DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DECISION SUPPORT SYSTEM FOR A CUSTOMER RELATIONSHIP MANAGEMENT CASE STUDY

by

Özge KART

January, 2013

İZMİR

DECISION SUPPORT SYSTEM FOR A CUSTOMER RELATIONSHIP MANAGEMENT CASE STUDY

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Science
in Computer Engineering, Computer Engineering Program

by

Özge KART

January, 2013

İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"DECISION SUPPORT SYSTEM FOR A CUSTOMER RELATIONSHIP MANAGEMENT CASE STUDY"** completed by **ÖZGE KART** under supervision of **PROF. DR. ALP KUT** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Alp Kut

Supervisor

Y. Doç. Dr. Reyat YILMAZ

(Jury Member)

Yrd. Doç. Dr. Sevil ŞEN

(Jury Member)

Prof.Dr. Mustafa SABUNCU

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to thank to my thesis advisor Prof. Dr. Alp Kut for his help, suggestions and guidance.

I also thank to my family and my sincere friends and Aslan Türk for their patience and support.

Özge KART

DECISION SUPPORT SYSTEM FOR A CUSTOMER RELATIONSHIP MANAGEMENT CASE STUDY

ABSTRACT

Data mining which has become very common and gained importance recently, is a tool which provides to discover hidden and valuable information in large datasets. One of the widely used areas of data mining is Customer Relationship Management (CRM). CRM is an approach used to understand customer's behaviors and increase the customer satisfaction.

The aim of this study, researching the data minig and CRM concepts which has become widespread and important in recent years, in addition to applying a data mining model to banking sector as an example and representing the result in a mobile platform.

This study has built a classifier using the naive Bayesian classification. The accuracy rate of the model is determined by doing cross validation. The results demonstrated the applicability and effectiveness of the proposed model. Naive Bayesian classifier reported high acceptable accuracy. So the classification rules can be used to support decision making for achieving a good CRM for businesses.

Keywords: Data mining, customer relationship management, CRM, mobile

MÜŞTERİ İLİŞKİLERİ YÖNETİMİ İÇİN KARAR DESTEK SİSTEMİ OLUŞTURULMASI

ÖZ

Son zamanlarda oldukça yaygınlaşan ve önem kazanan veri madenciliği, büyük verilerin içerisinde gizli bulunan değerli bilgilerin keşfedilmesini sağlayan bir araçtır. Veri madenciliğinin yaygın kullanım alanlarından biri Müşteri İlişkileri Yönetimi (CRM)'dir. CRM, müşterinin davranışlarını anlamak ve müşteri memnuniyetini artırmak için kullanılan bir yaklaşımdır.

Bu çalışmanın amacı, son yıllarda çok yaygın ve önemli hale gelen veri madenciliği ve CRM kavramlarını incelemek, bunun yanında bir veri madenciliği modelini örnek olarak bankacılık sektöründe uygulamak ve sonucunu da bir mobil platform üzerinden görüntülemektir.

Bu çalışmada naive Bayes sınıflandırma algoritması kullanarak bir sınıflandırıcı oluşturulmuştur. Oluşturulan modelin doğruluk oranını saptamak için çapraz doğrulama tekniğinden yararlanılmıştır. Sonuçlar oluşturulan modelin etkin ve uygulanabilir bir model olduğunu göstermektedir. Naive Bayes sınıflandırıcı yüksek ve kabul edilebilir bir doğruluk göstermiştir. Bu yüzden bu sınıflandırma kuralları işletmelerin iyi bir müşteri ilişkileri yönetimini başarabilmeleri için bir karar destek sistemi oluşturmaktadır.

Anahtar sözcükler: Veri madenciliği, müşteri ilişkileri yönetimi, CRM, mobil

CONTENTS

Page

CHAPTER ONE
INTRODUCTION

Today, with rapid advances in technology, relationships between companies and customers differentiated as well as relationships among people.

In this new era, especially for companies, competition has become more intense, geographical boundaries have lost the importance. Companies have begun to give more attention to customers' personal preferences. One to one marketing strategies began to come to the fore. Under these conditions, the companies could not guarantee the loyalty of customers and needed to be closer to their customers. This caused rising of Customer Relationship Management (CRM) concept.

Businesses need to use some tools when applying this management form. One of these tools is Data Mining. By using Data Mining tools, businesses learn customers' consumption behaviors, spending patterns and use these information for future decisions and strategies.

The main purpose of this thesis, with the help of Data Mining algorithms, extracting hidden information that companies can use in decision making process, from databases of companies. Then, accessing these valuable information via a WCF service and presenting on a mobile platform.

In the second chapter CRM is defined, the CRM process and architecture are explained. Benefits of CRM are discussed.

In the third chapter, the concept of Data Mining, Data Mining Process, widely used areas of Data mining (especially CRM), the models and techniques used to perform data analysis, are explained in detail.

In the fourth chapter, Bayesian classification, which is the basis of the application, is explained.

In the fifth chapter, technologies used in this application are explained in detail. WCF service, Android and Phonegap technologies are presented.

In the last part of thesis, Bayesian Classification is implemented on a dataset obtained from a banking institution. A decision support system is generated to help the instutituon to predict the behavior of a new customer. This prediction is presented on a mobile platform “Android”. The decision support system is accessed via WCF service.

1.1 Literature Review

In the process of economic development, business management concept evolved from product-oriented idea to market-oriented idea, and then to customer-oriented idea. Whether enterprises can obtain, maintain and develop their own clients or not has become the most critical factor because customers are important strategic resources. Customer Relationship Management(CRM) is based on understanding customers and it can enable businesses to provide the customers with more personalized and more efficient services, and then it can improve customer satisfaction and loyalty, and increase the competitiveness of businesses finally. Customer segmentation is classifying the customers according to the customer’s attributes, behavior, needs, preferences, value and other features in a clear business strategy and specific market. Customer segmentation can provide appropriate products, services and marketing models to the customers.

A. Data Mining and CRM

Data mining can be used for both classification and regression problems. In classification problems it is predicted what category something falls into – for

example, whether or not a person is a good credit risk or which of several offers someone is most likely to accept. In regression problems, a number is predicted, such as the probability that a person will respond to an offer. In CRM, data mining is frequently used to assign a score to a particular customer or prospect indicating the likelihood that the individual behaves the way you want (Gao,2011)(Chung, Gray, 1999). For example, a score could measure the tendency to respond to a particular offer or to switch to a competitor's product. It is also frequently used to identify a set of characteristics (called a profile) that segments customers into groups with similar behaviors, such as buying a particular product. A special type of classification can recommend items based on similar interests held by groups of customers.

Data Mining is an important step in the Knowledge Discovery in Database (KDD) process that consists of applying data analysis and knowledge discovery algorithms to produce useful patterns (or rules) over the datasets. Data mining techniques can filtrate and classify customer resources of insurance, divide credit customers into several grades, to predict the customer risk, thus investigating customer material of the low forecasted degrees of comparison can avoid deceiving policy effectively, and avoid service risk. (Viveros, 1996) addressed the effectiveness of two data mining techniques in analyzing and retrieving unknown behavior patterns from gigabytes of data collected in the health insurance industry.

Wu et al (2005) presented that KDD/Data mining is utilized to explore decision rule to investigate the potential customers for an existing insurance product. Kanwal garg (2008) presented decision tree method for identifying customer behaviour of investment in life insurance sector. Rivers (2010) examined some of the benefits and challenges of using data mining processes within the healthcare arena. E.W.T.Ngai, L. Xiu, D.C.K.Chau, (2009) analysis provides roadmap to guide future research and facilitate knowledge accumulation and creation concerning the application of data mining techniques in CRM.

Zhiyuan yao, Annika H.Holombom, Tomas Eklund and Barbro Back (2010) found that combined approach of SOM-Ward clustering and decision trees provide

more detailed information about customer base for tailoring actionable marketing strategies.Abrahams et.al (2009) used decision trees to create a marketing strategy for a pet insurance company.Anna Jurek considered application of Naïve Bayes model for evaluation of risk connected with Life insurance of customers. M.Staudt,J(1998) reports on a project initiated at swiss Life for mining its data resources from the life insurance business.it lies on establishing comfortable data preprocessing support for normalised relational databases and on the management of meta data. Young Moon Cha,et al(2001) examined the characteristics of the knowledge discovery and data mining algorithms to demonstrate how they can be used to predict health outcomes and provide policy information for hypertension management using the Korea Medical Insurance Corporation database. Young Moon Cha,et al(2004) examined characteristics of the mining time dependent patterns to demonstrate how they can be used to predict hypertension outcomes and provide lifestyle information in order to prevent hypertension using data mining approaches.

B. Baysian Classification

Bayesian classification is based on Bayes theorem. Studies about comparing classification algorithms have found a simple Bayesian classifier which is known as the naive Bayesian classifier to have higher performance comparing to decision and neural network classifiers. Bayesian classifiers have also showed high speed and accuracy when applied to databases.

The naive Bayesian classifier works as follows, as in (Kuykendall, 1999):

1) Each data sample is represented by an n-dimensional feature vector X = (x1, x2 ,..., xn) , depicting n measurements made on the sample from n attributes, respectively, A1, A2 ,..., An .

2. Suppose that there are m classes, C1,C2 ,...,Cm . Given an unknown data sample, X (having no class label) ,the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X .

That is, the naïve Bayesian classifier assigns an unknown sample X to the class Ci if and only if P(Ci X) > P(C j X) for $1 \le j \le m, j \ne i$.
Thus maximize P(Ci X) . The class Ci for which P(Ci X) is maximized is called the maximum posterior hypothesis. By Bayes theorem,

P(Ci/X)=P(X/Ci)P(Ci) / P(X)

3. As P(X) is constant for all classes, only P(X Ci)P(Ci) need to be maximized. If the class prior probabilities are not known, then it is assumed that the classes are equally likely, that is, P(C1) = P(C2) = ... = P(Cm) .

4. Given data sets with many attributes, it would be extremely computationally expensive to compute P(X Ci). In order to reduce computation in evaluating P(X Ci), the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample, that is, there are nodependence relationships among the attributes. Unknown sample using naive Bayesian classification, given the sample training data as Figure 2.1. The class label attribute, creditcard_proposing has two distinct values(namely,{yes, no}).

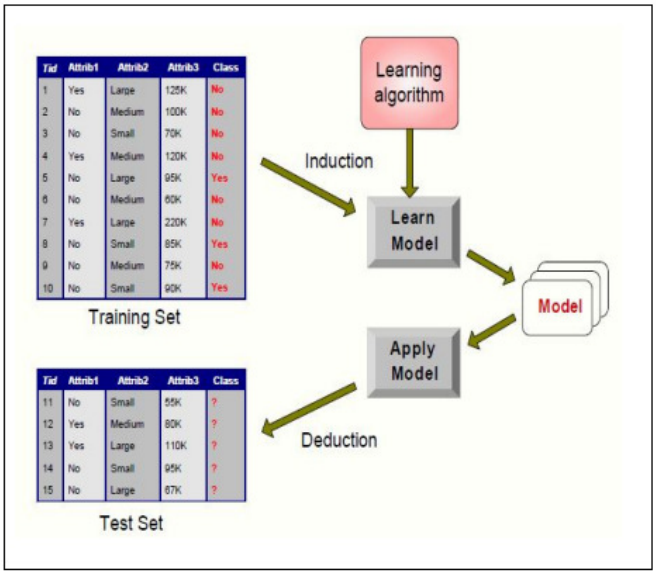


Figure 2.1 The architecture of CRM based on data mining

In previous studies, Naïve Bayesian classification assumes that the effect of an attribute value on a given class is independent of from other attribute values (Da, Hai-guang, Jian-he, 2010). When the assumption is true, the Bayesian classification is the most accurate in comparison with all other classification methods.

