FRAUD DETECTION

IN MOBILE COMMUNICATION NETWORKS

USING DATA MINING

BÜLENT KUŞAKSIZOĞLU

SEPTEMBER 2006

FRAUD DETECTION

IN MOBILE COMMUNICATION NETWORKS

USING DATA MINING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL
OF
THE UNIVERSITY of BAHCESEHIR
BY
BÜLENT KUŞAKSIZOĞLU

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF COMPUTER ENGINEERING

SEPTEMBER 2006

Approval of the Graduate School of (Name of the Graduate School)

_____

(Title and Name)

Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science

_____

(Title and Name)

Head of Department

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

_____ _____

(Title and Name) (Title and Name)

Co-Supervisor Supervisor

Examining Committee Members

..... _____
..... _____
..... _____
..... _____
..... _____

ABSTRACT

FRAUD DETECTION
IN MOBILE COMMUNICATION NETWORKS
USING DATA MINING

Kuşaksızoğlu, Bülent

M.S Department of Computer Engineering

Supervisor: Asst. Prof. Dr. Adem Karahoca

September 2006, 67 pages

Fraud is a significant source of lost revenue to the telecom industry. Efficient fraud detection systems and analysis system can save telecom operators a lot of money. Automated fraud detection systems enable operators to respond to fraud by detection, service denial and prosecutions against fraud.

In this study, we examine the call detail records (CDR's), demographic data and payment data of mobile subscribers in order to develop models of normal and fraudulent behavior via data mining techniques. First we have done some Exploratory Data Analysis (EDA) on the data set and discovered that some variables like Account length, Package type, Gender, Type, Total Charged Amount showed important tendency for fraudulent use and then we applied k-means cluster method to cluster the customer, based on their call behaviors. Standard variables with ranked attributes and variables obtained from factor analysis due to some correlated variables were used as two different set of variables.

Finally we performed the data mining techniques – Decision trees, Rule based methods, and Neural Networks- for both training and test sets and then discussed the collected results based on performance measures such as accuracy, sensitivity, specificity, precision and RMSE.

Key words: fraud, mobile communication, data mining, machine learning

ÖZET

VERİ MADENCİLİĞİ YARDIMIYLA MOBİL TELEKOMÜNİKASYON ŞEBEKELERİNDE SAHTEKARLIK TESPİTİ

Kuşaksızoğlu, Bülent

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi: Yrd. Doç. Dr. Adem Karahoca

Eylül 2006, 67 sayfa

Sahtekarlık/kötü niyetli kullanım telekom endüstrisinde kayıp gelir'in önemli bir kaynağıdır. Etkin sahtekarlık keşfetme sistemleri ve analiz sistemleri telekom operatörlerine çok para tasarruf ettirebilir. Otomatik sahtekarlık sistemleri, operatörlere sahtekarlık yapanları keşfetme, servislerini reddetme ve kovuşturma olanağı vermektedir.

Bu çalışmada, veri madenciliği yöntemleriyle normal kullanıcıları, sahtekearlık yapanlardan ayıran bir model geliştirmek için mobil abonelerin konuşma detay kayıtları(CDR's), demografik verileri ve ödeme verileri incelenmiştir. Önce açıklayıcı veri analizi ile veri seti incelenmiş ve abonelik süresi, paket tipi, cinsiyet, abonelik tipi, toplam fatura tutarı gibi değişkenlerin kötü niyetli kullanımın tespitinde önemli oldukları ortaya çıkmıştır. Daha sona k-means algoritması ile konuşma alışkanlıklarına gore abone kümelemesi yapılmıştır. Önem sırasına gore sıralanmış değişken seti ile ilişkili/bağlantılı değişkenler nedeni ile factor analizi sonuçu elde edilen değişken seti olmak üzere iki farklı değişken seti kullanılmıştır.

Son olarakta eğitim ve test setleri üzerinde karar ağaçları, kural tabanlı methodlar, yapay sinir ağları gibi veri madenciliği teknikleri uygulanmış ve çıkan sonuçlar doğruluk, duyarlılık, özgüllük, hassaslık ve hata kareleri ortalamalarının karekökü(HKOK) gibi performans ölçümlerine gore tartışılmıştır.

Anahtar Kelimeler: Sahtekarlık, Mobil telekomünikasyon , Veri madenciliği, Yapay öğrenme

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor **Asst. Prof. Dr. Adem Karahoca,** for his valuable guidance and advice. He has been very supportive and patient throughout the progress of my thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GSM	Global System for Mobile Communications
IMSI	International Mobile Subscriber Identity
SIM	Subscriber Identity Module
CDR	Call Detail Record
PBX	Public Branch Exchange
IMEI	International Mobile Equipment Identity

1 INTRODUCTION TO TELECOM FRAUD

There are many different types of telecom fraud. It could be to steal a phone and make calls or a retailer that reports an incorrect number of subscriptions sold in order to get better commission. Probably every company in telecom business has their own definition of what telecom fraud is. In general fraud can be defined as: **"Every attempt to use the operator's network with no intention of paying for it"**

There are an important difference between **bad debt** and **fraud**. Bad debt concerns people with occasional difficulties in paying for their invoices. This happens only once or twice per person. If the subscriber really can't pay, he or she will most probably be suspended and denied to open a new subscription in the future.

A fraudster, however never has the intention to pay. A fraudster is also more likely to repeat a committed crime. If a subscription is disconnected, the fraudster will probably find ways to obtain a new subscription and continue the fraudulent activities.

The difference between bad debt and fraud can be defined as :"Fraud and bad debt both have to do with network users not paying for the used services. Fraud always includes a lie, and there is no intention to pay involved. Bad debt simply is normal people without resources to pay for the used services."

Telecom fraud emerged in the late 1980's. In the beginning, the only fraud committed was subscription fraud. Subscription fraud is a concept involving different ways to obtain subscription under false identities. Later on, fraudster found out how easy to commit tumbling within analogue networks. Tumbling means to change the

subscription identification information between calls to facilitate calls using other people's subscription. Tumbling quickly became the most common form of telecom fraud. When the operators discovered this, they replied by setting up databases containing valid identity combinations.

This fraudster then started with cloning of analogue handsets. i.e. to copy the complete handset identity and enter it into another handset. Identities in analogue networks could be obtained by scanning the air with a simple scanning device. There have been cases where the subscription and handset identities could be read on the handset packages. A fraudster could simply walk into a store and get the information.

In 1998 a new phase of cloning of subscriptions emerged: Cloning of subscription identities with GSM networks.

Today's fraud situation differs some between analogue and digital networks. In analogue networks, the greatest problems still are cloning and tumbling. GSM networks have the encryption and authentications embedded into the system, and therefore haven't experienced any severe problems with tumbling and cloning fraud. Cloning within GSM requires hard-to-get information and special equipment. However, there are currently strong indications on that this is changing. Cloning equipment is now offered from various sources, at constantly dropping prices, and the time required to clone a SIM card is said to get shorter and shorter.

Telecom operators are investing quite significant sums of money in technology and staff training in their efforts to prevent and detect and analyze fraud. Technology investments include authentication, encryption, fingerprinting, profiling systems,

fraud detection systems etc. These investments are costly, but are often required to stop losses as a result of increasing fraud.

In many markets, the actions taken to prevent tumbling and cloning have forced the fraudsters to go back to committing subscription fraud. This is hence another area where many operators are focusing on preventing and detecting fraud, both in analogue and digital networks.

The costs of fraud do not only include unpaid invoices. Fraud might also lead to a potential loss of new and existing customers, as well as bad publicity. Fraud case investigations involve a lot of manpower. Roaming fraud generates roaming and interconnect charges. If it is a common knowledge that an operator does not make any attempts to prevent fraud, customers will turn to another operator for subscriptions. The fraudsters will most probably go to other way.

No one can be certain about the future. One thing that we do know is that fraud will continue to increase if operators do not implement an anti-fraud strategy consisting of different technical solutions, fraud management polices, and procedural solutions.

Another area that is interesting for the future is 3G. The 3G expansion involves a complex integration of different types of network entities and services, in combination with content billing as a complement to the existing usage billing. This opens up vulnerability for operators, but opportunities for the fraudsters, to be compared with same evolution the computer industry just experienced with massive virus threats when the internet was introduced. Neither users nor the industry were prepared for the new threats and concentrated more on features and functions than on security. This is very logical since features and functions are the driving forces

within most development. 3G opens up new ways of using the networks, making services and content more valuable than itself.

It is probably true that it is impossible to totally eliminate fraud. The fraudsters will always seek a way to beat the system and any fraud detection mechanism has to be cost effective.

There are many different fraud types. Here we explain the most important ones.

Subscription Fraud: This is by far the most common fraud encountered on the GSM network. Subscription fraud can be performed in different ways. – By using own identity (with changing some of his personal information) and by using false, or other people's identities- Subscription fraud can be further subdivided into two categories. The first is for personal usage by the fraudster, or someone he passes the phone on to. The second is for real profit; here the fraudster claims to be a small business to obtain a number of handsets for Direct Call Selling purposes. The fraudster, who has no intention of paying his bill, now sells on the airtime, probably for cash, to people wishing to make cheap long distance calls.

Cloning: The purpose of cloning is to facilitate calls where another person's subscription is used. The benefit of cloning is that calls that are made from a cloned phone will be charged to the person having the original subscription. Cloning has mainly considered as related to analogue networks. However, cloning is now reality also within GSM. Cloning of GSM phones has been considered as a very complicated operation. Each subscriber is identified by the IMSI number. The IMSI is stored on the SIM card. The SIM card also holds a secret key that is required when

authenticating the subscriber in the network. To further enhance security, the authentication process involves encryption of the information.

Cloning of a GSM phone means that a copy of the SIM card is made. What makes this complicated is that the fraudster needs both the IMSI and the encryption key. There is no use of scanning the air for it since the encryption key is never transmitted. The required information may be obtained from the operator's network (internal fraud) or from the SIM supplier (supplier fraud) but to read it from SIM card, the encryption algorithm has to be cracked. However, now it can be done by using off-the-shelf equipment. Cloning of SIM cards, as it is known to be performed today, requires the actual SIM card that should be copied. A smart card reader, connected to a computer, is used to read the information on it and than copied to an empty SIM card.

Premium Rate Fraud: This involves the abuse of the premium rate services and can occur in different ways. For example, a person could set up a premium rate line with a national operator. The operator is obliged to pay the owner of the line a proportion of the revenue generated. The fraudster then uses a fraudulent mobile phone to dial this number for long period. He may also get other people to do the same. The fraudster then gets the revenue without paying for his own calls. A further way in which premium rate services can be abused is by setting up a fraudulent mobile to divert calls to a popular premium rate line. The caller then only pays normal rates whilst the fraudulent mobile picks up the tab at premium rate. Characteristics are again long back-to-back calls.

Roaming Fraud: Roaming means that operators let visiting subscribers use their networks and their own subscribers use their partners' networks. This makes it possible for subscribers to use their mobile phones also in foreign countries, or in areas not covered by their home operators' networks.

Roaming fraud is a considerable threat for an operator of two reasons.

- Roaming fraud is more interesting from a fraud point of view, since this is an area where the potential profit is higher
- The call data has to be provided from the roaming partners, and this causes a delay in the analysis process. Roaming fraud is therefore harder to detect for an operator lacking the proper equipment.

A solution to detect roaming fraud at an earlier stage would be if roaming partners (operators) exchange roaming CDR's more often, and faster.

Another aspect that has to be considered is that roaming fraud causes actual loss of money for the operators. The roaming partners have to be paid for providing their networks, no matter if the customer pays in the end.

Prepaid Fraud: Many operators believe that prepaid subscriptions are the solution to solve telecom fraud. This is not true. Well some fraud types do not exist within the prepaid area, but prepaid subscriptions open up doors for some new fraud types.

A commonly used payment method for prepaid subscription is payment via vouchers. The voucher represents a value. This value is added to the prepaid account when a certain number is called, and the PIN code that follows with voucher is entered. The PIN code is hidden when voucher is purchased e.g. by a layer that can

be removed by scratching it off. The voucher systems is an example of what can be weakness with prepaid subscriptions. Identified instances include cheque fraud, credit card fraud, voucher theft, voucher ID duplication, faulty vouchers, network access fraud, network attack, long duration calls, handset theft, and roaming fraud.

Fixed Line Fraud : Fraud is a problem also in fixed line networks. There are numerous examples of public phones that have been misused for making free calls, e.g. by using simple tone generators. Another concrete example of fraud in fixed line networks is misuse of company PBXs. Companies may e.g. provide toll free numbers that route calls long distance or international. These numbers, enter a PIN code, and then be routed to their final destination. This enables the staff to make these calls without being charged for more than a local call.

The toll free numbers and PIN codes can be obtained in numerous ways, e.g. via the companies' staff or by hackers. Once this information is known it will spread quickly, at the same time as the misuse will be difficult to detect for the company. Without proper routines, they might not become aware of it until to receive the bill or when the PBX is congested with unauthorized traffic.

Dealer Fraud: Dealers often get commissions on the number of subscriptions they have sold. It is important to have rules surrounding the payment of the commissions to avoid having dealer exploiting any weaknesses. Below is a list of some known commission fraud cases.

- Subscriptions were first sold fictitiously to get the commission. The handsets that were included in the subscriptions were sold again or exported.
- Dealers reported a number of sold units that exceeded the true amount.

- Dealers purchased stolen units, changed the IMEI number and connected the units to the network.

Dealers are also potential source for GSM cloning cases. They have not only physical access to SIM cards, but also have possession of them the amount of time that is required to copy them.

1.1 Motivation

Huge amounts of data are being collected and kept in the warehouses as a result of increased use of mobile communication services. Insight information and knowledge derived from the databases in order to give operators a competitive edge in terms of customer care and retention, marketing and fraud detection. Thus telecommunication fraud has become a high priority item on the agenda of most telecom operators.

Fraud is a significant source of lost revenue to the telecom industry. Efficient fraud detection systems and analysis systems can save telecom operators a lot of money. Automated fraud detection systems enable operators to respond to fraud by detection, service denial and prosecutions against fraudulent users.

In general, the more advanced a service, the more it is vulnerable to fraud. In the future, operators will need to adopt rapidly to keep pace with new challenges posed by fraudulent users.

