNK 1461942371.9:T2U5Z48B                          *;
  *-------------------------------------------------------------*;
  *  END_INLINE                                                 *;
  *-------------------------------------------------------------*;
  *-------------------------------------------------------------*;
  *  START_INLINE                                               *;
  *-------------------------------------------------------------*;
  drop BONUSDAYLEFT;
  drop DIFFONNETRCVD_PM;
  drop INCSMSNUM_LM;
  drop PREVIOUS_PAIDAMT;
  drop TOT_AMTOFREFILL;
  *-------------------------------------------------------------*;
  *  END_CHUNK 1461942371.9:T2U5Z48B                            *;
  *-------------------------------------------------------------*;

  *-------------------------------------------------------------*;
  *  START_CHUNK 1461942380.5:T0L7KXAN                          *;
  *-------------------------------------------------------------*;
  *  END_INLINE                                                 *;
  *-------------------------------------------------------------*;
  *-------------------------------------------------------------*;
  *  START_INLINE                                               *;
  *-------------------------------------------------------------*;
  *-------------------------------------------------------------*;
  *                                                              ;
  *  TOOL : Replacement                                          ;
  *  TYPE : MODIFY                                               ;
  *  NODE : Replacement [T0L7KXAN]                               ;
  *                                                              ;
  *-------------------------------------------------------------*;
  *;
  *MOST FREQUENT VALUE (COUNT);
  *;
  if CURRENT_TARIFF_ID = . then CURRENT_TARIFF_ID = 1;
  if LASTREFILLAMT = . then LASTREFILLAMT = 100;
  if DIFFCMP2DIAL_LM = . then DIFFCMP2DIAL_LM = 0;
  if LOYALTY_FLAG = . then LOYALTY_FLAG = 1;
  if G_NUM_COMPLAINT_LM = . then G_NUM_COMPLAINT_LM = 1;
  if DIFFCMP2DIAL_PM = . then DIFFCMP2DIAL_PM = 0;
  if DIFFCMP2RCVD_LM = . then DIFFCMP2RCVD_LM = 0;
  if DIFFCMP2RCVD_PM = . then DIFFCMP2RCVD_PM = 0;
  *;
  *MEAN-MEDIAN-MIDRANGE AND ROBUST ESTIMATES;
  *;
  if AGE = . then AGE = 33.4257921067259;
  if NUM_ACTIVEPHONES = . then NUM_ACTIVEPHONES = 2.27432655305112;
  if LAST_PAIDAMT = . then LAST_PAIDAMT = 6.62860696999788;
  if AVG_REFILLDAY = . then AVG_REFILLDAY = 46.4725773170411;