In last years, many researchers have made achievements in the application of Bayesian Classification. For example, Huang (2009) introduced a construction method for university courses relationship Bayesian networks; Li (2009) constructed a Bayesian classification model for forecasting values of credit card customers; Leng (2011) made a learning dynamic Bayesian network structure based on the basic dependency relationship between variables and dependency analysis method.

CHAPTER TWO
CUSTOMER RELATIONSHIP MANAGEMENT (CRM)

2.1 What is CRM?

The term customer relationship management (CRM) is used just since 1990s. It emerged when some small- scale financial institutions in USA and England started to attract customers of large-scale banks by producing customized solutions. Some of Customer Relationship Management definitions are as follows:

- "CRM is an information industry term for methodologies, software and usually Internet capabilities that help an enterprise manage customer relationships in an organized way."(Berry, n.d)
- "CRM is the process of managing all aspects of interaction a company has with its customers, including prospecting, sales and service. CRM applications attempt to provide insight into and improve the company/customer relationship by combining all these views of customer interaction into one picture." (Sreedhar, Manthan, Ajay, Virendra, & Udupa, 2007)
- "CRM is an integrated information system that is used to plan, schedule and control the pre-sales and post-sales activities in an organization. CRM embraces all aspects of dealing with prospects and customers, including the call centre, sales-force, marketing, technical support and field service. The primary goal of CRM is to improve long-term growth and profitability through a better understanding of customer behaviour. CRM aims to provide more effective feedback and improved integration to better gauge the return on investment (ROI) in these areas." (Petersen,2003)
- "CRM is a business strategy that maximizes profitability, revenue and customer satisfaction by organizing around customer segments, fostering behaviour that satisfies customers and implementing customer centric processes." (Shaw, 2002)

In last few years, CRM became a new concept of business with spread of technology which developed within the framework of customer-orinted approach. In addition, the fast spread of Internet and related technologies raised the chance for marketing and made the relationships between companies and their customers are steerable.

Reasons for the emergence of the concept of CRM:

- Mass marketing is an increasingly expensive way to win a customer
- Increase in the importance of customer satisfaction and customer loyalty concepts
- Understanding the value of your existing customers and need for effort for the customer retention
- Along with the importance of marketing for individuals, need for the developing strategies to provide specific solutions to each customer
- Intense competition
- Developments in communication technologies and database management systems

Companies need more information about their customers since CRM has become important. In order to increase the effectiveness of CRM applications, they need to more information about their customers' charasteristics, behaviours, which product they need in which circumstances. It can be said that companies which understand and analyse their customers well, will be successful in application of CRM techniques. Companies that establish a CRM application for a bad analyzed customer portfolyo may cause increase in costs as well as customer dissatisfaction. If customer relationship management is defined as segmenting customers according to their similar characteristics, and managing this segments so as to increase the company's long term profit potential and the acquisation of customers, effective data analysis is the best way to achieve them.

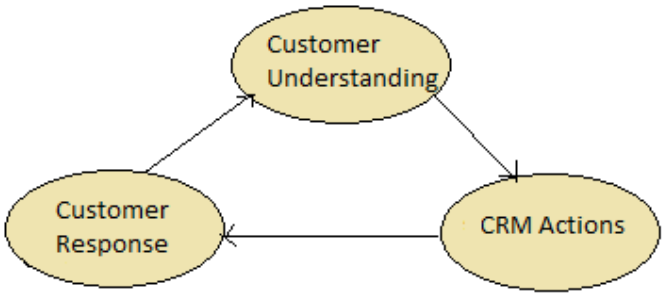


Figure 2.1 The CRM Cycle

CRM has following stages: (Swift, 2001; Parvatiyar & Sheth, 2001; Kracklauer, Mills, & Seifert, 2004)

1- Customer Identification
2- Customer Attraction
3- Customer Retention
4- Customer Development

These stages give rise to understand customers deeply to increase customer value to the company in the long term.

1- Customer Identification: The first stage of CRM is customer identification. The main goal of this stage is finding an answer to the question "Who is the most profitable customer?" To answer this questions following studies should be carried out:
   - Customer segmentation: segmenting all customers into smaller groups, according to their similarities.
   - Target customer analyzing: searching for the beneficial segments of customers by analyzing the customers' specific features.

   In addition, Customer Identification consists of finding customers who may be lost to the competition and how they can be won again (Kracklauer, Mills & Seifert, 2004).

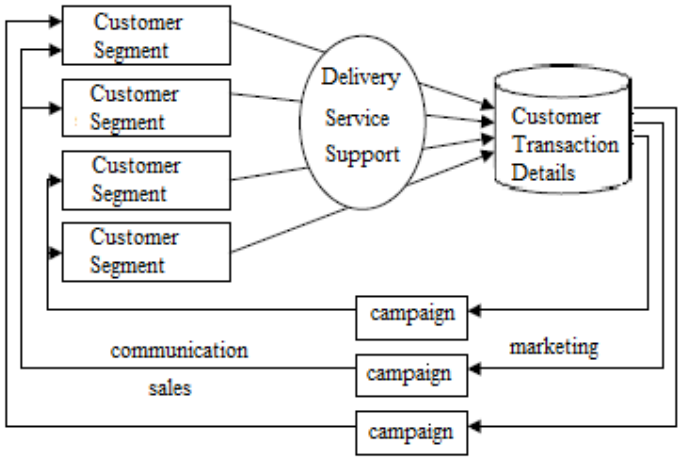


Figure 2.2 Campaign Management for Customer Segments

As seen in the Figure 2.2, companies obtain valuable data from transactions of customers such as purchasing through delivery, service or support units. Marketing units which evaluates this data from hundreds of thousands customers have chance to organize campaigns special to each customer segments. Aim of organizing campaigns in this way is supplying product or service that each customer segment purchases most likely on time.

2- Customer Attraction: The main purpose of this stage is to sale. The following steps must be taken in order to find the answer to the question "How to make specific sales to the target customer segments with the most effective way?"
   - Requirement analysis
   - Direct marketing: is a type of advertising that motivates customers to place orders through different channels such as e-mail, cell phone, text message, interactive user websites, (Cheung, Kwok, Law, & Tsui, 2003; Liao & Chen, 2004; Prinzie & Poel, 2005).

3- Customer retention: The main concern of Custoemer Relationship Management is customer retention. The question is “How long the customer keep on purchasing product or service from the company?” In customer retention stage, customer satisfaction is a very important case. Customer satisfaction means the measurement of to what degree the customer’s

expectations from the company's products or services are met. Some methods of customer retention are following:

- complaints management
- one-to-one marketing: personalized marketing campaigns conducted by analysing customers and detecting and estimating changes in customer attitudes (Chen, Chiu, & Chang, 2005; Jiang & Tuzhilin, 2006; Kim & Moon, 2006). For example,
  - – customer profiling
  - – replenishment systems
  - – recommender systems
- loyalty programs: organizing campaigns or making supporting activities for establishing long term relations with customers. For example,
  - – satisfaction
  - – credit scoring
  - – service quality
  - – credit scoring

4- Customer development: Customer development includes the steps for keeping profitability and loyalty of acquised customers for a long time and increasing the customer spending.These steps are:

- up/cross selling:It is defined as the promotion activities made for increasing the associated or closely related services that an existing customer purchases (Prinzie & Poel, 2006).
- customer lifetime value analysis: It refers to the estimation of the amount of net earning a company expects getting from a customer. (Etzion, Fisher, & Wasserkrug, 2005;).
- market basket analysis: It involves increasing the customer purchase density and customer transaction value by showing up the patterns in the trading behaviour of customers (Chen, Tang, Shen, & Hu, 2005).

2.2 Advantages of CRM

With the help of current and complete databases, customer expectations can be identified and met in the best way.

Organizations' CRM strategies remove the problems about customer relationships. The tools which support these strategies save employers from doing inefficient and unnecessary work. Classification of data into categories, ease of control data, mobile devices.

2.3 CRM Architecture

CRM Architecture consists of three elements; Operational, Analytical and Colloborative CRM.

2.3.1 Operational CRM

The automation and improvement of business processes with front-office customer contact points. With the help of CRM software applications, selling, marketing, and service functions may be automated. Operational CRM has advantages to a company in these areas:

- **customer service automation:** Service automation helps companies/firms managing their service operations.
- **marketing automation:** Marketing automation applies technology to marketing processes. Campaign management modules provide marketers to use information about customers for developing, implementing and evaluating customer-specific offers.

- **sales force automation**: refers to make all the sales related functions automated. Main idea of it, is increasing the productivity of sales department and as a result, increasing company's sales transaction.

2.3.2 Analytical CRM

Analytical CRM is providing tools necessary for analyzing customer behaviors. Analytical CRM includes collecting, warehousing, analyzing of data created at operational side. The aim of the Analiytical CRM is predicting the future by extracting knowledge from past data. Analytical CRM provides better planning and management by making analysis. Without analytical CRM it is hard to maintain operational CRM and integrated projects.

Anaytical CRM solutions encompass a range of technologies, including data warehousing, data mining, statistical analysis and predictive modelling and multidimensional reporting. Without customer analytics, companies will be unable to effectively leverage their operational CRM efforts and make the most of thier investment in CRM.

Furthermore, as market conditions get tougher companies want to focus even more heavily on this area in order to rationalize their data structures and to use insights gained from analyzing cusyomer data to optimize the return from their current investments in CRM processes and technology.

2.3.3 Collaborative CRM

Colloborative CRM, performs corporation among customers, suppliers and business partners, provides faster response for customers and helps to increase efficiency at supply chain.

Customer contact points (phone, internet, faks, mail, etc.) management takes place in this context.

These solutions which provide to occur co-ordination and interaction between customers and organizations, transform information coming from different communication channels into valuable knowledge.

Colloborative CRM solutions contains all functions which allow interaction with customer. These interactions can take place through multiple channels and media, ranging from the website, e-mail and inbound and outbound telephone calls through online chat, kiosks etc.

CHAPTER THREE
DATA MINING

3.1 What is Data Mining

Data mining is extraction of previously unknown, significiant and applicable information from large databases and use of these information in making business decisions. With classical methods, it is very difficult or impossible to extract this knowledge from large databases.

Data mining process involves transforming large amount of collected data into knowledge by making various analysis.

Figure 3.1 Data to knowledge

Data mining is an interdisciplinar area which involves disciplines like database systems, statistics, visualization, machine learning and pattern recognition.