1.2 Related Work

In this section, we review published work with relevance to fraud detection in telecommunication networks. Phua, Lee, Smith, and Gayler(2005) made a comprehensive survey of data mining techniques applied to fraud detection. The biggest revenue leakage area in the telecom industry is fraud (Wieland, 2004). Global telecommunications fraud losses are estimated in the tens of billions of dollars every year (FML, 2003). Some authors have emphasized the importance of distinguishing between fraud prevention and fraud detection (Bolton & Hand 2002). Fraud prevention describes measures to avoid fraud to occur in the first place. In contrast, fraud detection involves identifying fraud as quickly as possible once it has been committed. Bolton & Hand(2002) reviewed the statistical and machine learning technologies for fraud detection including their application to detect activities in money laundering, e-commerce, credit card fraud, telecommunication fraud and computer intrusion.

Cahill-Lambert-Pinheiro-Sun, (2000) The basis of the approach to detection is an account summary which is called an account signature, that is designed to track legitimate calling behavior for an account. An account signature might describe which call durations, times between calls, days of week and times of day, terminating numbers and payment methods are likely for the account and which are unlikely for the account. Signatures evolve with each new call that is not considered fraudulent, so each established customer eventually has its own signature. Likewise fraud signatures are defined for each kind of fraud using the same structure as an account signature. A call is scored by comparing its probability to belong to the account signature and fraud signature. For new accounts the first calls are used to assign

signature components associating them with calling patterns of a given segment of customers with similar initial information. The history of telecommunication crime, including several types of fraudulent activities, was reviewed by (Collins 1999a,b-2000).

Shawe-Taylor(1999) distinguished six different fraud scenarios; subscription fraud, the manipulation of PBX , frees phone fraud, handset theft, roaming fraud, premium rate service fraud. Subscription fraud which is defined as the use of telephone services with no intention of paying is most significant and prevalent telecom fraud. Subscription fraud is difficult to distinguish from bad debt, particularly if the fraud is personal usage.

Burge-Shawe-Taylor(1997)and Fawcett-Provost (1997) presented adaptive fraud detection. Fraudster adapt to new prevention and detection measures, so fraud detection needs to be adaptive and evolve over time. However, legitimate account users may gradually change their behavior over a longer period of time, and it is important to avoid spurious alarms. Models can be updated at fixed time points or continuously over time.

Moreau-Vandewalle (1997), designing a multilingual information system, the most important success factors of a multilingual information system should be examined that are the degree of authoring automation and cultural customization it offers and cross-lingual processing capability.

ACTS AC095 (1996-1997), The detection of fraud in mobile telecommunications was investigated in European project Advanced Security for Personal Communications Technologies (ASPeCT). The ASPeCT fraud detection tool is

based on investigating sequences of call detail records which contain the details of each mobile phone call attempt for billing purposes. The information produced for billing also contains usage behavior information valuable for fraud detection. A differential analysis is performed to identify a fraudster through profiling the behavior of a user. The analysis of user profiles are based on comparison of recent and longer term histories derived form the toll ticket data. ASPeCT fraud detection tool utilizes a rule based system for identifying certain frauds and neural networks to deal with novel or abnormal instances. Fawcett-Provost (1997) developed a method for choosing account specific thresholds rather than universal thresholds. Their procedure takes daily traffic summaries for a set of accounts that experienced at least 30 days of fraud free traffic activity followed by a period of fraud. This method was applied to cellular cloning, in which fraudulent usage is superimposed upon legitimate usage of an account. For each account a set of rules that distinguish fraud from non-fraud was developed. The superset of the rules for all accounts was then pruned by keeping only those that cover many accounts with possibly different thresholds for different accounts.

1.3 Roadmap

This study examines the Fraud Detection via use of Data Mining Techniques. First we take a look at the various Data Mining Techniques, then we use the Exploratory Data Analysis (EDA) in order to get to know the data set, then apply k-means cluster method to segment the customer based on their call behaviors. Two different set of variables were used. 1) Standard Variables, 2) Variables obtained from Factor Analysis due to some correlated variables. Finally, we perform the data mining methods for both training and test sets, and then we discuss the collected results.

2 REVIEW OF DATA MINING

2.1 Introduction

Data mining is predicted to be "one of the most revolutionary developments of the next decade" according to the online technology magazine ZDNET News (February 8, 2001). In fact, the MIT Technology Review chose data mining as one of ten emerging technologies that will change the world. According to the Gartner Group, "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques."

Databases today can range in size into the terabytes. Within these masses of data lies hidden information of strategic importance. But when there are so many trees, how do you draw meaningful conclusions about the forest.

The newest answer is data mining, which is being used both to increase revenues and to reduce costs. The potential returns are enormous. Innovative organizations worldwide are already using data mining to locate and appeal to higher value customers, to reconfigure their product offering to increase sales, and to minimize losses due to errors or fraud.

Data mining is a process that users a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid prediction.

The first and simplest analytical step in data mining is to **describe** the data – summarize its statistical attributes (such as means and standard deviations), visually review it using charts and graphs, and look for potentially meaningful links among variables(such as values that often occur together) .

But the data description alone cannot provide an action plan. You must **build a predictive model** based on patterns determined from results, and then test that model on results outside the original sample.

The final step is to empirically **verify** the model. For example, from a database of customers who have already responded to a particular offer, you've built a model predicting which prospects are likeliest to respond to the same offer.

Data mining takes advantages in the fields of artificial intelligence(AI) and statistics. Both disciplines have been working on problems of pattern recognition and classification. Both communities have made great contributions to the understanding and application of neural nets and decision trees.

Data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in the statistics community. The development of most statistical techniques was, until recently on elegant theory and analytical methods that worked quite well on the modest amounts of data being analyzed. The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows have allowed the development of new techniques based on a brute force exploration of possible solutions.

New techniques include relatively recent algorithms like neural nets and decision trees, and new approaches to older algorithms such as discriminant analysis. By virtue of bringing to bear the increased computer power on the huge volumes of available data, these techniques can approximate almost any functional form or

interaction on their own. Traditional statistical techniques rely on the modeler to specify the functional form and interactions.

The key point is that data mining is the application of these and other and statistical techniques to common business problems in a fashion that makes these techniques available to the skilled knowledge worker as well as the trained statistics professional. Data mining is a tool for increasing the productivity of people trying to build predictive models.

Data mining is increasingly popular because of the substantial contribution it can make. It can be used to control costs as well as contribute to revenue increases.

Many organization are using data mining to help manage all phase of customer life cycle, including acquiring new customers, increasing revenue from exiting customers, and retaining good customers. By determining characteristics of good customers (profiling), a company can target prospects with similar characteristics. By profiling customers who have bought particular product it can focus attention on similar customers who have not bought that product (cross-selling). By profiling customers who have left, a company can act to retain customers who are at risk for leaving (reducing churn or attrition), because it is usually far less expensive to retain a customer than acquire a new one.

Data mining offers value across a broad spectrum of industries. Telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services. Insurance companies and stock exchanges are also interested in applying this technology to reduce fraud.

Medical applications are another fruitful area: data mining can be used to predict the effectiveness of surgical procedures, medical tests or medications. Companies active in the financial markets use data mining to determine market and industry characteristics as well as to predict individual company and stock performances. Retailers are making more use of data mining to decide which products to stock in particular stores(and even how to place them within a store), as well as to assess the effectiveness of promotions and coupons. Pharmaceutical firms are mining large databases of chemical compounds and of genetic material to discover substances that might be candidates for development as agents for the treatments of disease.

There are two keys to success in data mining. First is coming up with a precise formulation of the problem you are trying to solve. A focused statement usually results in the best payoff. The second key is using the right data. After choosing from the data available to you, or perhaps buying external data, you may need to transform and combine it in significant ways.

The more the model builder can “play” with the data, build models, evaluate results, and work with the data some more (in a given unit of time), the better the resulting model will be. Consequently, the degree to which a data mining tool supports this interactive data exploration is more important than the algorithms it uses.

Ideally, the data exploration tools (graphics/visualization, query/OLAP) are well-integrated with the analytics or algorithms that build the models.

2.2 Data Description for Data Mining

2.2.1 Summaries and visualization

Before we can build good predictive model, we must understand our data. We can start by gathering a variety of numerical summaries (including descriptive statistics such as averages, standard deviations, and so forth) and looking at the distribution of the data. We may want to produce cross tabulations (pivot tables) for multi-dimensional data.

Data can be continuous, having any numerical value (e.g., quantity sold) or categorical, fitting into discrete classes (e.g., red, blue, green). Categorical data can be further defined as either ordinal, having a meaningful order (e.g., high/medium/low) , or nominal, that is unordered(e.g., postal codes). Graphing and visualization tools are vital aid in data preparation and their importance to effective data analysis cannot be overemphasized. Data visualization most often provides the Aha! leading to new insights and success. Some of the common and very useful graphical displays of data are histograms or box plots that display distributions of values. We may also want to look at scatter plots in two or three dimensions of different pairs of variables. The ability to add a third, overlay variable greatly increases the usefulness of some types of graphs.

Visualization works because it exploits the broader information bandwidth of graphics as opposed to text or numbers. It allows people to see the forest and zoom in on the trees. Patterns, relationships, exceptional values and missing values are often easier to perceive when shown graphically, rather than as lists of numbers and text. The problem in using visualization stems from the fact that models have many

dimensions or variables, but we are restricted to showing these dimensions on a two-dimension computer screen or paper.

2.2.2 Clustering

Clustering divides a database into different groups. The goal of clustering is to find groups that are different from each other, and whose members are very similar to each other. Consequently, someone who is knowledgeable in the business must interpret the clusters. Often it is necessary to modify the clustering by excluding variables that have been employed to group instances, because upon examination by the user identifies them as irrelevant or not meaningful. After you have found clusters that reasonably segment your database, these clusters may be used to classify new data. Some of the common algorithms used to perform clustering include Kohonen feature maps and K-means.

2.2.3 Link Analysis

Link analysis is a descriptive approach to exploring data that can help identify relationships among values in a database. The most common approaches to link analysis are **association discovery** and **sequence discovery**. Association discovery finds rules about items that appear together in an event such as a purchase transaction. Market-basket analysis is a well-known example of association discovery. Remember that association or sequence rules are not really rules, but rather descriptions of relationships in a particular database. There is no formal testing of models on other data to increase the predictive power of these rules. Rather there is an implicit assumption that the past behavior will continue in the future. It is often difficult to decide what to do with association rules you are discovered. In store planning, for example, putting associated items physically close together may reduce

the total value of market basket- customers may buy less overall because they no longer pick up unplanned items while walking through the store in search of desired items. Insight, analysis and experimentation are usually required to achieve any benefit from association rules .Graphical methods may also be very useful in seeing the structure of links. For instances, looking at an insurance database to detect potential fraud might reveal that a particular doctor and lawyer work together on an unusually large number of cases.

2.3 Predictive Data Mining

The goal of data mining is to produce new knowledge that the user can act upon. It does this by building a model of the real world based on data collected from variety of sources which may include corporate transactions, customer histories and demographic information, process control data, and relevant external databases such as credit bureau information or weather data. The result of the model building is a description of patterns and relationships in the data that can be confidently used for prediction.

To avoid confusing the different aspects of data mining. It helps to envision a hierarchy of the choices and decisions you need to make before you start;

- Business Goals
- Type of prediction
- Model type
- Algorithm
- Product

At the highest level is the **business goal**: what is the ultimate purpose of mining this data? For example, seeking patterns in your data to help you retain good customers, you might build one model to predict customer profitability and a second model to identify customers likely to leave (attrition). Your knowledge of your organization's needs and objectives will guide you in formulating the goal of your models.

The next step is deciding on the **type of prediction** that's most appropriate:

(1) classification: predicting into what category or a class a case falls, or *(2) regression*: predicting what number value a variable will have (if it's a variable that varies with time, it is called time series prediction), in the example above, you might use regression to forecast the amount of profitability, and classification to predict which customers might leave. These are discussed in more detail below.

Now you can choose the **model type**: a neural net to perform the regression, perhaps, and a decision tree for the classification. There are also traditional statistical models to choose from such as logistic regression, discriminant analysis, or general linear models.

Many **algorithms** are available to build your models. You might build the neural net using back propagation or radial basis functions. For the decision tree, you might choose among CART, C5.0, QUEST, or CHAID.

When selecting a data mining **product**, be aware that they generally have different implementations of a particular algorithm even they identify it with the same name. These implementation differences can affect operational characteristics such as

memory usage and data storage, as well as performance characteristics such as speed and accuracy.

Many business goals are best met by building multiple model types using a variety of algorithms. You may not able to determine which model type is best until you've tried several approaches.

In predictive models, the values or classes we are predicting are called the response, dependent or target variables. The values used to make the prediction are called the predictor or independent variables.

Predictive models are built, or trained, using data for which the value of response variable is already known. This kind of training is sometimes referred to as supervised learning because calculated or estimated values are compared with the known results. By contrast, descriptive techniques such as clustering are sometimes referred to as unsupervised learning because there is no already-known result to guide the algorithms

2.4 Data Mining Models and Algorithms

Let's look at some of the types of models and algorithms used to mine data. Most products use variations of algorithms that have been published in computer science or statistics journals.

2.4.1 Neural Networks

Neural networks are of particular interest because they offer a means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions. (Actual biological neural networks are

incomparably complex.) Neural nets may be used in classification problems (where the output is a categorical variable) or for regression (where the output variable is continuous).

A neural network starts with an *input layer*, where each node corresponds to a predictor variable. These 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.

One of the advantages of neural network models is that they can be implemented to run on massively parallel computers with each node simultaneously doing its own calculations. Users must be conscious of several facts about neural networks. First, neural networks are not easily interpreted. There is no explicit rationale given for the decisions or predictions a neural network makes.

Second, they tend to over fit the training data unless very stringent measures, such as weight decay and/or cross validation, are used judiciously. This is due to the very large number of parameters of the neural network which, if allowed to be of sufficient size, will fit any data set arbitrarily well when allowed to train convergence.

Third, neural networks require an extensive amount of training time unless the problem is very small. Once trained, however they can provide predictions very quickly.

Fourth, they require no less data preparation than any other method, which is to say they require a lot of data preparation. One myth of neural networks is that data of any quality can be used to provide reasonable predictions. The most successful implementations of neural networks (or decision trees, or logistic regression, or any other method) involve very careful data cleansing, selection, preparation and pre-processing. For instance, neural nets require that all variables be numeric. Therefore categorical data such as 'state' is usually broken up into multiple dichotomous variables, each with "1"(yes) or "0" (no) value. The resulting increase in variables is called the categorical explosion.