  if TOT_NUMOFREFILL = . then TOT_NUMOFREFILL = 6.12204507971413;
  if LASTREFILLDAY_EOM = . then LASTREFILLDAY_EOM =
29.9395272127542;
  if LASTREFILLDAY_CTT = . then LASTREFILLDAY_CTT =
35.9395272127542;
  if BALANCE_EOM = . then BALANCE_EOM = 54.6301172638354;
  if BONUSCNTLEFT = . then BONUSCNTLEFT = 44.7019311502938;
  if BONUSDAYPASSED = . then BONUSDAYPASSED = 32.5298068849706;
  if DIFFCMP1DIAL_LM = . then DIFFCMP1DIAL_LM = 1.02353616532721;
  if DIFFCMP1RCVD_LM = . then DIFFCMP1RCVD_LM = 1.07921928817451;
  if DIFFONNETDIAL_LM = . then DIFFONNETDIAL_LM = 8.40355912743972;
  if DIFFONNETRCVD_LM = . then DIFFONNETRCVD_LM = 9.96842709529276;
  if DIFFCMP1DIAL_PM = . then DIFFCMP1DIAL_PM = 1.10650546919976;
  if DIFFCMP1RCVD_PM = . then DIFFCMP1RCVD_PM = 1.22279792746114;
  if DIFFONNETDIAL_PM = . then DIFFONNETDIAL_PM = 9.55670696603339;
  if OUTCALLMIN_LM = . then OUTCALLMIN_LM = 17.8854348531432;
  if INCCALLMIN_LM = . then INCCALLMIN_LM = 36.7284274968242;
  if OUTSMSNUM_LM = . then OUTSMSNUM_LM = 9.20065970313359;
  if VASPAIDAMT_LM = . then VASPAIDAMT_LM = 0.34200307651445;
  if OUTCALLMIN_PM = . then OUTCALLMIN_PM = 29.4188928471732;
  if INCCALLMIN_PM = . then INCCALLMIN_PM = 54.6403529188334;
  if OUTSMSNUM_PM = . then OUTSMSNUM_PM = 14.690489279824;
  if INCSMSNUM_PM = . then INCSMSNUM_PM = 17.2622319956019;
  if VASPAIDAMT_PM = . then VASPAIDAMT_PM = 0.40447085338861;
  if CONTRACTDURATION = . then CONTRACTDURATION = 23.0786146234194;
  if PAIDAMTPERMINUTE_LM = . then PAIDAMTPERMINUTE_LM =
0.19357028096712;
  if PAIDAMTPERMINUTE_PM = . then PAIDAMTPERMINUTE_PM =
0.17348460925492;
  if PAIDAMTPERMINUTE_2M = . then PAIDAMTPERMINUTE_2M =
0.16368197403496;
  *;
  *Replacement (Class Variables);
  *;
  length _RFormat $200;
  drop _RFormat;
  _RFormat = '';
  _RFormat = put(CURRENT_TARIFF_ID, BEST12.);
  %DMNORMIP(_RFormat);
  _RFormat = put(LASTREFILLAMT, BEST12.);
  %DMNORMIP(_RFormat);
  _RFormat = put(DIFFCMP2DIAL_LM, BEST12.);
  %DMNORMIP(_RFormat);
  _RFormat = put(LOYALTY_FLAG, BEST12.);
  %DMNORMIP(_RFormat);
  _RFormat = put(G_NUM_COMPLAINT_LM, BEST12.);
  %DMNORMIP(_RFormat);
  _RFormat = put(DIFFCMP2DIAL_PM, BEST12.);
  %DMNORMIP(_RFormat);
  _RFormat = put(DIFFCMP2RCVD_LM, BEST12.);
  %DMNORMIP(_RFormat);
  _RFormat = put(DIFFCMP2RCVD_PM, BEST12.);
  %DMNORMIP(_RFormat);
  *;
  *Replacement (Interval Variables);
  *;
  *-------------------------------------------------------------*;
  *  END_CHUNK 1461942380.5:T0L7KXAN                            *;
  *-------------------------------------------------------------*;