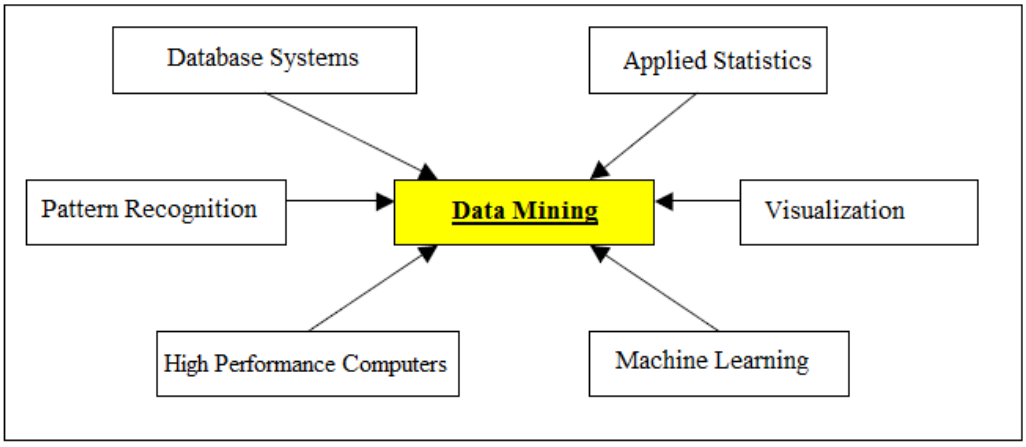


Figure 3.2 Disciplines related to data mining

Data mining is primarily used today by companies with a strong consumer focus retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among “internal” factors such as price, product positioning, or staff skills, and “external” factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits.

Data Mining is a stage in KDD (Knowledge Discovery in Databases) Process. Stages of KDD process:

1)Data Warehousing
2)Data Selection
3)Data Preprocessing
4)Data Transformation
5)*Data Mining*
6)Interpretation/Evaluation

3.2 Data Warehouse

Data warehouse contains data collected from different resources, with different views, from different time frames. Unnecessary parts of fhis data is removed. Some necessary transformations are made. After all these processes, all data is integrated in a single schema.

3.2.1 Properties of Data Warehouse

- Time variant
- Non-volatile
- Subject-oriented
- Integrated

3.3 Motivation of Data Mining

- The term “Data mining” emerged with changes in the business environment.
- Database sizes raised to terabytes.
- Organizations should take decisions quickly and with high knowledge
- Databases are expanding very rapidly.

3.4 Data Mining Applications

3.4.1 Marketing

- Retail Marketing
- Target Marketing (example: finding clusters of “model” customers who have the similar behaviors, interests, the level of income, spending habits, etc., example 2: Identifying customer purchasing patterns over time.)
- Customer Relationship Management (CRM) (example: Determining which of the customers are more loyal, and which are about to lost for a competitor?)
- Market Basket Analysis (example: finding which items are most frequently purchased together.)
- Cross Market Analysis (example: finding associations/co-relations between product sales, & predict based on such association)
- Market Segmentation
  - Dividing a larger market into submarkets based upon different needs or product preferences.
- Customer Profiling
  - What types of customers buy what products (clustering or classification)
- Customer Requirement Analysis
  - identifying customer requirements
  - identifying the best products for different customers
  - Predict what factors will attract new customers

- Campaign Analysis
  - identifying likely responders to promotions
  - identifying potential customers

3.4.2 Health

- Analysis of test results
- Medical diagnostic
- Disease detection according to the symptoms
- Determination of the treatment process
- The creation of gene maps
- Detection of gene sequences
- Associating genes with the disease
- Detection of genetic disorders
- The doctor - patient relationship management
- Hospital management decision support

3.4.3 Medicine

- Product development
- Research effects of drugs on diseases
- Definition of the side effects of drugs

3.4.4 Banking

- Detection of fraud
- Credit card fraud
- Anti-money laundering
- Evaluation of loan requests
- Determination of customer groups based on credit card spending
- Risk analysis, risk rating

- Find hidden correlations between different financial indicators

3.4.5 Insurance

- Determination of customer patterns of risk
- Estimating the customers who will require to new policy
- Insurance fraud detection

3.4.6 Stock Market

- Stock price prediction
- General market analysis
- Optimization of trading strategies

3.4.7 Telecommunication

Analysis of telecommunication data

- Density of the lines
- Figure out the related business
- Finding tricky activities
- Advance the quality of service
- Determine the telecommunication patterns
- Using resources in a better way.

Multidimensional analysis

- Originally multidimensional: location of callee, calling-time, calling duration, type of call

3.5 Data Mining Process

3.5.1 Goal Identification

The most important step in the data mining process. The first requirement to be successful in data mining applications, defining what business purpose the application has. This purpose must be clearly expressed and focused on the business' problem.

- Application Domain Understanding
  - Determine data mining objective
  - Determine success criteria
  - Inventory resources
  - Constraints
  - Cost and benefits
- Data Understanding
  - Decribe data

3.5.2 Data Preparation

The most inportant step in data mining is data preparation. Because the problems occur while the model is constructed cause turning back to this step. Data preparation contains data integration, data selection, data preprocessing and data transformation steps.

3.5.2.1 Data Integration

This step involves obtain and collect data from various sources. For example outside or inside of the company. Data from various sources are combined into a compatible store. This data may be collected from tables such as Excel Tables, Database Tables (Access, SQL Server, Oracle ...) or files (unstructured or structured

files) such as XML Files or Web Pages, Data Cubes, Data Marts (specialized version of a DW).

Some problems may occur during data integration process. Different spellings may be used for same person for example Agarwal, Agrawal, Aggarwal etc. Moreover there may be more than one way to denote an object like Data Mining , DM, D. Mining. Using different names is also a problem in data integration. (e.g Mumbai, Bombay). Required fields may be left blank. Some data may be inconsistent such as Invalid product codes collected at point of sale. Manually entry also may lead to mistakes.

3.5.2.2 Data Selection

Database may store terabytes of data and it may take too much time to run on the whole dataset. Need for complex data analysis / mining also caused data selection.

Selecting a target dataset is also a part of data selection. Data selection obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results.

Some data reduction techniques are as following:

1) Data Aggregation (sum, average)
2) Dimensionality Reduction (remove unimportant attributes)
3) Data Compression (encoding mechanisms)
4) Sampling (fit data into models)
5) Clustering (cluster data)
6) Concept hierarchy generation (street < city < state < country)

3.5.2.3 Data Preprocessing

In real world, data is not clean. Data may be incomplete, which means lacking attribute values or containing only aggregate data. They also may be containing errors or outliers or inconsistent, noisy, containing inconsistencies in codes or names. Data preprocessing is done by filling in missing values, removing noisy data, identifying and removing outliers, resolving inconsistencies and removing dublicate data.

3.5.2.4 Data Transformation

Data transformation is the process of transformation of some attributes into the way appropriate to the model. For example collectiong data like yes/no an instead of 0/1 may be more useful in terms of results.

Within the data transformation process, data may be converted into common format (Transform data into new format).Attribute construction may be needed. New attributes are constructed from the given ones. Discretization can be done as a part of data reduction especially for categorizing numerical data. Some aggregation and normalization process may be applied.

3.5.3 Data Mining

Data Mining is a step of Knowledge Discovery Process. According to the data mining objective,data mining algorithm and and the tool which implements the data mining algorithm is selected. Classification, Clustering, associaton rule mining are some of techniques used for data mining. The algorithm choice directly affects the success of implementation. So selection of data mining algorithm which will be implemented is very important process.

Data miners should also decide on which data mining tool sould be used, which one is the most appropriate to the constructed model. Some data mining tools are MS SQL Analysis Service, IBM Intelligent Miner.

3.5.4 Presentation and Interpretation of Knowledge

Last step of the Knowledge Discovery process is interpretation of data mining results, extracting valuable information from large amounts of data. Interesting and useful patterns are identified.

This phase of KDD process involves analyzing the data mining results, examining how well they performed on test data, deciding if the results are important or not, who needs to use them .

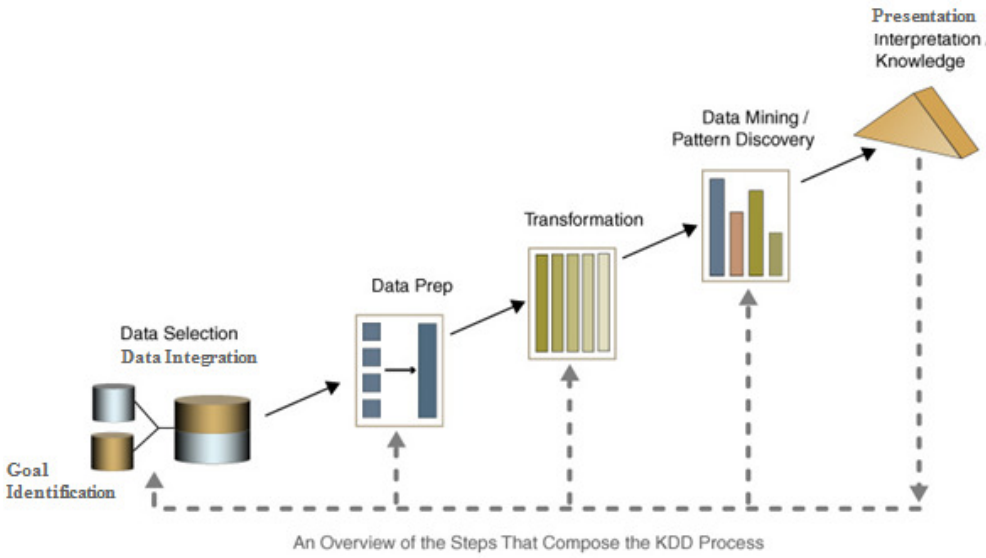


Figure 3.3 Steps of data mining process (Data Mining Process, n.d.)

3.6 Data Mining in CRM

Today, CRM is a concept that used to be quite common. A good CRM system requires identifying the needs of customers and understanding what customers like and what they do not like in a best way. CRM involves not only determining the

customers' needs and expectations, but also developing strategies according to these expectations and needs.

The core part of CRM activities is to understand customer requirements and retain profitable customers. To reach it in a highly competitive market, satisfying customer's needs is the key to business success (Gao, 2011). Unprecedented growth of competition has raised the importance of retaining current customers. Retaining existing customers is much less expensive and difficult than recruiting new customers in a mature market. So customer retention is a significant stage in Customer Relation Management, which is also the most important growth point of profit (Chung, Gray, 1999). Marketing literature states that it is more costly to engage a new customer than to retain an existing loyal customer. Churn prediction models are developed by academics and practitioners to effectively manage and control customer churn in order to retain existing customers (Padmanabhan, Tuzhilin, 2003). So, Customer satisfaction is important.

Data mining (DM) methodology has a great contribution for researchers to extract the hidden knowledge and information which have been inherited in the data used by researchers (Kuykendall, 1999). Data mining has a great contribution to the extraction of knowledge and information which have been hidden in a large volume of data (Han, Kamber ,2001). The concept of customer satisfaction and loyalty (CS&L) has attracted much attention in recent years. A key motivation for the fast growing emphasis on CS&L can be attributed to the fact that higher customer satisfaction and loyalty can lead to stronger competitive position resulting in larger market share and profitability (Kuykendall, 1999).