Finally, neural networks tend to work best when the data set is sufficiently large and the signal-to noise ratio is reasonable high. Because they are so flexible, they will find many false patterns in a low signal-to-noise ratio situation.

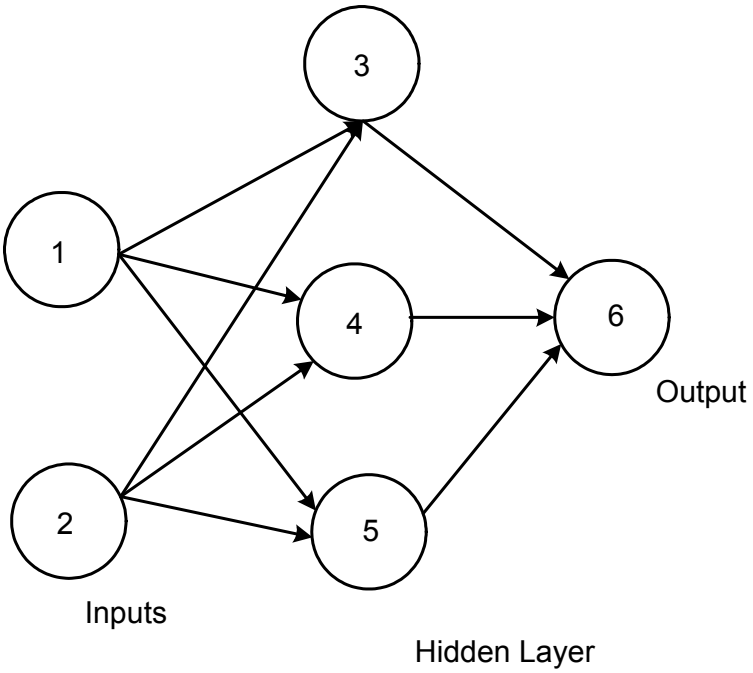

Figure 2. 1 A Neural Network with a hidden layer

2.4.2 Decision Trees

Decision trees are a way of representing a series of rules that lead to a class or value. For example, you may wish to classify loan applications as good or bad credit risks. Figure 2.2 shows a simple decision tree that solves this problem while illustrating all the basic components of a decision tree: the decision node, branches and leaves.

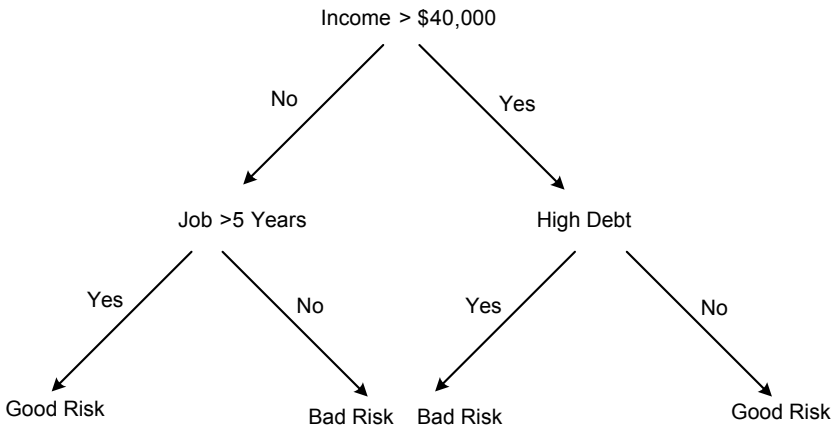

Figure 2. 2 A Simple Classification tree

The first component is the top decision node, or root node, which specifies a test to be carried out. The root in this example is "Income > $ 40,000." The results of this test cause the tree to split into branches, each representing one of the possible answers. In this case, the test “Income> $40,000” can be answered either “yes” or “no”, so we get two branches.

Depending on the algorithm, each node may have two or more branches. For example, CART generates trees only two branches at each node. Such a tree is called a binary tree. When more than two branches it is called a multi way tree.

Decision trees models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions. A number of different algorithms may be used for building decision trees including CHAID (Chi-squared Automatic Interaction Detection), CART(Classification And Regression Trees), QUEST, and C5.0. Decision trees which are used to predict categorical variables are called *classification trees* because they place instances in categories or classes. Decision trees used to predict continuous variables are called *regression trees*.

2.4.3 Multivariate Adaptive Regression Splines (MARS)

In the mid-1980's one of the inventors of CART, Jerome H. Friedman, developed a method designed to address its shortcomings.

The main disadvantages he wanted to eliminate were:

- Discontinuous predictions (hard splits).
- Dependence of all splits on previous ones.
- Reduced interpretability due to interactions, especially high-order interactions.

To this end he developed the MARS algorithm. The basic idea of MARS is quite simple, while the algorithm itself is rather involved. Very briefly, the CART disadvantages are taken care of by:

- Replacing the discontinuous branching at a node with a continuous transition modeled by a pair of straight lines. At the end of the

model-building process, the straight lines at each node are replaces with a very smooth function called spline.

- Not requiring that new splits be dependent on previous splits.

Unfortunately, this means is a method for deriving a set of rules to clarify and MARS loses the tree structure of CART and cannot produce rules. On the other hand, MARS automatically finds and lists the most important predictor variables as well as the interactions among predictor variables. MARS also plots the dependence of the response on each predictor. The result is an automatic non-linear step-wise regression tool.

2.4.4 Rule Induction

Rule induction is a method for deriving a set of rules to classify cases. Although decision trees can produce a set of rules, rule induction methods generate a set of independent rules which do not necessarily (and are unlikely to) form a tree. Because the rule inducer is not forcing splits at each level, and can look ahead, it may be able to find different and sometimes better patterns for classification. Unlike trees, the rules are generated may not cover all possible situations. Also unlike trees, rules may sometimes conflict in their predictions, in which case it is necessary to chose which rule to follow. One common method to resolve conflicts is to assign a confidence to rules and use the one in which you are most confident. Alternatively, if more than two rules conflict, you may let them vote, perhaps weighting theirs votes by confidence you have in rule.

2.4.5 K-Nearest neighbor and memory-based reasoning (MBR)

When trying to solve new problems, people often look at situation to similar problems that they have been previously solved. K-nearest neighbor (K-NN) is a classification technique that uses a version of this same method. It decides in which class to place a new case by examining some number – the "k" in k-nearest neighbor – of the most similar cases of neighbor. It counts the number of cases for each class, and assigns the new case to the same class to which most of its neighbors belong.

The first thing you must do to apply k-NN is to find a measure of the distances between attributes in the data and then calculate it. While this is easy for numeric data, categorical variables need special handling. For example, what is the distance between blue and green? You must then have a way of summing the distance measures for the attributes. Once you can calculate the distance between cases, you then select a set of already classified cases to use as the basis for classifying new cases, decide how large a neighborhood in which to do the comparisons, and also decide how to count the neighbors themselves. (e.g., you might give more weight to nearer neighbors than farther neighbors).

K-NN puts a large computational load on computer because the calculation time increases as the factorial of the total number of points. K-NN models are very easy to understand when there are few predictor variables. They are also useful for building models that involve non-standard data types, such as text. The only requirement for being able to include a data type is the existence of an appropriate metric.

2.4.6 Logistic Regression

Logistic regression is a generalization of linear regression. It is used primarily binary variables (with values such as yes/no or 0/1) and occasionally multi-class variables. Because the response variable is discrete, it cannot be modeled directly by linear regression. Therefore, rather than predict whether the event itself (the response variable) will occur, we build the model to predict the logarithm of odds of its occurrence. This logarithm is called the log odds or the logit transformation.

The odds ratio:

Probability of an event occurring / probability of the event not occurring

Having predicted the log odds, you then take the anti-log of this number to find the odds.

While logistic regression is a very powerful modeling tool, it assumes that the response variable (the log odds, not the event itself) is linear in the coefficients of the predictor variables. Furthermore, the modeler, based on his or her experience with the data and data analysis, must choose the right inputs and specify their functional relationship to the response variable. It is up to model builder to search for the right variables, find their correct expression, and account for their possible interactions.

2.4.7 Discriminant analysis

Discriminant analysis is the oldest mathematical classification technique, having been first published by R.A. Fisher in 1936 to classify the famous Iris botanical data into three species. It finds hyper-planes (e.g., lines in two dimension, planes in

three etc.) that separate the classes. The resultant model is very easy to interpret because all the user has to do is determine on which side of the line (or hyper-plane) a point falls. Training is simple and scalable. The technique is very sensitive to patterns in the data. It is used very often in certain disciplines such as medicine, the social science, and field biology.

Discriminant analysis is not very popular in data mining, however, for three main reasons. First, it assumes that all of the predictor variables are normally distributed (i.e., their histograms look like bell-shaped curves), which may not be the case. Second, unordered categorical predictor variables (e.g., red/blue/green) cannot be used at all. Third, the boundaries that separate the classes are all linear forms(such as lines or planes), but sometimes the data just can't be separated that way. Recent versions of discriminant analysis address some of these problems.

2.4.8 Generalized Additive Models (GAM)

There is a class of models extending both linear and logistic regression; known as generalized additive models or GAM. They are called additive because we assume that the model can be written as the sum of possibly non-linear functions, one for each predictor. GAM can be used either for regression or for classification of a binary response. The response variable can be virtually any function of the predictors as long as there are not discontinuous steps. GAM, using computer power in place of theory or knowledge of the functional form, will produce a smooth curve, summarizing the relationship as described above. The most common estimation procedure is backfitting. Instead of estimating large numbers of parameters as neural nets do, GAM goes one step further and estimates a value of the output for each value of the input. – one point, one estimate. As with the

neural net, GAM generates a curve automatically, choosing the amount of complexity based on the data.

2.4.9 Boosting

If you were to build a model using one sample data, and then build a new model using the same algorithm but on a different sample, you might get a different result. After validating the two models, you could choose the one best met your objectives. Even better results might be achieved if you build several models and let them vote, making a prediction based on what the majority recommended. Of course, any interpretability of the prediction would be lost, but the improved results might be worth it.

This is exactly the approach taken by boosting, a technique first published by Freund and Schapire in 1996. Basically, boosting takes multiple random samples from the data and builds a classification model for each. The training set is changed based on the results of the previous models. The final classification is the class assigned most often by the models. The exact algorithms for boasting have evolved from the original, but the underlying idea is the same. Boosting has become a very popular addition to data mining packages.

2.4.10 Genetic Algorithms

Genetic algorithms are not used to find patterns, but rather to guide the learning process of data mining algorithms such as neural nets. Essentially, genetic algorithms act as a method for performing a guided search for good models in the solution space.

They are called genetic algorithms because they loosely follow the pattern of biological evolution in which the members of one generation (of models) compete to pass on their characteristics to the next generation (of models), until the best (model) is found. The information to be passed on is contained in "chromosomes," which contain the parameters for building the model.

For example, in building a neural net, genetic algorithms can replaces backpropagation as a way to adjust the weights. The chromosome in this case would contain the weights. Alternatively, genetic algorithms might be used to find the best architecture, and chromosomes would contain the number of hidden layers and the number of nodes in each layer.

While genetic algorithms are an interesting approach to optimizing models, they add a lot of computational overhead.

2.5 The Data Mining Process

Recognizing that a systematic approach is essential to successful data mining, many vendor and consulting organizations have specified a process model designed to guide the user (especially someone new to building predictive models) through a sequence of steps that will lead to good results. SPSS used 5A's – Assess, Access, Analyze, Act and Automate- and SAS uses SEMMA – Sample, Explore, Modify, Model, Assess.

A consortium of vendors and users consisting of NCR systems engineering (Copenhagen-Denmark), Daimler-Benz AG (Germany), SPSS/Integral Solutions Ltd(England) and OHRA Verzekeringen en Bank Groep B.V (The Netherlands) has

been developed a specification called CRISP-DM – Cross Industry Standard Process for Data Mining.

The general CRISP-DM process model includes six phases that address the main issues in data mining. The six phases fit together in a cyclical process, illustrated in the following figure.

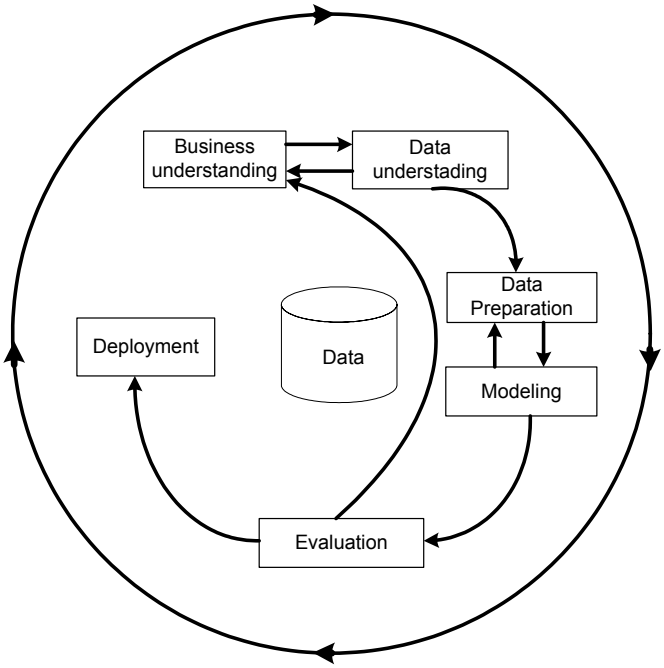

Figure 2. 3 Phases of the CRISP-DM reference model

These six phases cover the full data mining process, including how to incorporate data mining into your larger practices. The six phases include:

- **Business Understanding:** This is perhaps the most important phase of data mining. Business understanding includes determining business objectives, assessing the situation, determining data mining goals, and producing a project plan.
- **Data Understanding:** Data provides the "raw materials" of data mining. This phase addresses the need to understand what your data resources are

and the characteristics of these resources. It includes collecting initial data, describing data, exploring data, and verifying data quality.

- **Data Preparation:** After cataloging your data resources, you will need to prepare your data for mining. Preparations include selecting, cleaning, constructing, integrating, and formatting data.
- **Modeling:** This is, of course, the flashy part of data mining, where sophisticated analysis methods are used to extract information from the data. This phase involves selecting modeling techniques, generating test designs and building and assessing models.
- **Evaluation:** Once you have chosen your models, you are ready to evaluate how the data mining results can help you achieve your business objectives. Elements of this phase include evaluating results, reviewing the data mining process, and determining the next steps.
- **Deployment:**. This phase focuses on integrating your new knowledge into your everyday business processes to solve your original business problem. This phase includes plan deployment, monitoring, and maintenance, producing a final report, and reviewing the project.