  *-------------------------------------------------------------*;
  *  START_CHUNK 1461942391.5:T3ILB1JJ                          *;
  *-------------------------------------------------------------*;
  *                                                              ;
  *  TOOL : Regression                                           ;
  *  TYPE : MODEL                                                ;
  *  NODE : Stepwise Regression [T3ILB1JJ]                       ;
  *                                                              ;
```

```
*-------------------------------------------------------------*;
*                                                              ;
*  MODEL NAME : Stepwise                                       ;
*  DESCRIPTION : Stepwise Regression                           ;
*                                                              ;
*  TARGET : CHURN_FLAG                                         ;
*-------------------------------------------------------------*;
*************************************;
*** begin scoring code for regression;
*************************************;

length _WARN_ $4;
label _WARN_ = 'Warnings' ;

length I_Churn_Flag $ 12;
label I_Churn_Flag = 'Into: Churn_Flag' ;
*** Target Values;
array A1111 [2] $12 _temporary_ ('1' '0' );
label U_Churn_Flag = 'Unnormalized Into: Churn_Flag' ;
*** Unnormalized target values;
ARRAY A4582[2]  _TEMPORARY_ (1 0);

*** Generate dummy variables for Churn_Flag ;
drop _Y ;
label F_Churn_Flag = 'From: Churn_Flag' ;
length F_Churn_Flag $ 12;
F_Churn_Flag = put( Churn_Flag , BEST12. );
%DMNORMIP( F_Churn_Flag )
if missing( Churn_Flag ) then do;
   _Y = .;
end;
else do;
   if F_Churn_Flag = '0'  then do;
      _Y = 1;
   end;
   else if F_Churn_Flag = '1'  then do;
      _Y = 0;
   end;
   else do;
      _Y = .;
   end;
end;

drop _DM_BAD;
_DM_BAD=0;

*** Check DiffOnNetDial_LM for missing values ;
if missing( DiffOnNetDial_LM ) then do;
   substr(_warn_,1,1) = 'M';
   _DM_BAD = 1;
end;

*** Check Tot_NumofRefill for missing values ;
if missing( Tot_NumofRefill ) then do;
   substr(_warn_,1,1) = 'M';
   _DM_BAD = 1;
end;

*** Generate dummy variables for G_Num_Complaint_LM ;
drop _6_0 _6_1 ;
if missing( G_Num_Complaint_LM ) then do;
```

```
     _6_0 = .;
     _6_1 = .;
     substr(_warn_,1,1) = 'M';
     _DM_BAD = 1;
  end;
  else do;
     length _dm12 $ 12; drop _dm12 ;
     _dm12 = put( G_Num_Complaint_LM , BEST12. );
     %DMNORMIP( _dm12 )
     if _dm12 = '1'  then do;
        _6_0 = 0;
        _6_1 = 1;
     end;
     else if _dm12 = '2'  then do;
        _6_0 = -1;
        _6_1 = -1;
     end;
     else if _dm12 = '0'  then do;
        _6_0 = 1;
        _6_1 = 0;
     end;
     else do;
        _6_0 = .;
        _6_1 = .;
        substr(_warn_,2,1) = 'U';
        _DM_BAD = 1;
     end;
  end;

  *** Generate dummy variables for Loyalty_Flag ;
  drop _8_0 ;
  if missing( Loyalty_Flag ) then do;
     _8_0 = .;
     substr(_warn_,1,1) = 'M';
     _DM_BAD = 1;
  end;
  else do;
     length _dm12 $ 12; drop _dm12 ;
     _dm12 = put( Loyalty_Flag , BEST12. );
     %DMNORMIP( _dm12 )
     if _dm12 = '1'  then do;
        _8_0 = -1;
     end;
     else if _dm12 = '0'  then do;
        _8_0 = 1;
     end;
     else do;
        _8_0 = .;
        substr(_warn_,2,1) = 'U';
        _DM_BAD = 1;
     end;
  end;

  *** If missing inputs, use averages;
  if _DM_BAD > 0 then do;
     _P0 = 0.4997251237;
     _P1 = 0.5002748763;
     goto G0265;
  end;

  *** Compute Linear Predictor;
```

```
drop _TEMP;
drop _LP0;
_LP0 = 0;

***  Effect: DiffOnNetDial_LM ;
_TEMP = DiffOnNetDial_LM ;
_LP0 = _LP0 + (   -0.12398668418386 * _TEMP);

***  Effect: Loyalty_Flag ;
_TEMP = 1;
_LP0 = _LP0 + (    0.88965916825934) * _TEMP * _8_0;

***  Effect: G_Num_Complaint_LM ;
_TEMP = 1;
_LP0 = _LP0 + (    5.96405353439198) * _TEMP * _6_0;
_LP0 = _LP0 + (   -6.00222261516505) * _TEMP * _6_1;

***  Effect: Tot_NumofRefill ;
_TEMP = Tot_NumofRefill ;
_LP0 = _LP0 + (    0.08544902233899 * _TEMP);

*** Naive Posterior Probabilities;
drop _MAXP _IY _P0 _P1;
_TEMP =     5.57628917840599 + _LP0;
if (_TEMP < 0) then do;
   _TEMP = exp(_TEMP);
   _P0 = _TEMP / (1 + _TEMP);
end;
else _P0 = 1 / (1 + exp(-_TEMP));
_P1 = 1.0 - _P0;