Data mining techniques, have important role to play in CRM applications. With data mining applications, databases, records in large companies can be converted into meaningful information.

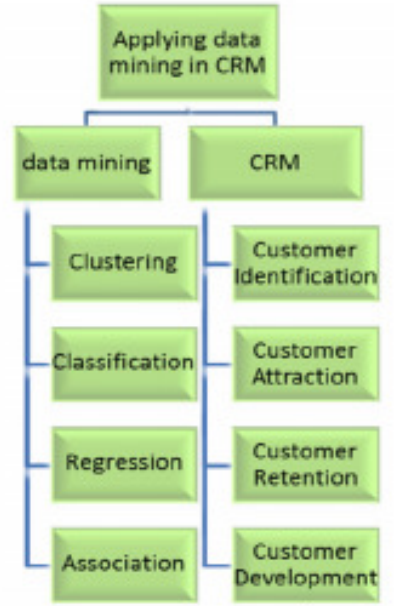


Figure 3.4 Classification of data minig techniques in CRM (Rodpysh, Aghai& Majdi,2012)

With data collected within the company in the past, different models such as the customer profile identification, campaign management, and customer loyalty determining models can be developed. With the help of these models, the information for decision support can be reached quickly.

For example determining which customers purchased which product combinations with market basket analysis is an important work in this subject. The results obtained at the end of this work, provide an important decision support for determining the target customer group at promotions, placing the products on shelves.

In Marketing and retail sectors, data can be systematically collected through sales terminals and coding systems. Customer and shopping he/she made is associated with each other with credit cards and shopping cards. Data collected in marketing and retail sector is extremely important at getting an advantage in a competitive environment.

Through data mining methods, information hidden in these data are discovered and discovered qualified information have a critical importance in terms of corporate competitiveness and success.

Firms prefer to provide services and product to the customers appropriate to profitability criteria rather than meet customers' all expectations equally. To do so, by considering the customers' past behavior, lifestyle and demographic characteristics, prediction models for future behavior are formed.

Customer segmentation is segmenting customers into homogenous groups according to their common features. One of the goals of customer segmentation is increasing the customer loyalty and profitability. The stage after customer segmentation is customer profiling. Customer behavior model or customer profile is the most important tool for target marketing applications.

Customer profiling is the process of determining the customers' characteristics like age, income level, and lifestyle. Customer profile is generated with customer behavior models. The simplest form of customer profiles is made up from customer's name, surname, address, city, postal codes etc. Customer profile information are like customer's cultural background, economic structure, frequency of shopping, frequency of complaints, level of satisfaction, references, age, education level, lifestyle, media tools that the customer use, the way of first contact with the company.

For such applications, various tools of information technology are needed. Especially for the collection of data, the creation of market-oriented strategic information and planning the marketing campaigns, data mining is used as an important tool.

Data mining helps defining the behavior surrounding a particular lifecycle event as shown in Figure 3.5:

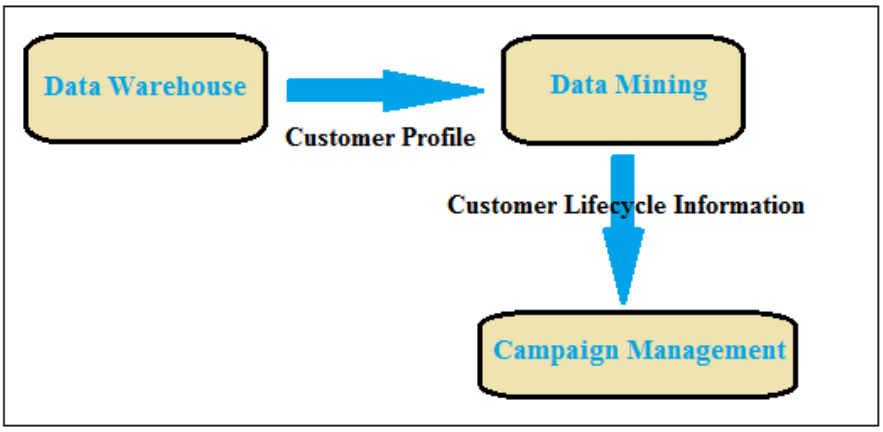


Figure 3.5 Data mining in CRM

3.7 Data Mining Techniques in CRM

3.7.1 Predictive Models

Classification and regression are two data analysis methods which revealing important data classes or constructing models of predicting trends in future data.

3.7.1.1 Classification

The purpose of the classification is constructing a classifier which determines a new object belongs to which class. Predefined classes are used for model generation to classify data collected from data warehouse or database.

A classification is used to match the characteristics of the product with the customer. Thus , the ideal product for a customer or ideal customer profile for a product may be found. For example young women purchase small car, old and rich men purchase big and luxurious car. If A car company finds a rule like that, gives advertisements of small cars to the magazine that young women read.

Common classification techniques are

- bayesian classification,
- k-nearest neighbor,

- decision trees,
- artificial neural networks,
- genetic algorithms,
- support vector machines,
- fuzzy set approach,
- rough set approach.

3.7.1.2 Regression

Regression refers to a type of statistical prediction model that is used to map every data object to a. actual value supply prediction value (Carrier & Povel, 2003). Regression is used to predict continuous values. Regression model may be constructed for example to estimate their expenditures of potential customers, whose income and job is known, while they are purchasing computer products.

3.7.2 Descriptive Models

3.7.2.1 Clustering

Clustering is keeping similar records together. Clustering is usually looking from the top to the database. Clustering is a multi-variable statistical technique, primary aim of which is grouping observation units according to their similarities. The observation units in a cluster obtained from the cluster analysis, similar to each other in terms of a predetermined property. So the observation units at the obtained cluster are homogenous. The goal of clustering model, obtaining clusters that properties of each cluster members are very similar to each other, but properties of clusters are very different from each other, and dividing records in database into these clusters. Here, the most important features of each cluster are common.

Some algorithms for clustering:

- K-means,
- DBSCAN,

- SOM(Self Organizing Map).

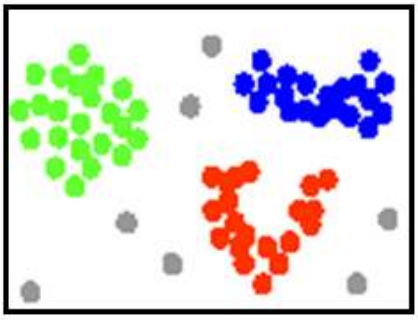


Figure 3.6 Clustering

3.7.2.2 Association

Association rules determines the products the customers purchased together during any shopping or the other products they purchased as another shopping process and behavior models related to these purchasing patterns. This analysis is also called Market Basket Analysis. For example if a person who is travelling, buys an introductory book about the country he travels to, he will buy a dictionary also with a 20% percent probability.

Apriori algorithm, fp growth tree are common algorithms for association technique.

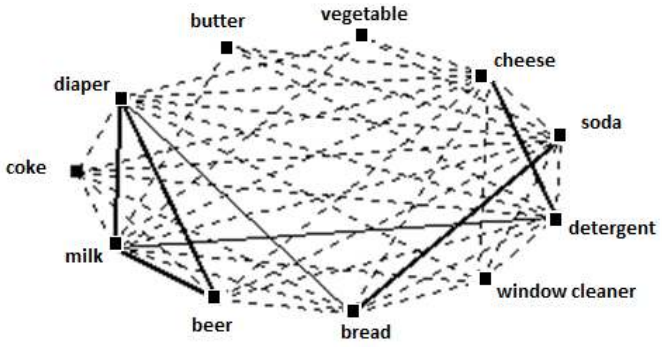


Figure 3.7 Association among supermarket products

CHAPTER FOUR
BAYESIAN CLASSIFICATION

4.1 Bayesian Classification

Bayes' Theorem is named after Thomas Bayes, an 18th century British mathematician and minister. He worked on probability and decision theory.

Bayesian classification is a statistical classified method. It can predict class membership probabilities, such as the probability that a given data belongs to a particular class. Naïve Bayesian classifiers are based on Bayes' theorem. They are useful in data mining and decision support.

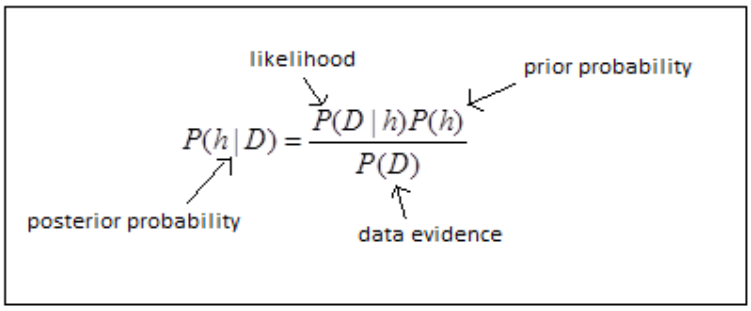
Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. Bayesian belief networks are graphical models, which unlike naïve Bayesian classifiers allow the representation of dependencies among subsets of attributes. Bayesian belief networks can be used for classification as well.

Bayesian technique is used in supervised learning. The number of classes is known before. The posterior distribution is computed using the training samples.

Learning is formulated as a form of probabilistic inference, using the observations to update a prior distribution over hypotheses in Bayes classification. The probability of each hypothesis, given the data is calculated.

After that estimations are made using all hypotheses weighted by their probabilities. The training of the Bayesian Classifier consists of the estimation of the conditional probability distribution of each attribute, given the class.
Given a hypothesis h and data D which concerned with the hypothesis:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$



P(h): independent probability of h (prior probability)

P(D): independent probability of D (data evidence)

P(D|h): conditional probability of D given h (likelihood)

P(h|D): conditional probability of h given D (posterior probability)

The goal of Bayes Theorem is to speficy the most probable hypothesis from the given data D.

Prior probability of h, P(h): is the probability of being h is a correct hypothesis. Prior probability of D, P(D): is the probability of training data D will be observed. Conditional Probability of observation D, P(D|h): is the probability of observing data D given some world in which hypothesis h holds. (Barber, 2010)

Probability Theory

A Random variable (RV): a variable that takes on values from a set of mutually exclusive and exhaustive values.

- A=a: a proposition, variable A has a particular value a

This can correspond to a percept or feature (e.g. Wind=Weak)

- P(A=a): single probability of RV A=a, which is the degree of belief in a proposition in the absence of any other relevant information (e.g. P(Wind=Weak))
- P(A): probability distribution, i.e. set of P(A=ai) for all i

(e.g. P(Wind) = { P(Wind=Weak), P(Wind=Strong) })

Conditional (posterior) probabilities:

- Formalize the process of accumulating evidence and updating probabilities based on new evidence
- Specify the belief in one proposition (event, conclusion, diagnosis, etc.) conditioned on another proposition (evidence, feature, symptom, etc.)
- P(A | B) is the conditional probability of A given evidence B is known to be true:

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

Bayes' Rule is the basis for efficiently computing unknown conditional probabilities, as derived from the product rule:

$$P(A \wedge B) =$$

$$P(A \mid B) \times P(B) = P(B \mid A) \times P(A)$$

$$\boxed{P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}}$$

4.2 Properties of Bayes Classifier

- Incrementality: The predicted probability may be increased or decreased incrementally by every training data. The prior and the likelihood may be updated dynamically by each training sample.
- Combines prior knowledge and observed data: To specify the last probability of a hypothesis, observed data and prior knowledge may be combined. Prior probability of a hypothesis multiplied with probability of the hypothesis given the training data.