3 Methods

3.1 Data Sets

In this thesis, fraud detection is based on the calling activity of mobile phone subscriber (CDR's), demographic data, and payment data. In order to develop models of normal and fraudulent behavior and to be able to diagnostic accuracy of the model, call data, demographics and payment data exhibiting both kinds of behavior are needed.

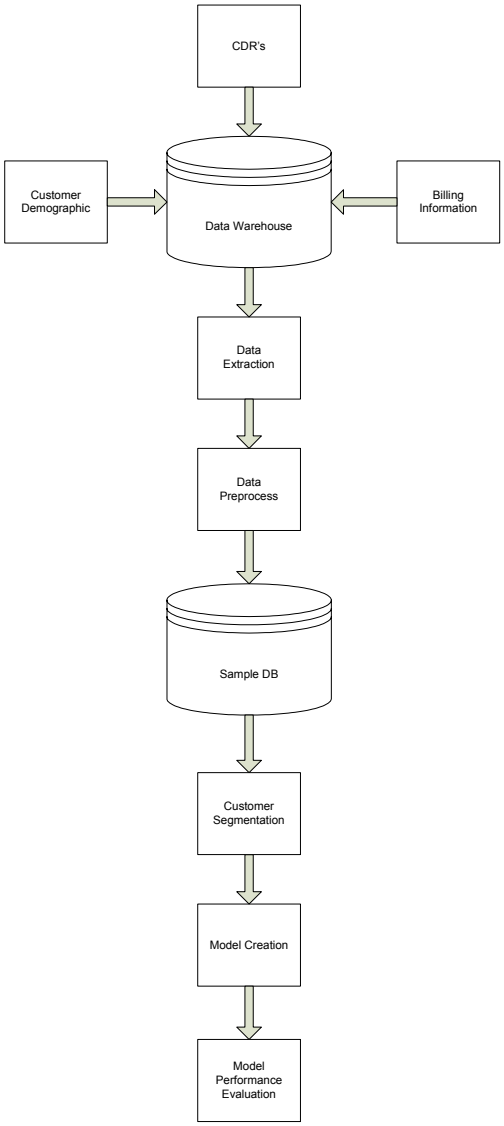

Figure 3. 1 Process of fraud detection with Data Mining

The data of 560 post-paid subscribers from 2002 is obtained from a Mobile Phone Operator's data warehouse. Normal and fraudulent ratio was one to one. First daily CRD's are extracted and aggregated weekly, and then demographic data and payment data are extracted. All of the extracted data are preprocessed and inserted into a table in the database. Then data set is divided into training and test sets. The training set consists of 2/3 of 560 subscribers which is 374 and the test set consists of 1/3 which is 186.

3.2 Exploratory Data Analysis

3.2.1 Getting to know the Data Set

Simple (or-not-so-simple) graphs, plots, and table often uncover important relationships that could indicate fecund areas for further investigation. We use exploratory methods to delve into the fraud data set. We use the Clementine data mining software package from SPSS Inc. for the Exploratory Data Analysis (EDA).The data set contains 40 variables worth of information about 560 customers, along with an indication (status flag A/I) for fraud.

Table 3. 1 Subscriber information (call, demographic and, payment data)

CallerId	Categorical	unique subscriber ID
GenTotMin	Integer	total minutes customer used
GenTotCharge	Continuous	total charge of calls
GenTotCalls	Integer	total number of calls
SMSCharge	Continuous	total charge of SMS calls
SMSCalls	Integer	total number of SMS calls
PSTNTotMin	Integer	total minutes of PSTN calls
PSTNTotCharge	Continuous	total charge of PSTN calls
PSTNTotCalls	Integer	total number of PSTN calls
INTTotMin	Integer	total minutes of International calls
INTTotCharge	Continuous	total charge of international calls
INTTotCalls	Integer	total number of int. calls
VASTotMin	Integer	total minutes of VAS calls
VASTotCharge	Continuous	total charge of VAS calls
VASTotCalls	Integer	total number of VAS calls
ROATotMin	Integer	total minutes of Roaming calls
ROATotCharge	Continuous	total charge of Roaming calls
ROATotCalls	Integer	total number of Roaming calls
OGSMTotMin	Integer	total minutes of other GSM
OGSMTotCharge	Continuous	total charge of other GSM
OGSMTotCalls	Integer	total number of OGSM calls
GSMTotMin	Integer	total minutes of GSM calls
GSMTotCharge	Continuous	total charge of GSM calls
GSMTotCalls	Integer	total number of GSM calls
PRETotMin	Integer	total minutes of Premium calls
PRETotCharge	Continuous	total charge of Premium calls
PRETotCalls	Integer	total number of PRE. Calls
ReasonCode	Integer	reason code for fraud type
ActInvNum	Integer	active number of invoices
ActInvCost	Integer	amount of active invoices
AddressCity	Integer	city name (1-81)
PackageID	Integer	package Type (1-53)
AccountLength	Integer	duration of the subscription
Type	Integer	1= Corporate, 2= Person
Gender	Integer	1= Male, 2= Female
Age	Integer	subscriber age
AlarmNum	Integer	number of raised alarms
AlarmScore	Integer	value of alarms
Cluster	Integer	customer segments
Status	Categorical	A=Active, I=Inactive

3.2.2 Dealing with Correlated Variables

One should take care to avoid feeding correlated variables to one's data mining and statistical models. At best, using correlated variables will overemphasize one data component: at worst, using correlated variables will cause the model to become unstable and deliver unreliable results.

The call behavior data set contains three variables: *mins, calls*, and *charge*. The data description indicates that the charge variable may be a function of minutes and calls, with the result that the variables would be correlated.

There does seem to be relationship between Mins and Calls or between Charge and Calls. One may have expected that the number of calls increased, the number of minutes would tend to increase (and similarly for charge), resulting in a positive correlation these fields. The Table 3.2 shows the Pearson correlation for these fields.

The only difference is TotCharge, TotMins and TotCalls which does not show positive strong correlations like others. Since these fields are derived from the sum of other call type. We can eliminate these there fields for the data mining.

The other call types which have strong positive correlation, we should eliminate the correlated values in order not to get incoherent results.

After ranking the attributes with Gain Attribute Evaluator in Weka and only getting the attributes with higher impact on the outcome, there were not any correlated attributes anymore.

Meantime, we will also apply the Factor Analysis in order to reduce the number of variables as an alternative way.

Table 3. 2 Pearson Correlation Table

Pearson Correlation Table			
	Charge	Mins	Calls
TotCharge		0.664	0.421
TotMins	0.664		0.517
TotCalls	0.421	0.517	
SMSCharge			0.916
SMSCalls	0.916		
PSTNCharge		0.777	0.516
PSTNMins	0.777		0.721
PSTNCals	0.516	0.721	
INTCharge		0.990	0.963
INTMins	0.990		0.955
INTCalls	0.963	0.955	
VASCharge		0.536	0.929
VASMins	0.536		0.399
VASCalls	0.929	0.399	
ROACharge		0.993	0.890
ROAMins	0.993		0.930
ROACalls	0.890	0.930	
OGSMCharge		0.989	0.711
OGSMMins	0.989		0.722
OGSMCalls	0.722	0.722	
GSMCharge		0.911	0.647
GSMMins	0.911		0.718
GSMCalls	0.647	0.718	
PRECharge		0.874	0.749
PREMins	0.874		0.850
PRECalls	0.749	0.850	

3.2.3 Exploring Variables

One of the primary reasons for performing exploratory data analysis is to investigate the variables, look at histograms of numeric variables, examine the distributions of categorical variables, and explore the relationships among sets of variables.

Figure 3.2 shows that there is an association of fraud with the account length. We may say that fraudsters are usually from the new subscribers.

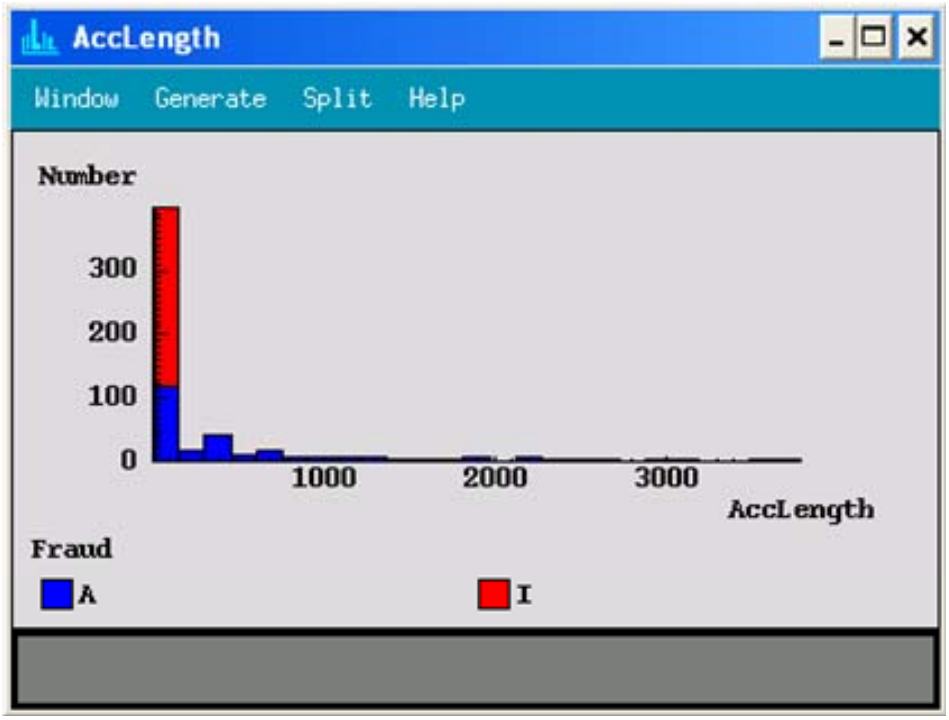

Figure 3. 2 Histogram of subscribers' account length with status overlay

Figure 3.3 show that there are more fraudsters in some regions.

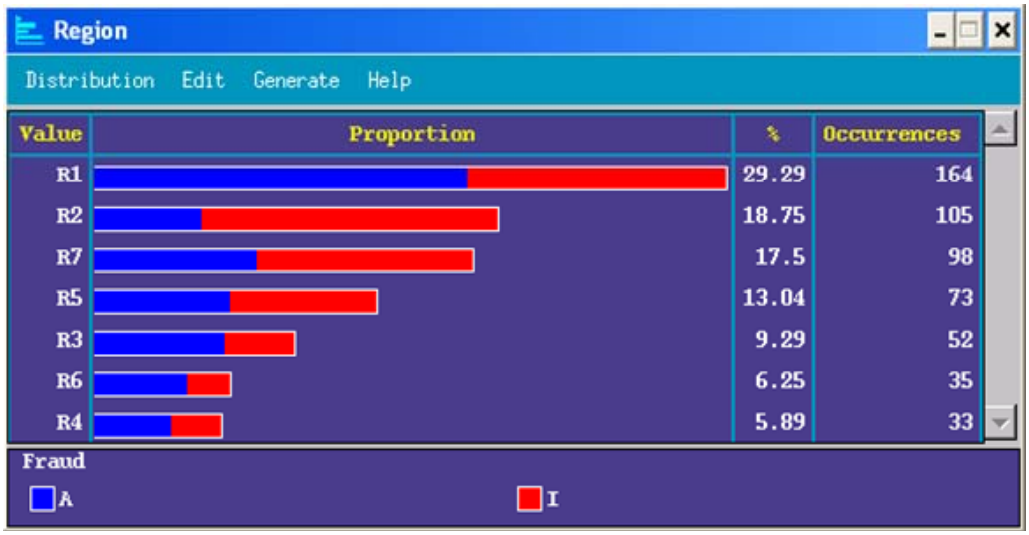

Figure 3. 3 Distribution of Region attribute with status overlay

Figure 3.4 shows that, the number of the fraudsters in some tariff packages is higher.

Figure 3. 4 Distribution of PackID attribute with status overlay

Figure 3.5 shows that Male fraudster are much higher than the female.

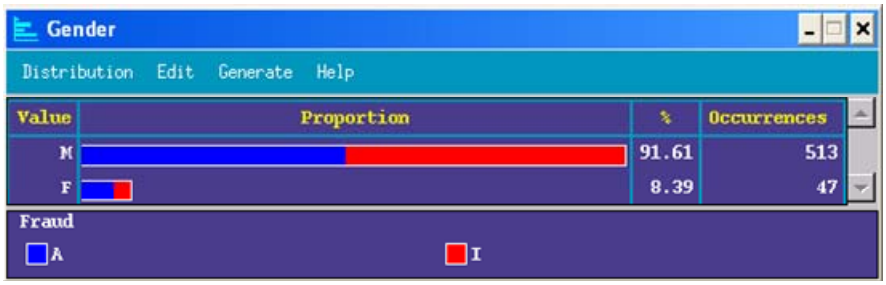

Figure 3. 5 Distribution of Gender attribute with status overlay

Figure 3. 6 Distribution of Type attribute with status overlay

Figure 3.6 show those fraudsters are mainly individual subscriber rather than corporate subscribers.

Figure 3.7 shows that TotCharge, OGSMCharge, GSMCharge and PSTNCharge variables are important variable to detect fraudsters.