G0265:

*** Residuals;
if (_Y = .) then do;
   R_Churn_Flag1 = .;
   R_Churn_Flag0 = .;
end;
else do;
    label R_Churn_Flag1 = 'Residual: Churn_Flag=1' ;
    label R_Churn_Flag0 = 'Residual: Churn_Flag=0' ;
   R_Churn_Flag1 = - _P0;
   R_Churn_Flag0 = - _P1;
   select( _Y );
      when (0)  R_Churn_Flag1 = R_Churn_Flag1 + 1;
      when (1)  R_Churn_Flag0 = R_Churn_Flag0 + 1;
   end;
end;


*** Update Posterior Probabilities;

*** Decision Processing;
label D_CHURN_FLAG_ = 'Decision: CHURN_FLAG' ;
label EP_CHURN_FLAG_ = 'Expected Profit: CHURN_FLAG' ;
label BP_CHURN_FLAG_ = 'Best Profit: CHURN_FLAG' ;
label CP_CHURN_FLAG_ = 'Computed Profit: CHURN_FLAG' ;

length D_CHURN_FLAG_ $ 5;

D_CHURN_FLAG_ = ' ';
```

```
  EP_CHURN_FLAG_ = .;
  BP_CHURN_FLAG_ = .;
  CP_CHURN_FLAG_ = .;

  *** Compute Expected Consequences and Choose Decision;
  _decnum = 1; drop _decnum;

  D_CHURN_FLAG_ = '1' ;
  EP_CHURN_FLAG_ = _P0 * 1 + _P1 * 0;

  *** Decision Matrix;
  array A37484 [2,1] _temporary_ (
  /* row 1 */  1
  /* row 2 */  0
  );

  *** Find Index of Target Category;
  drop _tarnum; select( F_Churn_Flag );
     when('1' ) _tarnum = 1;
     when('0' ) _tarnum = 2;
     otherwise _tarnum = 0;
  end;
  if _tarnum <= 0 then goto G22894;

  *** Computed Consequence of Chosen Decision;
  CP_CHURN_FLAG_ = A37484 [_tarnum,_decnum];

  *** Best Possible Consequence of Any Decision without Cost;
  array A00419 [2] _temporary_ ( 1 0);
  BP_CHURN_FLAG_ = A00419 [_tarnum];


  G22894:;

  *** End Decision Processing ;

  *** Posterior Probabilities and Predicted Level;
  label P_Churn_Flag1 = 'Predicted: Churn_Flag=1' ;
  label P_Churn_Flag0 = 'Predicted: Churn_Flag=0' ;
  P_Churn_Flag1 = _P0;
  _MAXP = _P0;
  _IY = 1;
  P_Churn_Flag0 = _P1;
  if (_P1 - _MAXP > 1e-8) then do;
     _MAXP = _P1;
     _IY = 2;
  end;
  I_Churn_Flag = A1111[_IY];
  U_Churn_Flag = A4582[_IY];

  *************************************;
  ***** end scoring code for regression;
  *************************************;
  *-------------------------------------------------------------*;
  *  END_CHUNK 1461942391.5:T3ILB1JJ                            *;
  *-------------------------------------------------------------*;
  RUN  ;
  QUIT ;
 /*--------------------------------------------------------------*/
 /*  ENTERPRISE MINER: END SCORE CODE                            */
 /*--------------------------------------------------------------*/
```

CURRICULUM VITAE

Müge Özmen was born on 11th of June, 1974 in İstanbul. She has graduated from Kaşgarlı Mahmut College in 1992 and then started studying at Mathematical Engineering of İstanbul Technical University.

She has attended Management Engineering master program of Istanbul Technical University in 1997 and studied until 1999. She had to freze her master degree until 2005 due to working abroad and then restarted studying with the law of amnesty issued in 2005 by the government.