- Probabilistic hypothesis: Hypotheses with probabilities can be accommodated. Distribution of probabilities over each classes is the result of Bayes Classifier beside the classification.
- Variables have to be assumed to be independent.

4.3 Naive Bayesian Categorization

Naive Bayes aims to simplify the estimation problem by assuming that the different input features are conditionally independent. That is, they are assumed to be independent when conditioned on the class. Mathematically, for inputs x ∈ R^d, it is expressed as:

$$p(\mathbf{x}|C) = \prod_{i=1}^{d} p(x_i|C)$$

For this reason, it is only needed to get P(Xi | C) for every possible couple of a category and a feature-value.

Bayes Classification Example:

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Figure 4.1 Sample data set

The problem is prediction of play for the day <sunny, cool, high, strong>.

$$P(PlayTennis = yes) = 9/14 = 0.64$$
$$P(PlayTennis = no) = 5/14 = 0.36$$
$$P(Wind = strong \mid PlayTennis = yes) = 3/9 = 0.33$$
$$P(Wind = strong \mid PlayTennis = no) = 3/5 = 0.60$$
$$etc.$$
$$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes) = 0.0053$$
$$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no) = \mathbf{0.0206}$$
$$\Rightarrow answer : PlayTennis(x) = no$$

Incomplete databases seriously compromise the computational efficiency of Bayesian classifiers. One approach is to throw away all the incomplete entries. Another approach is to try to complete the database by allowing the user to specify the pattern of the data.

4.4 Zero Value Problem

The **Laplacian correction** is the solution of zero value problem. Remember the formula used in Bayesian classification:

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

If there is not an attribute value xk in Ci, P(xk|Ci) becomes equal to zero. So multiplication of all attributes of X becomes equal to zero (P(X|Ci)=0).

To solve this zero probability value problem, it is assumed that the training data is large enough that adding one to each count of attribute value affects the estimated probabilities very little and it can be ignored but prevents the zero probability condition. This method is known as Laplacian correction.

For example, assume that for the class play = yes in a training set which includes 100 samples, There are 10 samples with wind = low, 90 samples with wind = medium, and 0 samples with wind = high.

The probabilities of these events, without the Laplacian correction, are 0.10 (from 10/100), 0.90 (from 90/100) and 0 respectively. After the Laplacian correction, '1' is added to each wind-value pair. In the end, following probabilities are obtained respectively.

$$\frac{11}{103}=0{,}106 \quad \frac{91}{103}=0{,}883 \quad \frac{1}{103}=0{,}009$$

4.5 Continuous Data

If attribute X has continuous values instead of categorical values, for calculating P(Xi | Y), Gaussian distribution is used to calculate the probability of X (P(X|Y)).

For each combination of a continuous value Xi and a class value for Y, yk, has a mean, μik , and standard deviation(variance) σik based on values Xi in class yk.
For estimation P(Xi | Y=yk) of this example, Gaussian distribution of Xi is defined by μik(mean) and σik (variance) depends on Y:

$$P(X_i \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \exp\left(\frac{-(X_i - \mu_{ik})^2}{2\sigma_{ik}^{\ 2}}\right)$$

Formulas of mean and variance:

Mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Variance:

$$\sigma^2 = \frac{1}{(N-1)}\sum_{i=1}^{N} (x_i - \mu)^2$$

4.6 Advantages of Bayesian Classification

- One of the most convenient learning methods along with decision trees, k nearest neighbor, neural networks.
- Model is incrementally updated with training examples
- Bayesian Classification can classify new instances through combining predictions of multiple hypotheses.

CHAPTER FIVE
WCF SERVICES, ANDROID AND PHONEGAP

5. 1 WCF Services

Windows Communication Foundation (WCF) is a Software Development Kit for developing and deploying services on Windows (Löwy, 2007). WCF is a structure for building service-oriented applications. Services could be built without WCF, but services could be built significantly easier with WCF. WCF maintains interoperability between services. WCF provides lots of beneficial facilities for developing services, like security, reliability, hosting, service instance management, asynchronous calls, disconnected queued calls and transaction management. WCF has also an extensibility model. Actually, WCF is developed by using this extensibility model.

Properties of WCF :

- Security
- AJAX and REST Support
- Interoperability
- Attribute-based programming
- Service Orientation
- Service Metadata
- Data Contracts
- Extensibility and Location Transparency

5.1.1 Differences between Web Service and WCF Service

- WCF has more flexibility and transportability than old ASMX. Because WCF design is a summary of whole different distributed programming infrastructures presented by Microsoft.

- Web Services can only be reached over HTTP. But WCF is flexible. WCF services may be hosted in such as IIS ,WAS (Windows Activation Service).
- One of the main differences is the web services use XmlSerializer while WCF services use DataContractSerializer. Because DataContractSerializer has a better performance than XMLSerializer.

  Some differences between DataContractSerializer and XMLSerializer.
  - DataContractSerializer has better performance than Xmlserializer.
  - The DataContractSerializer can convert the hash table into XML.
  - DataCotratSerializer specifies which fields or properties of type are serialized into XML, but XmlSerializer not.

Creating a Sample Service in WCF:

ServiceContract indicates that an interface defines a WCF service contract. OperationContract specifies which methods are the operations of the service contract.

```
 [ServiceContract]
public interface IService1 {
[OperationContract]
int MultiplyWithTen(int value);
}
public class Service : IService1 {
public int MultiplyWithTen (int value) {
int result = value*10;
return result;
}
}
```

5.1.2 Hosting The WCF Service

WCF services need to be hosted in a host process. A single host process can host multiple services, as well as multiple host processes can host the same service type.

WCF services are compiled into a class library. All libraries need a host to run in. WCF Service can be hosted with IIS(Internet Information Server) or WAS (WindowsActivationService).

5.1.2.1 IIS (Internet Information Server) Hosting

The fundamental benefit of IIS hosting is that the host process is started automatically when the first client makes request, and IIS web server manages the life cycle of the host process.

Hosting a WCF service in IIS is very similar to hosting a traditional ASMX web service.

- Compile the service into a class library.
- Create a virtual directory under IIS
- Copy the service (.svc) file into the directory. (the syntax of a sample .svc file is in the example below)

Example: A .svc file syntax

```
<%@ ServiceHost Language="C#" Debug="true"
Service="WcfService1.Service1"
CodeBehind="Service1.svc.cs" %>
```

When hosting with IIS, the base address used for the service must be the same as the address of the .svc file.

5.1.2.2 Endpoints

A WCF service has at least one endpoint. Endpoints provides communication to WCF service and determines the communication rules. There may be one endpoint at client site, and n endpoints at server site. This means n different communication ways. The endpoint is the union of the (A)ddress, (B)inding and (C)ontract. ABC of an endpoint determines the endpoint's characteristics.

(A)ddress: Basically the address of the service. Actually, address determines how to be reached at the service. Addresses are unique and have format like that:

"Protocol://<MachineName>[:port]/Path"
Protocol is meant to the way the service is reached. Such as HTTP, TCP, etc.

(B)inding: Binding determines how to communicate with service. It includes the information of how the message will be sent.

(C)ontract: Contract determines what the service does. Contract indicates what the message contains. It specifies what the object is (Service Contract, Data Contract, Fault Contract, Message Contract).

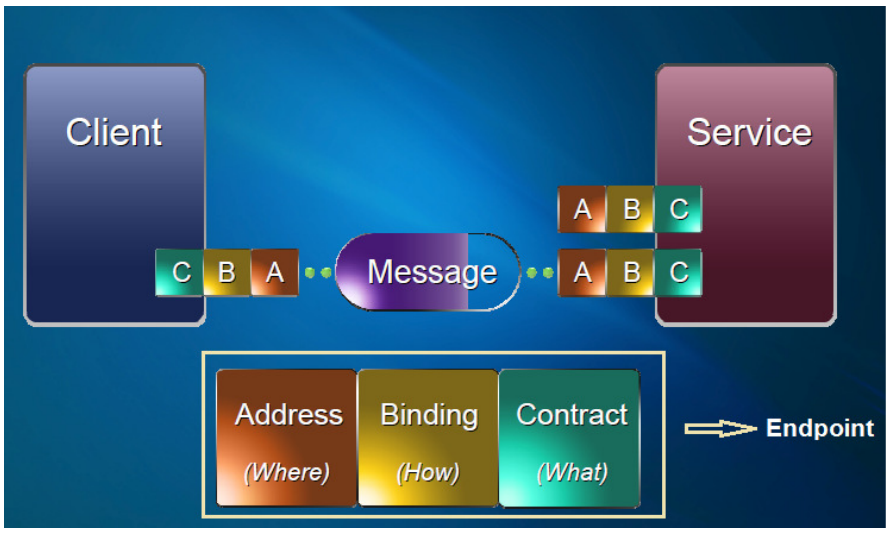


Figure 5.1 Structure of endpoints of client and service (Cohen-Yashar ,n.d.)

5.2 Android

5.2.1 Android Operating System

Android is an open source operating system that was developed using the Linux kernel. The use of the operating system especially on mobile phones and tablet PCs is expanding day by day. With this popular technology, which has a SDK for software developers, development can be done on Windows, Linux ve Mac OS X.

Advantages of Android

- A simple and powerful SDK
- No pay for licensing, development or distribution
- Development over many platform
  - Windows, Linux, Mac OS,
- Good documentation
- Growing developer community

5.2.2 Android Development Tools

Android Development Tools (ADT) is provided by Google for developing Android applications with Eclipse.

Android applications can be created, compiled, debugged and deployed from the Eclipse IDE and from command line with ADT. An Android device emulator is also provided by ADT for testing Android applications without a mobile device.

5.2.3 Android Application Architecture

5.2.3.1 AndroidManifest.xml

AndroidManifest.xml file describes the Android application. All components of the application for example Activities and Services are must be defined in this file.

The necessary permissions for the application are also must be specified in AndroidManifest.xml.