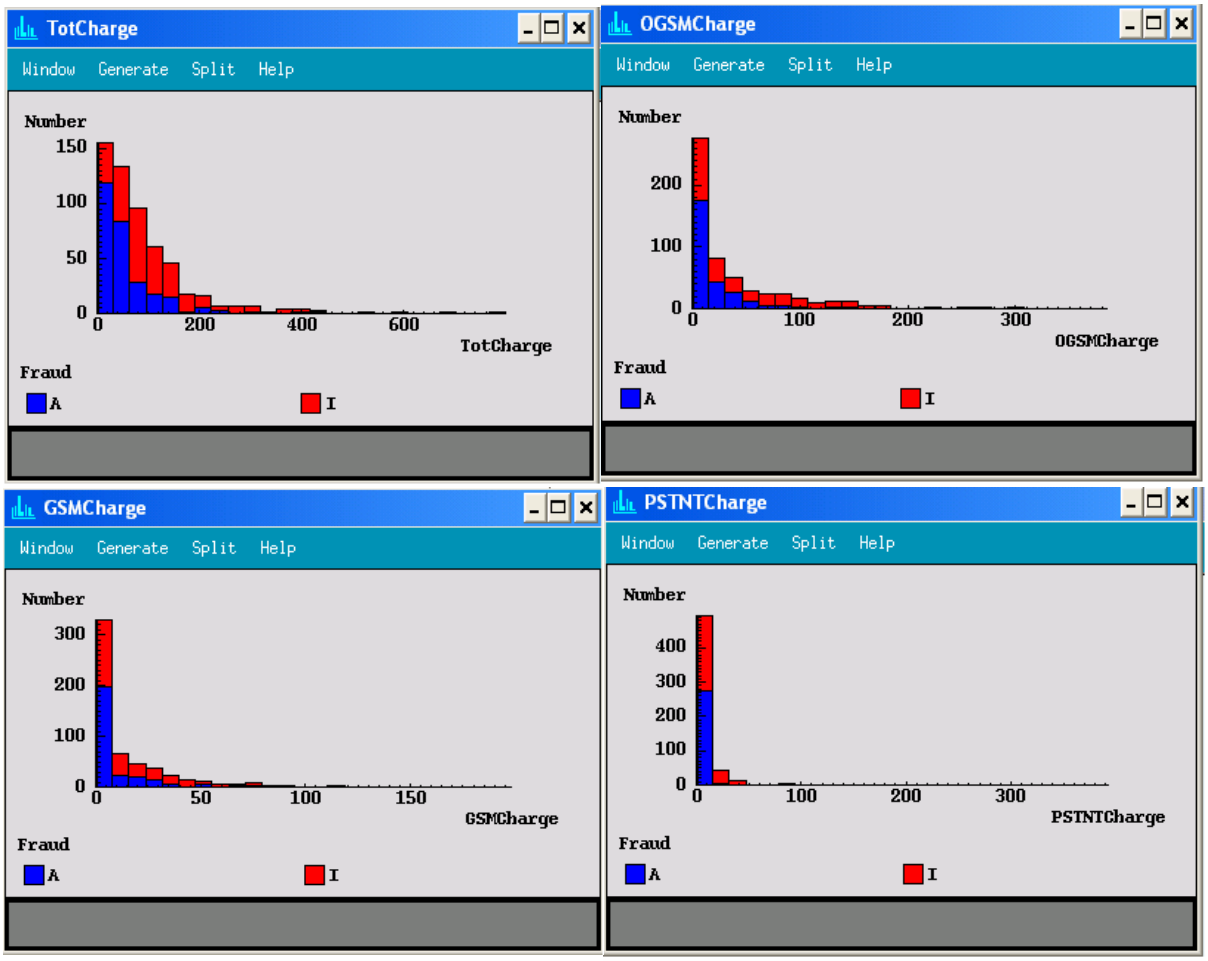

Figure 3. 7 Histogram of TotCharge, OGSMCharge, GSMCharge, PSTNCharge with fraud overlay

Figure 3.8 shows the links between Fraud and Clusters. While Clusters 3 and 5 have weak links, Clusters 1 and 2 have strong links, Cluster 4 has medium link, with fraudulent user.

Clusters 1 and 2 have strong links, Cluster 3, 4 and 5 have medium link with normal user.

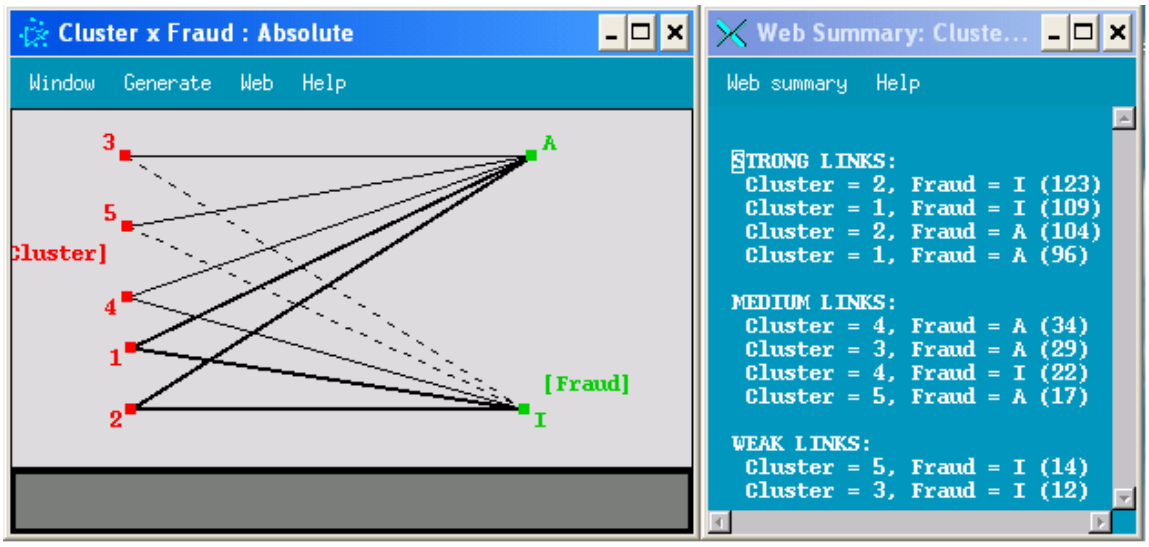

Figure 3. 8 Web graph of Cluster vs. Fraud

As a result of Exploratory Data Analysis, we have found that some variables like AccLentgh, Region, PackID, Gender, Type , TotCharge, OGSMCharge, GSMCharge and PSTNcharge with status overlay show important tendency for fraudulent use. While some of the other variables show slight tendency, others show no tendency at all.

3.3 Cluster Analysis

Cluster analysis is an exploratory data analysis technique designed to reveal natural groupings within a collection of data. The basic criterion used for this is distance, in the sense that cases close together should fall into the same cluster, while observations far apart should be in different clusters. Ideally the cases within a cluster would be relatively homogenous, but different from those contained in other clusters.

As cluster analysis is based on distance derived from the fields in the data, these fields are typically interval, ordinal or binary in scale. When clustering is successful, the results suggest separate segments within the overall set of data.

Given these characteristics, it is not surprising that cluster analysis often employed in market segmentation studies, since the aim is to find distinct types of customers towards whom more targeted and effective marketing and sales action may be taken. In addition, for modeling applications, clustering is sometimes performed first to identify subgroups in the data that might behave differently. These subgroups can then be modeled separately or the cluster membership variable can be included as a predictor.

There are many different clustering methods, but in the area of data mining two are in wide usage. This is because the large class of hierarchical clustering methods requires that distances between every pair of data records be stored (n*(n-1)/2) and updated, which places a substantial demand on memory and resources for the large files common in data mining. Instead clustering is typically performed using K-

means algorithm or using an unsupervised neural network method (Kohonen). Of the two, K-means clustering is considerably faster.

Cluster analysis is not an end in itself, but one step in a data mining project. We would like to use this clusters as predictors for late assistance in classifying customers as fraudulent or not. Therefore, we will not to include the demographics data and S*tatus field* among the variables used to build the clusters.

We used K-means clustering and function "k-means" partitions into K mutually exclusive clusters, and return a vector of indices indicating to which of the k clusters it has assigned each observation. Unlike the hierarchical clustering methods used, K-means does not create a tree structure to describe the groupings in your data, but rather creates a single level of clusters. K-means treats each observation in your data as an object having a location in space. It finds a partition in which objects within each clusters are as close to each other as possible, and as far from objects in other clusters as possible. Each cluster in the partition is defined by its member objects and by its centroid, or center. The centroid for each cluster is the point to which the sum of the distances from all objects in that cluster is minimized. (Mathworks 2002)

We apply the K-means algorithm only to the call behavior data in the fraud *data* set using Matlab Statistics Toolbox.

Attempts to cluster the fraud data set into 2,3,4,5,6 and 7 clusters are made. Silhouette plots and average silhouette values of 2,3,4,5,6,7 clusters made by k-means are compared. It is obviously seen that the best values are in cluster 5. In figure 3.9 silhouette plot for 5 clusters created by k-means are shown.

From the silhouette plot, we can see that most points in all clusters have a large silhouette value, greater than 0.6 indicating that those points are well-separated from neighboring clusters. However, some clusters also contain a few points with negative silhouette values, indicating that they are nearby to points from other clusters.

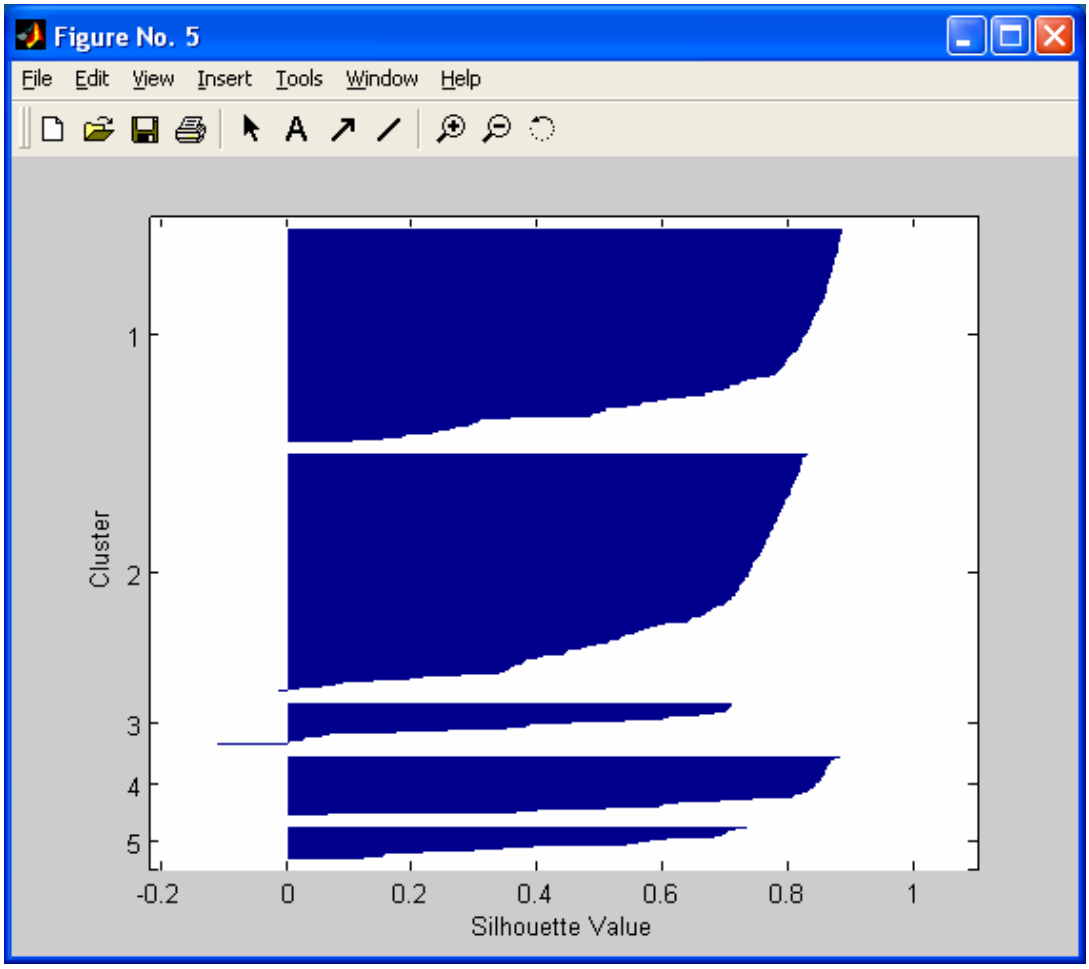

Figure 3. 9 Silhouette plot for 5 clusters

Table 3. 3 Mean () results for different clustering

Number of Clusters	Mean( )
2	0.4857
3	0.5978
4	0.6398
5	0.6570
6	0.1598
7	0.2007

Table 3. 4 the results of Cluster Analysis

Cluster/Status	1A	1I	2A	2I	3A	3I	4A	4I	5A	5I
N	96	109	104	123	29	12	34	22	17	14
SMSCharge	1,94	4,36	1,21	2,91	15,42	44,71	0,51	0,55	0,31	0,49
SMSCalls	12,58	27,52	7,78	18,76	136,86	277,67	2,79	3,32	1,71	3,07
PSTNMins	3,23	10,56	4,76	25,44	3,17	8,25	0,82	7,23	33,71	118,86
PSTNTCharge	1,48	5,49	1,77	12,60	1,31	3,51	0,38	5,72	12,36	84,93
PSTNCalls	2,98	9,66	4,23	18,09	3,76	7,83	1,06	4,27	24,47	35,93
INTMins	0,11	0,06	0,67	0,14	4,24	4,92	0,00	0,00	3,00	0,00
INTCharge	0,17	0,10	0,77	0,13	6,58	8,33	0,00	0,00	3,57	0,00
INTCalls	0,04	0,03	0,18	0,03	1,31	1,67	0,00	0,00	0,71	0,00
VASMins	0,45	1,81	2,88	4,07	27,31	11,67	1,65	15,73	0,06	0,79
VASCharge	1,04	1,19	0,40	2,67	9,90	9,92	0,29	5,00	0,65	2,64
VASCalls	2,30	7,34	1,31	8,24	43,07	57,83	2,44	18,64	0,47	16,79
ROAMins	0,84	0,00	0,00	0,00	3,48	0,00	0,00	0,00	0,47	0,00
ROACharge	1,33	0,00	0,00	0,00	7,10	0,00	0,00	0,00	1,06	0,00
ROACalls	1,52	0,00	0,00	0,00	3,28	0,00	0,00	0,00	0,59	0,00
OGSMMins	18,14	43,58	91,99	256,72	13,52	47,00	9,03	29,00	15,59	51,29
OGSMCharge	8,67	18,45	40,75	108,33	6,90	20,57	5,23	11,97	5,65	21,68
OGSMCalls	15,35	31,96	44,87	140,25	15,24	35,33	5,00	18,82	15,53	37,29
GSMMins	203,25	343,92	18,13	51,11	25,97	84,67	15,44	11,73	8,65	18,79
GSMCharge	24,43	38,66	3,36	6,29	2,97	9,33	2,97	2,00	1,94	2,29
GSMCalls	64,13	117,67	13,85	33,02	9,72	38,50	7,32	9,18	6,35	13,64
PREMins	3,11	3,15	0,17	9,70	0,00	0,25	184,06	232,00	0,00	7,86
PRECharge	2,21	1,60	0,22	5,76	0,00	0,17	152,41	188,73	0,00	4,00
PRECalls	1,40	2,22	0,10	3,50	0,00	0,33	64,53	57,91	0,00	3,29

Table 3.4 shows the results of the clustering. It shows the size of the each cluster and the mean value of each cluster for normal and fraudulent use. Here is a description of the mean profile of each cluster obtained by interpreting the Table 3.4. It is also clearly seen that for most of the call types, fraudsters have much higher mean value than normal users.

Cluster 1: Represents high usage of GSM Minutes, Calls, and Charge and significant use of OGSM Minutes and Calls.

Cluster 2: Represents high usage of OGSM Minutes, Calls and Charge and also significant use of GSM Minutes and Calls.

Cluster 3: Represents high usage of SMS calls, VAS Calls and VAS Minutes.

Cluster 4: Represents high usage of PRE Minutes, Charge and Calls.

Cluster 5: Represents high usage of PSTN Minutes, Calls and significant use of OGSM Minutes, Calls.

3.4 Factor Analysis

Factor analysis attempts to identify underlying variables, or factors, that explain the pattern of correlations within a set of observed variables. Factor analysis is often used in data reduction to identify a small number of factors that explain most of the variance observed in a much larger number of manifest variables. Factor analysis can also be used to generate hypotheses regarding causal mechanisms or to screen variables for subsequent analysis (for example, to identify collinearity prior to performing a linear regression analysis).

Factor Analysis is primarily used for data reduction or structure detection.

The purpose of data reduction is to remove the redundant (highly correlated) variables from the data file, perhaps replacing the entire data file with a smaller number of uncorrelated variables.

The purpose of structure detection is to examine the underlying (or latent) relationship between variables.

The Factor Analysis procedure has several extraction methods for constructing a solution.

For Data Reduction. The principal components method of extraction begins by finding a linear combination of variables. (a component) that accounts for a much variation in the original variables as possible. It then finds another component that accounts for as much of the remaining variation as possible and is uncorrelated with the previous component, continuing in this way until there are as many components as original variables. Usually, a few components will account for most of the variation, and these components can be used to replace the original variables. This method is most often used to reduce the number of variables in the data file.

For Structure Detection. Other Factor Analysis extraction methods go one step further by adding the assumption that some of the variability in the data cannot be explained by the components (usually called factors in other extraction methods). As a result, the total variance explained by the solution is smaller; however, the addition of this structure to the factor model makes these methods ideal for examining relationships between the variables.

In our fraud case, We used SPSS statistical software for the factor analysis in order to reduce the number of highly correlated variables, after appling the factor analysis ,as a result the number of variables reduced from 40 variables to 12 variables.

Table 3.5 shows the KMO and Barlett's Test results. When the value of Kaiser-Meyer-Olkin Measure of Sampling is close to 1 then it means data is suitable for the Factor Analysis and if it is under 0.5 then it means, data is not suitable.

The Barlett's Test of Sphericity shows the correlation between the variables. If the value of Sig. is greater than 0.10 then data is not suitable. In our case, both values are in the right ranges and our data is suitable for the Factor Analysis.

We used SPSS for the factor analysis in order to reduce the number of highly correlated variables, after appling the factor analysis ,as a result the number of variables reduced from 40 variables to 12 variables.

Table 3. 5 KMO and Barlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,645
Bartlett's Test of Sphericity	Approx. Chi-Square	17682,897
	Df	465
	Sig.	,000

Table 3.6 shows the Total Variance Explained. There are 12 variables which Eigenvalues are greater than 1 in the table. These 12 variables explain the %75.95 of the total variables. These value should not be under %50 according to many sources.

Table 3. 6 Total Variance explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,855	10,707	10,707	3,855	10,707	10,707	3,334	9,260	9,260
2	3,643	10,120	20,827	3,643	10,120	20,827	2,999	8,330	17,590
3	3,221	8,947	29,775	3,221	8,947	29,775	2,927	8,131	25,722
4	2,950	8,194	37,969	2,950	8,194	37,969	2,879	7,997	33,719
5	2,650	7,361	45,329	2,650	7,361	45,329	2,872	7,978	41,696
6	2,447	6,797	52,126	2,447	6,797	52,126	2,508	6,967	48,664
7	1,937	5,379	57,505	1,937	5,379	57,505	2,286	6,351	55,015
8	1,819	5,053	62,558	1,819	5,053	62,558	2,123	5,896	60,911
9	1,373	3,815	66,373	1,373	3,815	66,373	1,446	4,018	64,929
10	1,226	3,406	69,778	1,226	3,406	69,778	1,413	3,926	68,855
11	1,160	3,224	73,002	1,160	3,224	73,002	1,298	3,605	72,459
12	1,061	2,948	75,950	1,061	2,948	75,950	1,257	3,491	75,950
13	,935	2,597	78,548						
14	,918	2,549	81,096						
15	,861	2,391	83,488						
16	,842	2,338	85,826						
17	,788	2,188	88,014						
18	,636	1,766	89,780						
19	,588	1,634	91,415						
20	,524	1,454	92,869						
21	,459	1,274	94,143						
22	,418	1,160	95,303						
23	,354	,982	96,285						
24	,270	,749	97,034						
25	,263	,732	97,766						
26	,184	,512	98,278						
27	,145	,404	98,682						
28	,114	,317	99,000						
29	,086	,240	99,239						
30	,081	,226	99,466						
31	,075	,207	99,673						
32	,052	,143	99,816						
33	,045	,126	99,942						
34	,010	,027	99,969						
35	,009	,024	99,993						
36	,002	,007	100,000						

Extraction Method: Principal Component Analysis.

As listed in Table 3.8 - Rotated Matrix, 12 factors are able to explain 40 variables. We can rename these 12 factor variables as below. It is obvious that Call related variables for Mins, Cost and Duration became one variable

Table 3. 7 Renamed Factor Variables

Fact-1 : PreCalls	Fact-7 : VASCalls
Fact-2 : IntCalls	Fact-8 : SMSCalls
Fact-3 : RoaCalls	Fact-9 : ActInv&PackID
Fact-4 : GSMCalls	Fact-10 : AddressCity&Region
Fact-5 : OGSMCalls	Fact-11 : AccLegth&Gender
Fact-6 : PSTNCalls	Fact-12 : Age&Type

Table 3. 8 Rotated Compenent Matrix

	Component											
	1	2	3	4	5	6	7	8	9	10	11	12
PREMins	,924											
PRECalls	,908											
PRECharge	,898											
ACTInvCostYTL												
Alarm												
INTCharge		,989										
INTMins		,987										
INTCalls		,976										
ROAMins			,992									
ROACharge			,978									
ROACalls			,957									
GSMMins				,940								
GSMCharge				,927								
GSMCalls				,788								
Cluster				-,562								
OGSMMins					,947							
OGSMCharge					,945							
OGSMCalls					,836							
PSTNMins						,927						
PSTNTCharge						,864						
PSTNCalls						,791						
VASCharge							,962					
VASCalls							,913					
VASMins							,703					
SMSCharge								,934				
SMSCalls								,926				
Reason												
ACTInvNum									,816			
PackID									,631			
AddressCity										,796		
Region										,678		
AccLength											,645	
Gender											,618	
Age												,723
Type												,601
Score												

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 6 iterations.

4 RESULTS & DISCUSSIONS

4.1 Used Methods

We considered 11 different methods which are BayesNet, NaïveBayes, MLP, RBFN, SMO, ADTrees, J48, NBTrees, Jrip, Part and Ridor to predict fraudsters.

a. BayesNet: Bayesian network is a form of probabilistic graphical model, also known as Bayesian belief network. A Bayesian network can be represented by a graph with probabilities attached. Thus a Bayesian network represents a set of variables together with a joint probability distribution with explicit independency assumptions.
b. NaiveBayes: is a simple probabilistic classifier based on applying Bayes' Theorem with strong independence assumptions. Depending on the precise nature of the probability model, naive Bayes classifier can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naïve Bayes models uses the method of maximum likelihood.
c. MLP (Multi-layer perceptron) : Multi-layer networks use a variety of learning techniques, the most popular being back-propagation. Here the output values are compared with the correct answer to compute the value of some predefined error-function. By various techniques the error is then fed back through the network. Using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the error function by small amount. After repeating this process for a sufficiently large number of training cycles the network will usually converge to some state where the

error of the calculations is small. In this case one says that the network has learned a certain target function.

d. RBFN : Radial Basis Functions are powerful techniques for interpolation in multidimensional space. A RBF is a function which has built into a distances criterion with respect to a centre. Radial basis functions have been applied in the area of neural networks where they may be used as a replacement for the sigmoidal hidden layer transfer characteristics in multi-layer perceptrons.

e. SMO: The self-organizing map is a subtype of artificial neural networks. It is trained using unsupervised learning to produce low dimensional representation of the training samples while preserving the topological properties of the input space. This makes SOM especially good for visualizing high-dimensional data. The model was first described by the Finnish professor Teuvo Kohonen and is thus sometimes referred to as a Kohonen map.

f. ADTrees : All Dimension Trees are data structures used to accelerate conjunctive counting queries on a data set. AD Trees can be used for generating alternating decision trees. The number of boosting iterations needs to be manually tuned to suit the dataset and the desired complexity/accuracy tradeoff. Induction of the trees has been optimized and heuristic search methods have been introduced to speed learning.

g. J4.8 : Decision trees represent a supervised approach to classification. A decision tree is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes. J.R. Quinlan has popularized the decision tree approach with his research. The

latest public domain implementation of Quinlan's model is C4.5. The Weka classifier package has its own version of C4.5 known as J48.

h. NBTrees : Naïve Bayes Tree uses decision tree as the general structure and deploys naive Bayesian classifiers at the leaves. The intuition behind it is that naïve Bayesian classifier work better than decision trees when the sample data set is small. Therefore, after several attribute splits when constructing a decision tree, it is better to use naive Bayes classifier at the leaves than to continue splitting the attributes. Naïve Bayes Tree is used to improve classification accuracy and under the area curve.

i. Jrip: This class implements a propositional rule learner, Repeated Incremental Pruning to produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP.

j. Part : is a rule generator that uses J48 to generate pruned decision trees from which rules are extracted. Part is applicable for categorical classification and prediction. Input attributes can be categorical and numerical.

k. Ridor : is the implementation of a Ripple Down Rule learner. It generates a default rule first and then the exceptions for the default rule with the least (weighted) error rate. Then it generates the "best" exceptions for each exception and iterates until prune. Thus it performs a tree-like expansion of exceptions. The exceptions are a set of rules that predict classes other than the default.

4.2 Performance Measure

In order to evaluate the results, we have used the below metrics.

Accuracy (Correctness) = (TP + TN) / (TP + FN + FP + TN)

Sensitivity (True positive Rate) = TP / (TP+FN)

Specificity (True Negative Rate) = TN / (FP+TN)

Precision (Selectivity) = TP / (TP + FP)

$$\text{RMSE}= \sqrt{\frac{(p_1-a_1)^2+....+(p_n-a_n)^2}{n}}$$

4.3 Using Weka with Standard Variables

From the set of variables, some of the fields are ignored in order to reduce computational burden of the analysis. Gain Ratio Attribute Evaluator which is a supervised filter, and ranker search which ranks the attributes by their individual evaluators are used to select the attributes. The factors in table 4.1 are assumed to have the highest contribution to the ultimate decision about the subscriber.

Table 4. 1 Ranked Attributes

0.437	AccLength
0.222	PackID
0.191	TotMins
0.158	Type
0.156	OGSMCalls

Table 4.2 and Table 4.3 shows the Accuracy, and the misclassification matrix which are TP(True Positive), FN(False Negative), FP(False positive) and TN(True Negative) values both for Training set and Test set.

Table 4. 2 Misclassification Matrix for the Training Set

		Training Set (374)			
Method	Accuracy	TP	FN	FP	TN
BayesNet	0,88	161	26	20	167
NaiveBayes	0,86	134	53	1	186
MLP	0,83	125	62	0	187
RBFN	0,83	137	50	15	172
SMO	0,81	126	61	10	177
ADTree	0,94	170	17	4	183
J48	0,91	152	35	0	187
NBTree	0,94	168	19	3	184
Jrip	0,94	169	18	5	182
Part	0,91	152	35	0	187
Ridor	0,93	160	27	0	187

Table 4. 3 Misclassification Matrix for the Test Set

		Test Set (186)			
Method	Accuracy	TP	FN	FP	TN
BayesNet	0,92	82	11	3	90
NaiveBayes	0,91	76	17	0	93
MLP	0,81	67	26	10	83
RBFN	0,89	74	19	2	91
SMO	0,83	67	26	5	88
ADTree	0,91	81	12	5	88
J48	0,92	79	14	0	93
NBTree	0,93	84	9	4	89
Jrip	0,89	81	12	8	85
Part	0,92	79	14	0	93
Ridor	0,93	80	13	0	93

Table 4.4 and Table 4.5 show the Accuracy, Sensitivity, Specificity, Precision and RMSE results for the methods used for both Training and Test sets.

As listed Table 4.3, for the Training set, Decision Trees (AdTrees, NBTrees, J48) and Rule based methods (Jrip, Ridor, Part) produced good results. ADTree is the best method with minimal RMSE which is 0.21, Precision which is 0.98 and Accuracy which 0.94.

Table 4. 4 Training Results for the methods used

		Training Set (374)			
Method	Accuracy	Sensitivity	Specificity	Precision	RMSE
ADTree	**0,94**	**0,91**	**0,98**	**0,98**	**0,21**
NBTree	0,94	0,90	0,98	0,98	0,22
Jrip	0,94	0,90	0,97	0,97	0,23
Ridor	0,93	0,86	1,00	1,00	0,27
J48	0,91	0,81	1,00	1,00	0,28
Part	0,91	0,81	1,0	1,00	0,27
BayesNet	0,88	0,86	0,89	0,89	0,29
NaiveBayes	0,86	0,72	0,99	0,99	0,35
MLP	0,83	0,67	1,00	1,00	0,34
RBFN	0,83	0,73	0,92	0,90	0,35
SMO	0,81	0,67	0,95	0,93	0,44

As listed Table 4.5, for the Test set, Decision Trees and Rule Based methods gave good results again with an addition of Bayes Net method. Neural network methods (MLP,RBFN,SOM) didin‘t give good results both for training and test sets. NBTree is the best method with minimal RMSE which is 0.25, Precision which is 0.95, and the Accuracy which is 0.93. (Lower RMSE systems tend to make incorrect classification less than the others).