Example AndroidManifest.xml file

```
<?xml version="1.0" encoding="utf-8"?>
<manifest xmlns:android="http://schemas.android.com/apk/res/android"
    package="com.myproject.myphonegapproject"
    android:versionCode="1"
    android:versionName="1.0" >
    <uses-sdk android:minSdkVersion="16" />
    <supports-screens
        android:largeScreens="true"
        android:normalScreens="true"
        android:smallScreens="true"
        android:xlargeScreens="true"
        android:resizeable="true"
        android:anyDensity="true"
        />
    <uses-permission android:name="android.permission.INTERNET" />
    <application android:icon="@drawable/ic_launcher"
        android:label="@string/app_name" >
        <activity android:configChanges="orientation|keyboardHidden|keyboard|screenSize|locale"
            android:name=".MyPhoneGapActivity"
            android:label="@string/app_name" >
            <intent-filter>
                <action android:name="android.intent.action.MAIN" />
                <category android:name="android.intent.category.LAUNCHER" />
            </intent-filter>
        </activity>
    </application>
</manifest>
```

5.2.3.2 R.java and Resources

R.java is a generated class. R.java includes references to particular resources of the project. The resources must be defined in the “res” directory. The resources may be XML files, pictures or icons.

Eclipse creates R.java automatically. It is not needed to make changes manually.

5.2.3.3 Activities and Lifecycle

The Android system manages the life cycle of the application. The operating system defines a life cycle for activities through the pre-defined methods. Some of the important methods are:

onPause() - called when the Activity ends. The method is used to release resource or save data.

onResume() - called when the Activity is restarted The method can be used to initialize fields.

onSaveInstanceState() - called when the activity is stopped, The method is used to save data in order to the activity can restore its states when re-started.

5.2.3.4 Context

The Context class provides the connections to the Android system. It also provides access to Android Service.

5.2.4. Installation

Tools

- Eclipse (http://www.eclipse.org/downloads/)
  - Android Plugin (ADT)
- Android SDK (http://developer.android.com/sdk/index.html)

After downloading Eclipse, Android Plugin is installed by going to Help -> Install New Software -> Add and entering name: ADT Plugin and Location: https://dl-ssl.google.com/android/eclipse/. Then, the location of Android SDK downloaded before is selected from Preferences window.

5.2.5 Android Virtual Device – Emulator

The Android Development Tools (ADT) provides an Android device emulator to run an Android application. The emulator operates as a real Android device mostly and allows testing Android application without a mobile device.

The device may be selected from the emulator. Several devices may be started in parallel. These devices are named as Android Virtual Device (AVD).

Figure 5.2 Android Virtual Device Emulator

5.2.6 Creating a New Android Project

File - New - Project from Eclipse is selected. Then "Android Project" under "Android" is selected.

In the opening window, project name, application name and package name are written. Create “Activity” is selected and name of activity is typed. Build target is selected and finish button is clicked.

View of Package Explorer when a new android project is created as shown in figure 5.3:

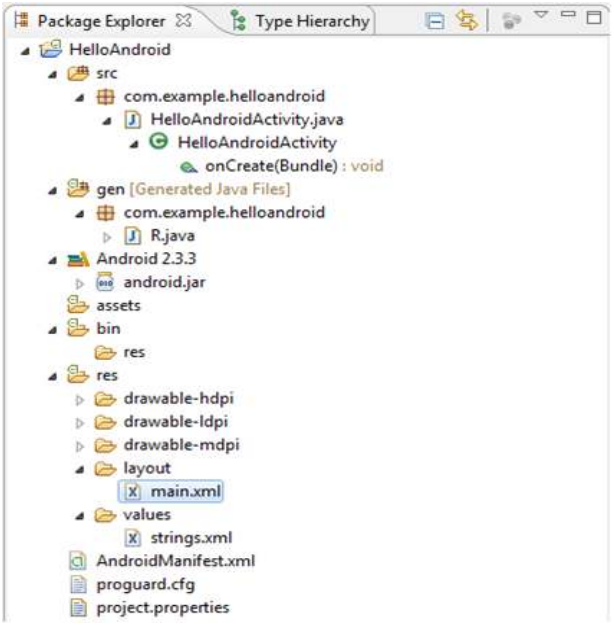


Figure 5.3 Package Explorer

HelloAndroidActivity.java

Activities parts of the application that performs actions. An application can contain a lot of activity. However, the user may interact with only one of them at the same time.

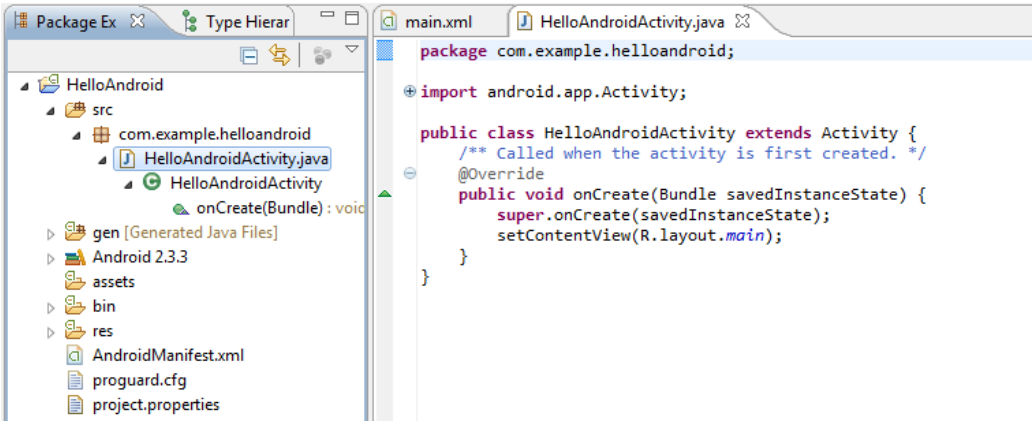


Figure 5.4 HelloAndroidActivity.java

AndroidManifest.xml

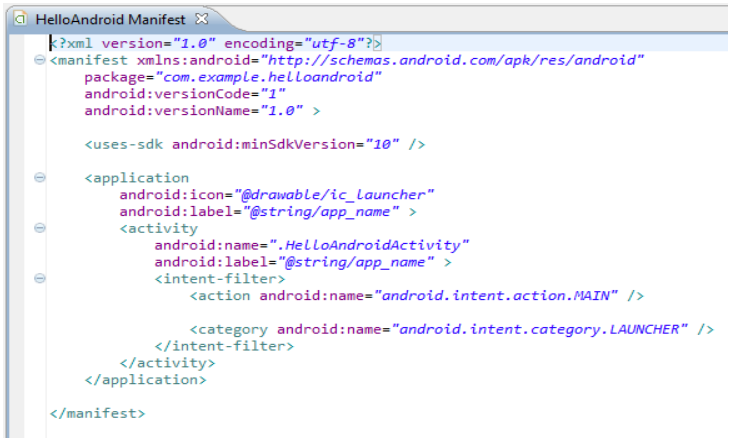


Figure 5.5 Android Manifest.xml

5.3 PhoneGap

PhoneGap is an open-source development tool for mobile cross-platform App publication that uses device-agnostic wrappers (Device agnosticism is the capacity of a computing component to work with various systems without requiring any special adaptations. The term can apply to either hardware or software. In an IT context, agnosticism refers to anything that is designed to be compatible across most common systems. A device-agnostic mobile application (app), for example, is compatible with most operating systems and may also work on different types of devices.) like HTML, Javascript, and CSS, that can be rapidly deployed on Android, Blackberry, and iPhone platforms. PhoneGap is a "develop once, publish anywhere" package/project.

Required Tools

- Latest PhoneGap project
- Sun Java SE JDK 6
- Eclipse IDE
- Android SDK
- ADT Plug-in

5.3.1 Creating a PhoneGap Project

- An android Project is created.
    - File -> New -> Android Project
- In the root directory of the project, two new directories are created.

/libs and /assets/www

- phonegap.js is copied from PhoneGap download earlier to /assets/www
- an index.html file in /assets/www is created.
- phonegap.jar is copied from your PhoneGap download earlier to /libs
- xml folder is copied from your PhoneGap download to /res
- the build path of the phonegap.jar is set.
- /libs folder is right clicked.
- Build Paths/ -> Configure Build Paths is selected.
- In the Libraries tab, phonegap-2.2.0.jar is added to the project.

After these steps all it is needed to do is to edit the index.html file. Folloqing figure (Figure 5.6) showa a sample index.html file.

```
<!DOCTYPE HTML>
<html>
  <head>
    <meta name="viewport" content="width=320; user-scalable=no" />
    <meta http-equiv="Content-type" content="text/html; charset=utf-8">
    <title>PhoneGap Android App</title>
              <script type="text/javascript" charset="utf-8" src="phonegap.js"></script>
              <script type="text/javascript" charset="utf-8">
                        var showMessageBox = function() {
                             navigator.notification.alert("Hello World of PhoneGap");
                        }
                        function init(){
                             document.addEventListener("deviceready", showMessageBox, true
                        }
  </script>
  </head>
  <body onload="init();"  >
  </body>
</html>
```

Figure 5.6 Index.html

CHAPTER SIX
IMPLEMENTATION

Because of high competition in the business field, it is important to consider the customer relationship management of the businesses. The massive volume of customer data is analyzed and classified based on the customer behaviours and prediction.

The classifier will predict the customers belongs to which class that should have highest posterior probability. The valuable customer information accumulated by a Portuguese banking institution, which is used to identify customers and provide decision support.

A data model is generated based upon the history of the customers in the bank. Then the sample data is classified by using the Naïve Bayesian classification algorithm and placed them into the appropriate class based upon the posterior probability and the percentage of subscribing a term deposit for the customers can be predicted.

In this application, the dataset is obtained from the UCI machine learning repository (http://archive.ics.uci.edu/ml/).

The data is obtained from a Portuguese banking institution. The dataset is about **direct marketing campaigns**. Aim of the classification is to predict if the client will subscribe a term deposit or not(Moro, Laureano, Cortez, 2011).

Information about attributes used in the application:

Inputs:

age (numeric)

job: type of job (categorical: 'admin.', 'unemployed', 'management', 'housemaid', 'entrepreneur', 'student', 'blue-collar', 'self-employed', 'retired', 'technician', 'services')
marital: marital status (categorical: 'married','divorced','single')
education (categorical:,'primary', 'secondary','tertiary')
default: has credit in default? (binary: 'yes','no')
balance: average yearly balance, in euros (numeric)
housing: has housing loan? (binary: 'yes','no')
loan: has personal loan? (binary: 'yes','no')

- attributes about the last contact of the current campaign:
contact: contact communication type (categorical:'telephone','cellular')
duration: last contact duration, in seconds (numeric)

- other attributes:
campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

Output (target):
y : has the client subscribed a term deposit? (binary: 'yes','no')

Note: Unknown values are ommited from dataset.

Banks have numerous individual retail customers. They uses CRM because of its analytical abilites. CRM helps the banks to increase the cross sell performance and manage the churn rates (customer defection rates). Data Mining models may be used to define the customers which are eager to confirm cross sell offers, which are about to be lost and what can be done to win them again.

Dataset in stored in Microsoft SQL Server 2008 inside the database BankDB. The Data Mining operation is applied to data. This operation is converted to a WCF service.

Architectural design of the application is as in the Figure 6.1. Wcf service connects to database and read data from database or write into the database. From the application WCF service is called for computing the bayesian classification. Phonegap which is a “write once, run everywhere” platform connects to wcf services. It enables to run the application on all operating systems like IOS, Android, Windows mobile etc...

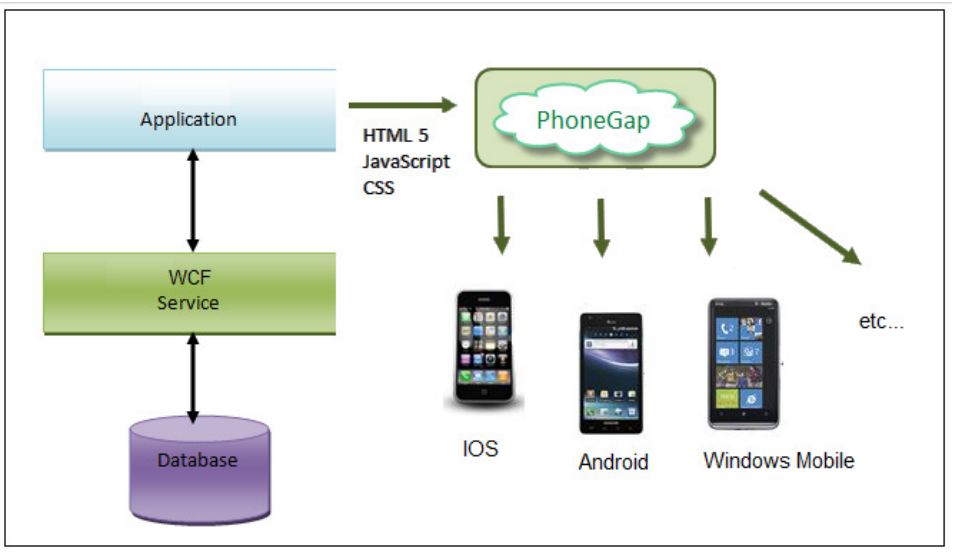


Figure 6.1 Architectural design

The application is developed on Android platform (Android 4.1.2) using PhoneGap (version 2.1.0) technology to connect to the service and return the result. JSON data type, which means JavaScript Object Notation, is used for sending and receiving data between WCF service and Phonegap application. JSON is language independent and JSON parsers are available for many different programming languages.

```
function CallWCFService(WCFServiceURL) {
    $.ajax({
        type: "GET",
        url: WCFServiceURL,
        contentType: "application/json; charset=utf-8",
        dataType: 'json',
        processdata: true,
        success: function (msg) {
            WCFServiceSucceeded(msg);
        },
        error: alert
    });
```

Figure 6.2 Json data type in an example javascript function

6.1 Implementation of Bayesian Classification

Bayesian calculation is done by using the dataset and the probabilies of each variables are written into database when the method BayesianCalculation() is called from WCF service.

At first, model is constructed with trainig test and tested with test set. Figure 6.3 shows the model construction diagram. In training set, output classes of samples are known. There are categorical and continuous attributes in dataset. Figure 6.3 includes some of them.

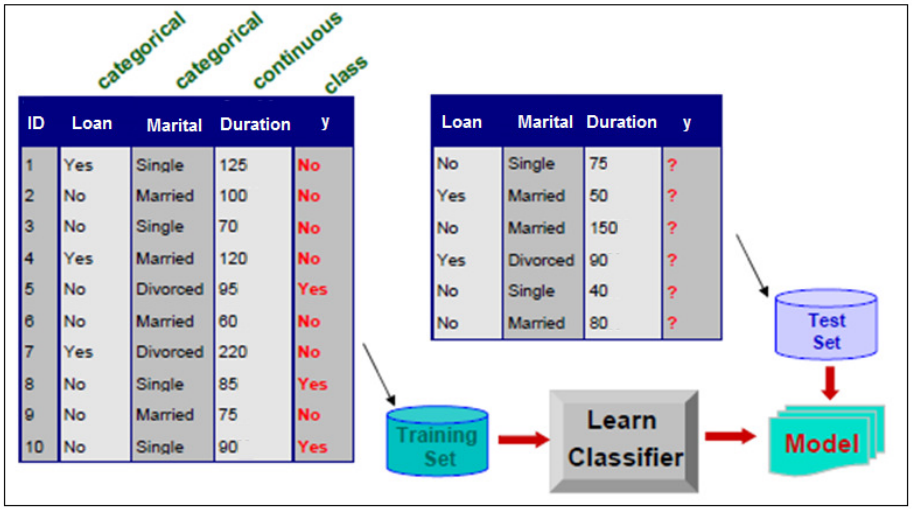


Figure 6.3 Model construction

Some of the inputs have continuous values in the dataset. Continuous values are handled by Gaussian Distribution which is explained in chapter four.

Probabilities of all categorical attributes when output class y is 'No' and output class y is 'Yes' are calculated and shown in Table 6.1 for the dataset used in the application.

Table 6.1 Probabilities of categorical attributes

		No	Yes
job	management	0,233450798	0,258865248
	blue-collar	0,189799553	0,116134752
	technician	0,182865371	0,163342199
	services	0,08798454	0,067375887
	admin.	0,113068849	0,122340426
	unemployed	0,029707097	0,040336879
	entrepreneur	0,034064643	0,022163121
	housemaid	0,02932818	0,020833333
	retired	0,045697397	0,103501773
	self-employed	0,037626464	0,035682624
	student	0,016407108	0,049423759
marital	single	0,279830245	0,365469858
	married	0,607328256	0,520611702
	divorced	0,112841499	0,11391844
education	secondary	0,524800121	0,476507092
	primary	0,142397029	0,111702128
	tertiary	0,332802849	0,41179078
default	no	0,981963548	0,992464539
	yes	0,018036452	0,007535461
housing	no	0,475995605	0,664893617
	yes	0,524004395	0,335106383
loan	no	0,8213785	0,91001773
	yes	0,1786215	0,08998227
contact	cellular	0,911181842	0,922429078
	telephone	0,088818158	0,077570922

Mean and variance of each continuous attributes when output class y is 'No' and output class y is 'Yes' are calculated and shown in Table 6.2 These values are used to calculate Gaussian distribution which has the following formula:

$$P(X_i \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \exp\left(\frac{-(X_i - \mu_{ik})^2}{2\sigma_{ik}^{\ 2}}\right)$$

Table 6.2 Mean and variance of continuous attributes

	No		Yes	
	Mean	Variance	Mean	Variance
age	40,77863666	10,37905095	41,73891844	13,64783701
balance	1402,718086	3086,941414	1870,782358	3593,489828
duration	218,3497783	205,4261557	507,03125	370,4182292
campaign	2,861960517	3,092719771	2,105718085	1,831433975

As an example, following sample customer data is an evidence for Bayesian Classifier.

Table 6.3 Example customer data

age	job	marital	education	default	balance	housing	loan	contact	duration	campaign	y
34	self-employed	married	tertiary	no	1045	yes	yes	telephone	65	2	?

Probabilites of each classes for given sample customer data and Gaussian distributions for continuous attributes and each general class probabilites (P(yes) and P(no)) are all multiplied for finding the probable class of given customer. All values and result of multiplication are shown in Table 6.4. According to the table, Probabiliy of No is higher than Probability of Yes.

Table 6.4 Mean and variance of continuous attributes

	No	Yes
self-employed	0,037626464	0,035682624
married	0,607328256	0,520611702
tertiary	0,332802849	0,41179078
no	0,981963548	0,992464539
yes	0,524004395	0,335106383
yes	0,1786215	0,08998227
telephone	0,088818158	0,077570922
G(age= 34)	0,031054774	0,024890009
G(balance= 1045)	0,000128371	0,000108125
G(duration= 65)	0,001469764	0,000528436
G(campaign= 2)	0,124080091	0,217467943
no/y - yes/y	0,853994758	0,146005242
RESULT	**3,85451E-14**	**8,01878E-16**

The percentage of results may be calculated as following:

$$\frac{X * 100}{X + Y}$$

When it is applied to the sample data :

(3,85451E-14 * 100) / (3,85451E-14 + 8,01878E-16)

According to result of the calculation it can be expressed that the output class of the sample customer is no with 97% probability.

When the application is run, following screen (Figure 6.4) appears and user enter the parameters.

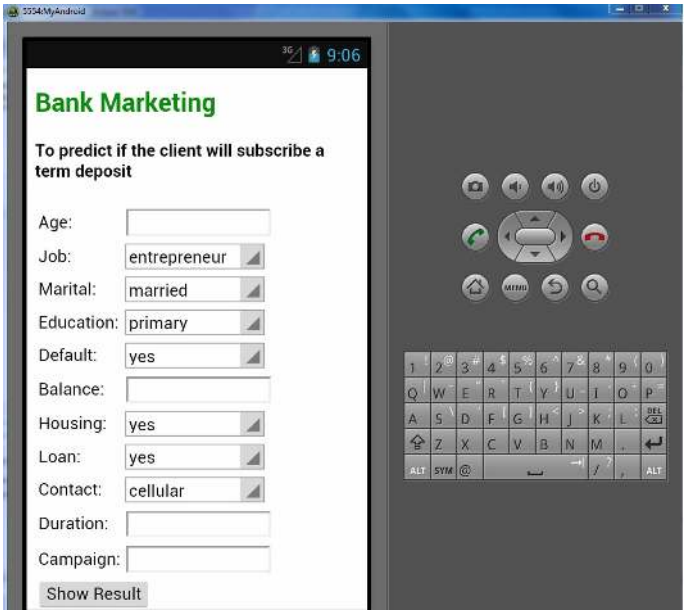


Figure 6.4 Interface of the Application

After clicking the Show Result button, program evaluates the values entered by user and shows the classification result. “Classify” method is called from WCF when the buton is clicked. And the service returns the result.

Figure 6.5 shows the result:

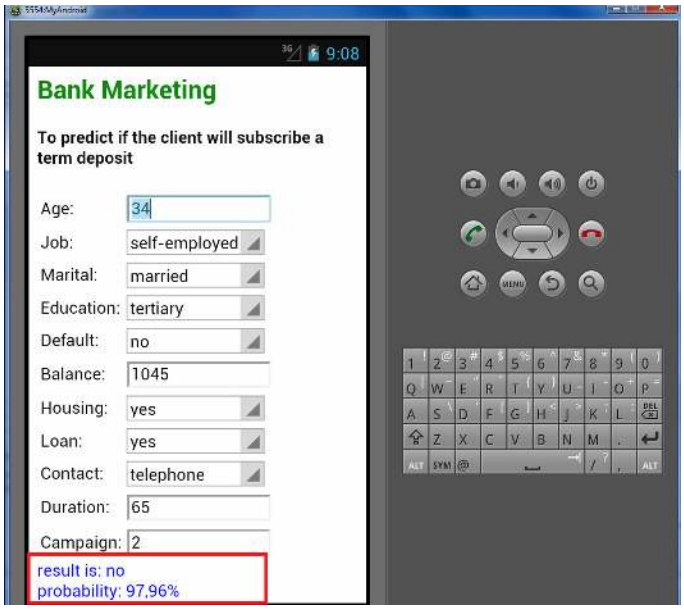


Figure 6.5 Result of the classification and probability of the result

6.2 Accuracy Rate Of The Model

Accuracy rate of the model used in this application, is computed by doing Cross Validation.