Table 4. 5 Testing Results for the methods used

		Test Set (186)			
Method	Accuracy	Sensitivity	Specificity	Precision	RMSE
NBTree	**0,93**	**0,90**	**0,96**	**0,95**	**0,25**
Ridor	0,93	0,86	1,00	1,00	0,26
BayesNet	0,92	0,88	0,97	0,96	0,26
J48	0,92	0,85	1,00	1,00	0,26
Part	0,92	0,85	1,00	1,00	0,27
NaiveBayes	0,91	0,82	1,00	1,00	0,28
ADTree	0,91	0,87	0,95	0,94	0,25
Jrip	0,89	0,87	0,91	0,91	0,31
RBFN	0,89	0,80	0,98	0,97	0,31
SMO	0,83	0,72	0,95	0,93	0,41
MLP	0,81	0,72	0,89	0,87	0,39

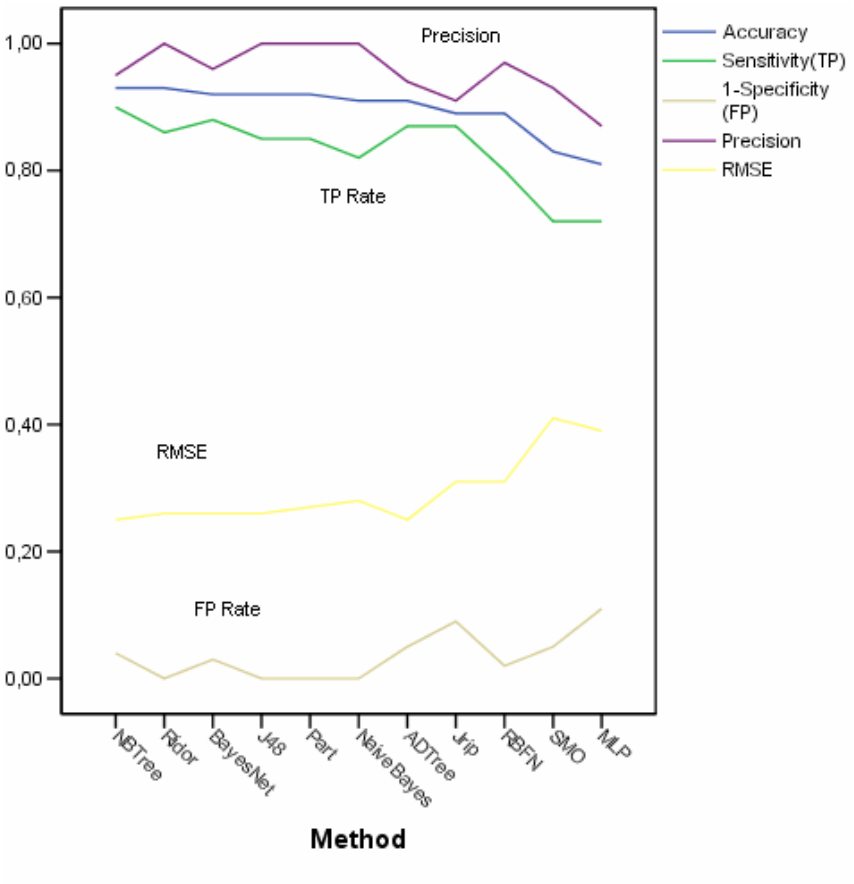

Figure 4. 1 Sequence graph of the methods for the Test Set for standard variables

The ROC Curve in Figure 4.2 shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity) for Training set with NBTree, Ridor and BayesNet.

The ROC Curve plots sensitivity vs. 1-Specificiy, or TP(True positive) rate vs. FP(False positive) rate. The closer the curve follows the left hand border and then the top border of the ROC spaces, the more accurate the test. The closer the curve comes to the 45 degree diagonal of the ROC space, the less accurate the test. We can also see that the best methods in order as NBTree, Ridor and BayesNet from the ROC curve.

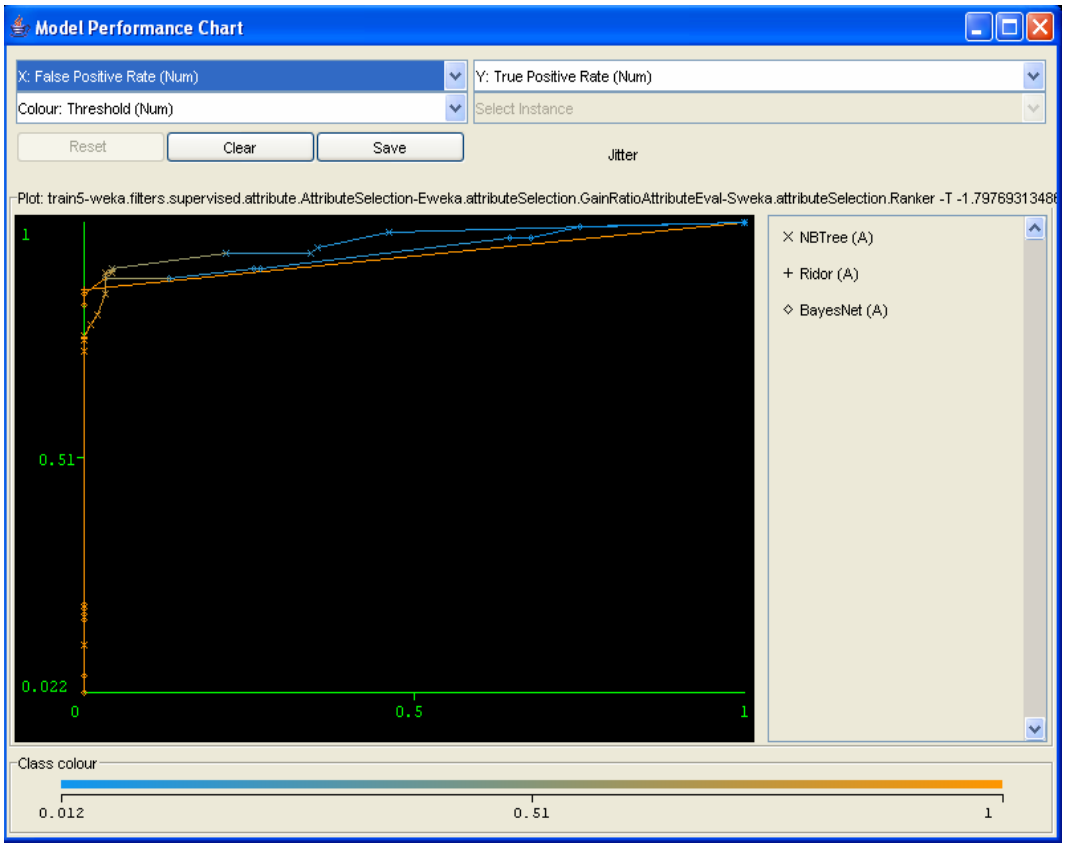

Figure 4. 2 ROC curve for Test Set with NBTree, Ridor and BayesNet

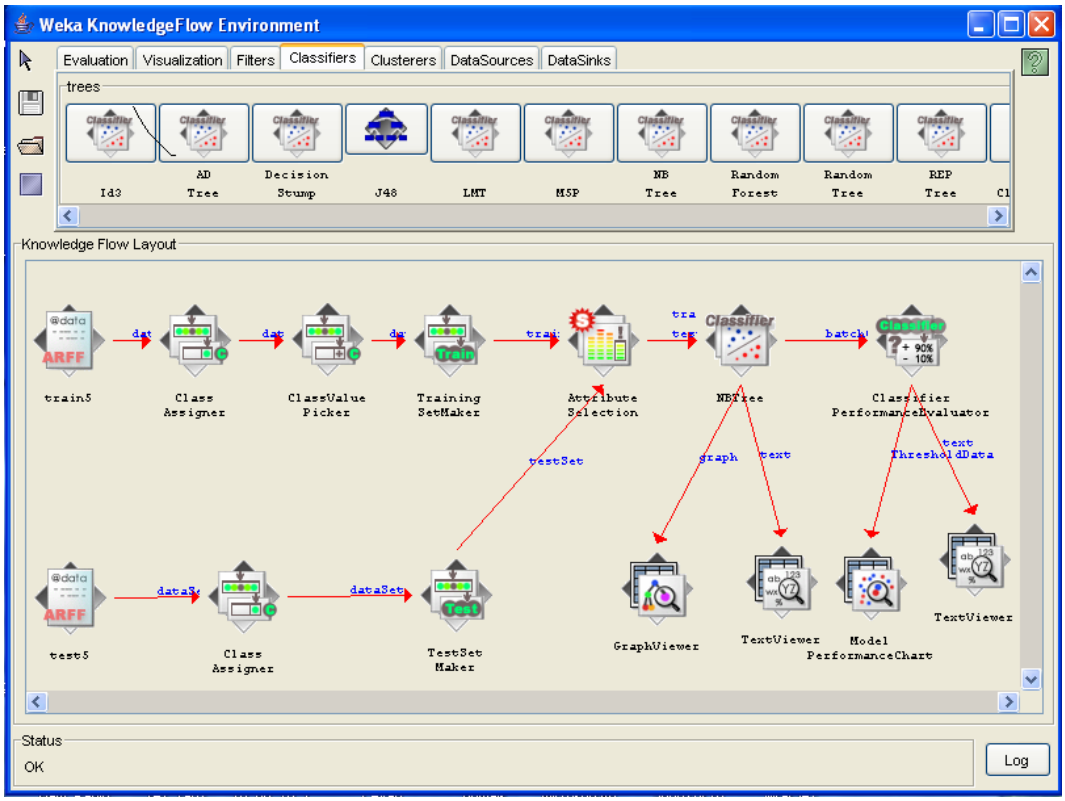

Figure 4. 3 Weka knowledgeFlow for NBTree method for Test Set

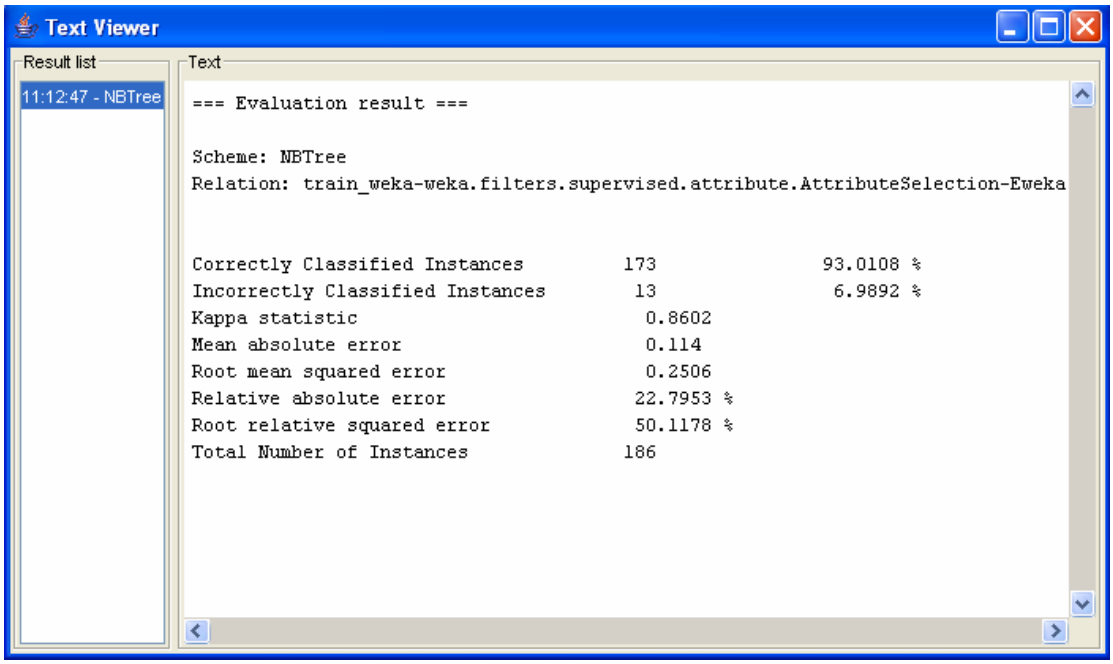

Figure 4. 4 Results for NBTree for Test set

4.4 Using Weka with Variables obtained from Factor Analysis

Table 4.6 and Table 4.7 shows the Accuracy, and the misclassification matrix which are TP(True Positive), FN(False Negative), FP(False positive) and TN(True Negative) values both for Training set and Test set.

Table 4. 6 Misclassification Matrix for the Training Set

		Training Set (374)			
Method	Accuracy	TP	FN	FP	TN
BayesNet	0,74	134	53	45	142
NaiveBayes	0,60	54	133	16	171
MLP	0,83	151	36	26	161
RBFN	0,68	99	88	32	155
SMO	0,78	144	43	40	147
ADTree	0,82	162	25	44	143
J48	0,79	122	65	13	174
NBTree	0,84	162	25	36	151
Jrip	0,80	149	38	38	149
Part	0,80	115	72	1	186
Ridor	0,86	152	35	17	170

Table 4. 7 Misclassification Matrix for the Test Set

		Test Set (186)			
Method	Accuracy	TP	FN	FP	TN
BayesNet	0,72	70	23	29	64
NaiveBayes	0,64	34	59	8	85
MLP	0,81	71	22	13	80
RBFN	0,70	48	45	11	82
SMO	0,81	69	24	12	81
ADTree	0,74	69	24	24	69
J48	0,70	56	37	19	74
NBTree	0,74	67	26	23	70
Jrip	0,77	69	24	19	74
Part	0,69	48	45	13	80
Ridor	0,70	58	35	21	72

As listed Table 4.8, for the Training set, Ridor, NBTree, and MLP methods produced good results. Ridor as a rule based method is the best method with minimal RMSE which is 0.37, precision 0.90 and the accuracy which is 0.86.

Table 4. 8 Training results for methods used

		Training Set (374)			
Method	Accuracy	Sensitivity	Specificity	Precision	RMSE
Ridor	**0,86**	**0,81**	**0,91**	**0,90**	**0,37**
NBTree	0,84	0,87	0,81	0,82	0,34
MLP	0,83	0,81	0,86	0,85	0,35
ADTree	0,82	0,87	0,76	0,79	0,38
Part	0,80	0,61	0,99	0,99	0,36
Jrip	0,80	0,80	0,80	0,80	0,39
J48	0,79	0,65	0,93	0,90	0,37
SMO	0,78	0,77	0,79	0,78	0,47
BayesNet	0,74	0,72	0,76	0,75	0,41
RBFN	0,68	0,53	0,83	0,76	0,45
NaiveBayes	0,60	0,29	0,91	0,77	0,57

Table 4. 9 Test Set results for methods used

		Test Set (186)			
Method	Accuracy	Sensitivity	Specificity	Precision	RMSE
MLP	**0,81**	**0,76**	**0,86**	**0,85**	**0,39**
SMO	0,81	0,74	0,87	0,85	0,44
Jrip	0,77	0,74	0,80	0,78	0,43
ADTree	0,74	0,74	0,74	0,74	0,43
NBTree	0,74	0,72	0,75	0,74	0,43
BayesNet	0,72	0,75	0,69	0,71	0,43
RBFN	0,70	0,52	0,88	0,81	0,45
J48	0,70	0,60	0,80	0,75	0,47
Ridor	0,70	0,62	0,77	0,73	0,55
Part	0,69	0,52	0,86	0,79	0,47
NaiveBayes	0,64	0,37	0,91	0,81	0,54

As listed Table 4.9, for the Test set, MLP, SMO (Neural network) and Jrip (Rule based) and ADtree, NNTree(Decision Trees) produced the best results.

MLP is the best method with minimal RMSE which is 0.39, Precision which is 0.85, and the Accuracy which is 0.81. (Since lower RMSE systems tend to make incorrect classification less than the others).

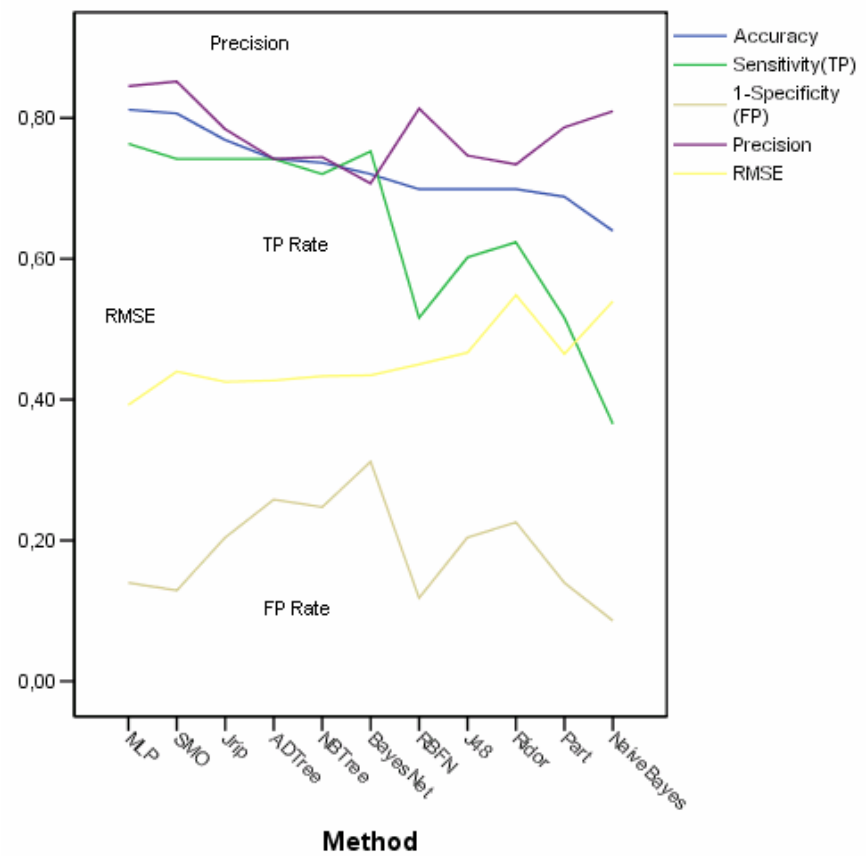

Figure 4. 5 Sequence graph of the methods for the Test Set

Figure 4.6 and Figure 4.7 shows the ROC curve for Test Set both for normal subscribers and fraudulent subscribers. The closer the curve follows the left hand border and then the top border of the ROC spaces, the more accurate the test. The closer the curve comes to the 45 degree diagonal of the ROC space, the less accurate the test

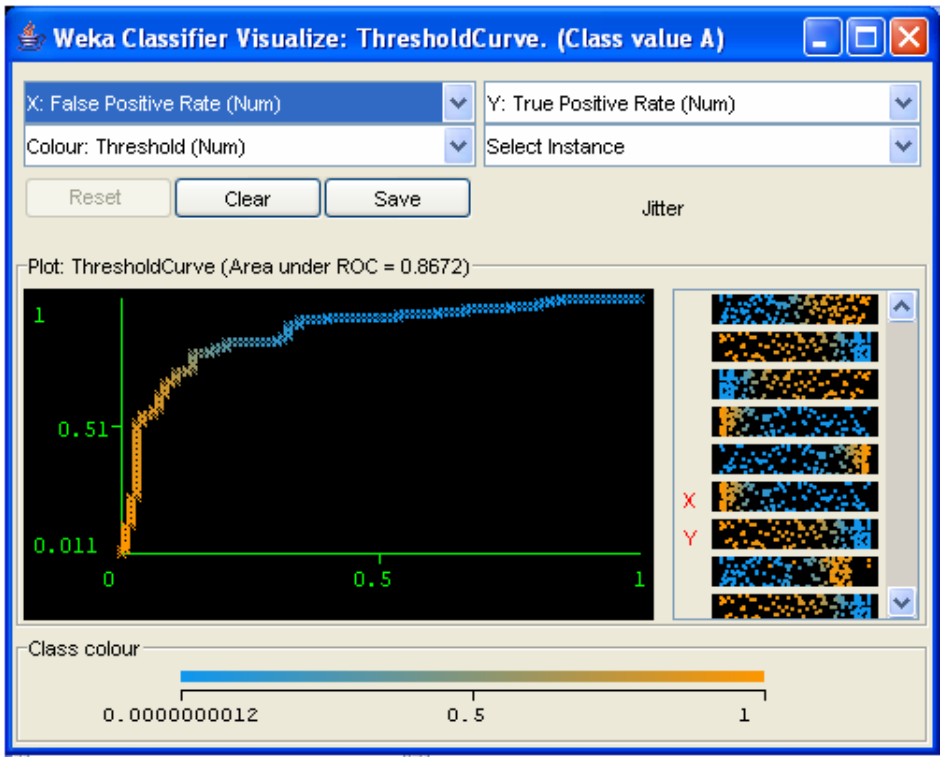

Figure 4. 6 ROC curve for Test Set with MLP for Class A (Normal Subscriber)

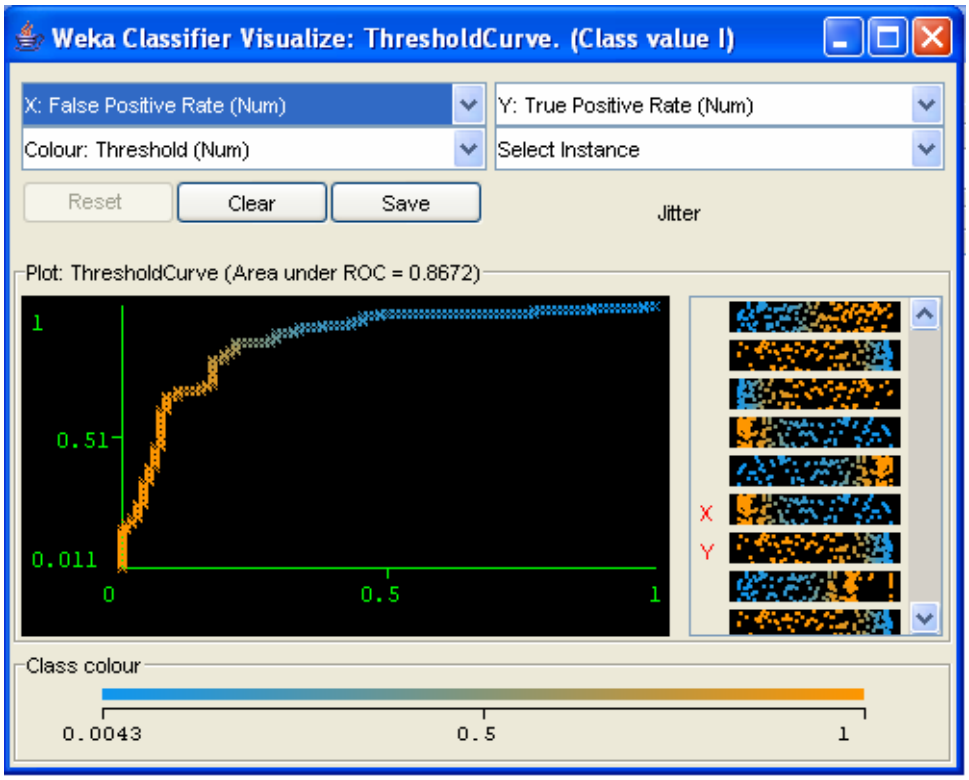

Figure 4. 7 ROC curve for Test Set with MLP for Class I (Fraudulent Subscriber)

5 CONCLUSION AND FUTURE DIRECTION

In section 4.3, we used standard variables with ranked attributes -which are assumed to have the highest contribution to the ultimate decision about the subscriber- We had good results mainly with Decision trees and Rule based methods.

In section 4.4, the variables obtained from factor analysis were used and we had good results with Neural networks and Rule based methods. But the results were not as good as results of standard variables with ranked attributes.

We prefer to use the results obtained from Standard variables. Since we have ranked the attributes with Gain Attribute Evaluator in Weka and only used the attributes with higher impact on the outcome, there were not any correlated attributes anymore.

Having determined that the model is reasonably good, we can then apply the model to other data files containing similar attributes. This process is often referred to as scoring which effectively can assist telecom service providers to detect fraudulent use more accurately.

It is well known that the pattern as well as the levels and costs of fraud change very quickly in time. Because of this complexity, any fraud system could become rapidly obsolete.

In the future, the number of instances in the training and test sets should be increased. The number of the attributes can be enhanced by adding information about used services, separation of the day, evening, night call, minutes and charges etc.

REFERENCESS

ACTS AC095(1996), project ASPeCT : Definition of Fraud Detection Concepts. ACTS ASPeCT, AC095/KUL/W22/DS/P/06/A

ACTS AC095(1997), project ASPeCT : Fraud Management Tool : evaluation report, ACTS ASPeCT, AC095/SAG/W22/DS/P/13/2

Bolton, R.,& Hand, D.(2002). Statistical Fraud Detection: A review. Statistical Science 17(3), 235-249

Bolton, R.,& Hand, D.(2002). Unsupervised Profiling Methods for Fraud Detection

Burge, P., & Shawe-Taylor J. (2001). An supervised Neural Network Approach to profiling the behaviour of Mobile Phone Users..

Burge, P., Shawe-Taylor J. (1997). Detecting Cellular Fraud Using Adaptive Protypes

Burge, P., Shawe-Taylor J.,Moreau Y., Verrelst, H., Störmann C. & Gosset, P. (1997) BRUTUS- A Hybrid Detection Tool.

Burge, P., Shawe-Taylor, J., Detecting Cellular Fraud Using Adaptive Protypes

Cahill, H. C,, Lambert, D., Pinheiro, J. C., Sun, D. X., (2000). Detecting Fruad in the Real World

Collins, M. (1999a). Telecommunications Crime Part 1. Computer& Security, 18, 577-586

Collins, M. (1999b). Telecommunications Crime Part 2. Computer& Security, 18, 683-692

Collins, M. (2000). Telecommunications Crime Part 3. Computer& Security, 19, 141-148

Estevez, P. A., Held, C. M., Perez, C. A. (2005) Subscription Fraud Prevention in Telecommunicaitons Using Fuzzy Rules and ....

Fawcett, T. & Provost, F. (1997). Adaptive Fraud Detection. Data Mining and Knowledge Discovery. 1(3):291-316

FML., (2003). FML revenue assurance and fraud management yearbook

Hand, D., Mannila, H., Smyth, P., Principles of Data Mining, The MIT Press

Hollmen, J. & Tresp, V (1998). Call Based Fraud Detection in Mobile Communication Networks Using a Hierarchical Regime-Switching Model

Karahoca, A., Sanver, M., Yalcin, S., Churn Management Using an Adaptive Neuro-Fuzzy Inference System

Karahoca, A.,(2004) Data Mining via Cellular Neural Networks in the GSM sector, Proceedings of 8th IASTED international Conference: Software Engineering Applications

Karahoca, A.,(2003) Strategic Data Mining by Neural Networks in the Telecommunication Sector, International XII. Turkish Symposium on Artificial Intelligence and Neural Networks

Larose D. T., Discovering Knowledge in Data : An Introduction to Data Mining, Wiley-Interscience

MathWorks Inc., Matlab 6.5 R13 Statistical Toolbox Help.pdf, June 2002

Moreau, Y. & Vandewalle, J (1997). Detection of Mobile Phone Fraud Using Supervised Neural Networks

Moreau, Y., Preneel, B., Burge, P., Shawe-Taylor, J., Stoerman., C., Cooke, C.,(1997) Novel Techniques for Fraud Detection in Mobile Telecommunication Network, ACTS Mobile Summit, Granada, Spain

Phua C., Lee, V., Smith, K., & Gayler, R.(2005). A Comprehensive Survey of Data Mining-based fraud detection

Shawe-Taylor, J., Howker, K.,&Burge, P.(1999). Detection of Fraud in Mobile Telecommunications

SPSS statistical software, SPSS Clementine data mining software suit, www.spss.com

Taniguchi, M., Haft, M., Hollmen, J. & Tresp, V. (1998) Fraud Detection in Communications Networks using Neural and Probalistic Methods

Two Crows Corporation (1999). Introduction to Data Mining and Knowledge Discovery, Third Edition

Weka, Waikato Environment for Knowledge Analysis, www.cs.waikato.ac.nz/ml/weka

Wieland, K. (2004). The last Taboo? Revenue leakage continues to hamper the telecom industry. Telecommunications (International Edition), 38, 10-11

Witten, I. H., Frank E., Data Mining: Practical Machine Learning Tools and Technologies with Java implementations, Morgan Kaufmann Publishers

VITA

Bülent Kuşaksızoğlu was born in Adapazarı. He received his B.Sc. degree in Mathematics Engineering with emphasis in System Analysis and M.Sc. degree in Management Engineering from the Technical University of Istanbul in 1982 and 1984 respectively. Since then he has worked as programmer and as IT manager in various companies. His main areas of interest are data warehouse, data mining and neural networks.