Cross validation allows using whole dataset in computing. Dataset is divided into two parts randomly. First part is used for model construction. The model is tested with second part of dataset and the accuracy rate is computed. After that, a model is constructed with second part and tested with first part. Accuracy rate is computed. Finally, the model is constructed using whole dataset. Average of the computed accuracy rates is the accuracy rate of the constructed model. Figure 6.6 shows the Cross Validation:

Figure 6.6 Cross Validation

In this study, the first half of dataset is used to train and second half is used to test. After that, the second half of dataset is used to train and first half is used to test the algorithm. Cross validation showed that the accuracy rate of the model is 84.5%.

First test results are showed in Table 6.5. The first part of dataset is used for testing. There are 906 samples which belong to actual class value 'yes'. After the test, 555 of them are classified as 'yes' which is the right result and 351 of them are classified as 'no'. There are also 14545 samples which belong to actual class value 'no. After the test, 13540 of them are classified as 'no' which is the right result and 1005 of them are classified as 'yes'.

As a result, accuracy rate of first training is ;

Truly classified samples: 555+13540=14095

All samples in dataset: 15451

Accuracy rate : (14095 * 100) / 15451 ~= 91%

Table 6.5 First cross validation results

		Classified As	
		Yes	No
Actual Class Value	Yes	555	351
	No	1005	13540

Second test results are showed in Table 6.6. The second part of dataset is used for testing. There are 3606 samples which belong to actual class value ‘yes’. After the test, 552 of them are classified as ‘yes’ which is the right result and 3054 of them are classified as ‘no’. There are also 11846 samples which belong to actual class value ‘no. After the test, 11514 of them are classified as ‘no’ which is the right result and 332 of them are classified as ‘yes’.

As a result, accuracy rate of second training is ;
Truly classified samples: 552+11514=12066
All samples in dataset: 15452
Accuracy rate : (12066* 100) / 15452 ~= 78%

Table 6.6 Second cross validation results

		Classified As	
		Yes	No
Actual Class Value	Yes	552	3054
	No	332	11514

Average of accuracy rates of first and second tests give the accuracy of the implementation which is 84.5 %.

CHAPTER SEVEN
CONCLUSION AND DISCUSSION

Today information is vital for companies and Customer Relationship Management and Data Mining techniques provide to determine customers with high profitability. The information obtained from these methods have a strategic importance for companies. In this way, they provide an important competitive advantage to competitors.

The massive volume of customer data is analyzed and classified based on the customer behaviours and prediction. This study has built a classifier using the naive Bayesian classification. Naive Bayesian classifier reported high acceptable accuracy like 84.5% for all the tested data by doing cross validation. The accuracy is determined as the percentage of the correctly classified instances from the test set. In other words, classification centers around exploring through data objects (training set) to find a set of rules which determine the class of each object according to its attributes. Since the accuracy of the model is acceptable, the model can be used to classify data tuples whose class labels are not known. The classification rules can be used to support decision making for achieving a good CRM for businesses.

In this thesis, data obtained from a banking institute is analyzed. Bayesian classification is applied to the data. Bayesian classification is implemented as a WCF service. From an Android device, this service is called and result of classification of a new customer data is showed.

Dataset is stored in Microsoft SQL Server 2008. Bayesian classification method is implemented on a WCF (Windows Communicaion Foundation) service project using Microsoft Visual Studio 2010. The result of the implementation is presented on Android platform using Phonegap technology.

With the information obtained from the application, the companies can predict the behavior of a customer. For this application, the bank can predict if a customer will subscribe a term deposit or not. So that they can manage direct marketing campaigns using this prediction. They can acces the results anytime and anywhere through a mobile device.

With application of CRM in mobile platform, changes and updates can be done seamlessly from anywhere and anytime with no delays.

REFERENCES

Abrahams, A. , Becker, S. , Sabido, A.B. , Souza, D.D. , Makriyiannis, R. & Krasnodebski, G. (2009). Inducing a Marketing strategy for a New Pet Insurance Company using Decision trees. *Expert systems with applications*. Vol.36,pp.1914-1921.

Barber, D. (2010). Prior, Likelihood and Posterior In *Bayesian Reasoning and Machine Learning* Retrieved November 29, 2012 from http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/090310.pdf

Berry, T. (n.d). *Customer Relationship Management (CRM)* Retreived November 29, 2012, from http://articles.bplans.com/growing-a-business/customer-relationship-management/92

Carrier, C. G. & Povel, O. (2003). Characterising data mining software. *Intelligent Data Analysis*, *7*, 181–192.

Chen, M. C., Chiu, A. L., & Chang, H. H. (2005). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications, 28,* 773–781.

Chen, Y. L., Tang, K., Shen, R. J., & Hu, Y. H. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems, 40,* 339–354.

Cheung, K. W., Kwok, J. T., Law, M. H., & Tsui, K. C. (2003). Mining customer product ratings for personalized marketing. *Decision Support Systems, 35,*231–243.

Chung, H.M. & Gray, P. (1999). "Data mining"[J]. *Journal of MIS*,16(1):11-13

Cohen-Yashar, M. (n.d.). Endpoint Retrieved November 29, 2012 from https://www.owasp.org/images/5/51/OWASP_IL_WCF_Security.pdf

Data Mining Process (n.d.). Retrieved November 25,2012 from http://www.quretec.com/u/vilo/edu/2004-05/Andmekaevandus/index.cgi?f=Intro

Danutazakrzewska, A.J. (2008). Improving Naïve Bayes models of Insurance risk by unsupervised classification. *Proceedings of the International Multiconference on Computer Science and Information Technology,* pp.137-144.

Etzion, O., Fisher, A., & Wasserkrug, S. (2005). E-CLV: A modeling approach for customer lifetime evaluation in e-commerce domains, with an application and case study for online auction. *Information Systems Frontiers, 7,* 421–434.

Gao, Hua. (2011). IEEE, "Customer Relationship Management Based on Data Mining Technique".

Garg, K. , Kumar, D. & Garg, M.C. (2008). Data mining techniques for identifying the customer behavior of investment in life insurance sector in India. *International journal information technology and knowledge management*, Vol.1,pp.51-56.

Glover, S. , Rivers, P.A. , Asoh, D.A. ,& Piper, C.N. & Murph, K. (2010). Data mining for health executive decision support: an imperative with a daunting future*., Health Services Management Research* ,Vol. 23, Number 1 > pp. 42-46

Han, J. & Kamber, M. (2001). *Data Mining: Concepts and Techniques[M].* CA: Morgan Kaufmann Publishers. San Francisco

Huang Jian-ming, Fang Jiao-li & Wang Xin-ping. (2009). Research on Bayesian NetWorks Model of University Courses [J]. *Journal of Guizhou University (Science)*, 4(2): 81-84.

Jiang, T., & Tuzhilin, A. (2006). Segmenting customers from population to individuals: Does 1-to-1 keep your customers forever. *IEEE Transactions on Knowledge and Data Engineering, 18,* 1297–1311.

Kim, Y. H., & Moon, B. R. (2006). Multicampaign assignment problem. *IEEE Transactions on Knowledge and Data Engineering, 18,* 405–414.

Kracklauer, A. H., Mills, D. Q., & Seifert, D. (2004). Customer management as the origin of collaborative customer relationship management. *Collaborative Customer Relationship Management - taking CRM to the next level*, 3–6.

Kuykendall, L. (1999). ”The data-mining toolbox”[J]. Credit Card Management,12(6):30-40.

Liao, S. H. & Chen, Y. J. (2004). Mining customer knowledge for electronic catalog marketing. *Expert Systems with Applications, 27,* 521–532.

Li Xiang & Zhou Li. (2009). Value Forecast Based on Bayesian Network for Credit Card Customers [J]. *Computer and Digital Engineering*, 37(3), 91-93.

Leng Cui-ping, Wang Shuang-cheng & Wang Hui. (2011) Study on Dependency Analysis Method for Learning Dynamic Bayesian Network Structure [J]. *Computer Engineering and Applications*, 47(3), 51-53.

Löwy, J. (2007). WCF Essentials In Programming WCF Services(1st. ed.)(1).USA: O’Reilly

Ma Da, Hu Hai-guang & Guan Jian-he.(2010). The Naïve Bayesian Approach in Classfying the Learner of Distance Education System. *The 2nd International Conference on Information Engineering and Computer Science*, 12(2): 978-981.

Moro, S., Laureano, R., & Cortez, P. (2011). Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. *European Simulation and Modelling Conference - ESM'2011*

Ngai, E.W.T. , L.Xiuand & D.C.KChau. (2009). Application of data mining techniques in customer relationship management, *Expert systems with applications*, Vol.36,pp.2592-2602.

Rodpysh, K.V. , Aghai, A., Majdi, M. (2012). Applyıng Data Mınıng In Customer Relatıonshıp Management. *International Journal of Information Technology, Control and Automation (IJITCA)* Vol.2, No.3

Shaw, R. (2002). Customer Relationship Management (CRM) : Overview. Retrieved November 29, 2012 from http://facweb.cs.depaul.edu/nsutcliffe/450Readings/Customer%20Relationship%20Management%20(CRM)%20Overview.htm

Sreedhar, D., Manthan, J., Ajay, P., Virendra, S., and Udupa N. (2007). Customer Relationship Management and Customer Managed Relationship -Need of the hour Retreived November 29, 2012, from http://www.pharmainfo.net/reviews/customer-relationship-management-and-customer-managed-relationship-need-hour

Staudt, M.J. , Kietz, U. & Reimer, U. (1998). A data mining support environment and its application on Insurance data, *American Journal for Aritifical Intelligence*, pp.15-19.

Swift, R. S. (2001). Accelarating customer relationships: Using CRM and relationship technologies. Upper saddle river. N.J.: Prentice Hall PTR.

Padmanabhan, B. & Tuzhilin, A. (2003) On the Use of Optimization for Data Mining: Theoretical Interactions and eCRM Opportunities[J], *Management Science* . Vol. 49, No. 10, October 2003, pp. 13271343.

Parvatiyar, A., & Sheth, J. N. (2001). Customer relationship management: Emergingpractice, process, and discipline. *Journal of Economic & Social Research, 3,* 1–34.

Petersen, G. (2003). Customer Relationship Management In ROI: Building the CRM Business Case. Xlibris.

Prinzie, A., & Poel, D. V. D. (2005). Constrained optimization of data-mining problems to improve model performance: A direct-marketing application. *Expert Systems with Applications, 29,* 630–640.

Viveros, M.S. (1996). BM Research Division T. J. Watson. Applying Data Mining Techniques to a Health Insurance Information System. *Proceedings of the 22nd VLDB Conference* Mumbai (Bombay), India.

Wu, Chien-Hsing, Kao, Shu-Chen, Su, Yann-Yean & Wu, Chuan-Chun. (2005). *Targeting customers via discovery knowledge for the insurance industry*, Vol.29 .pp.291-299. Elsevier Ltd.

Yao, Z. , H.Holmbom, A. , Eklund, T. & Back, B. (2010). Combining Unsupervised and Supervised Data Mining Techniques for Conducting Customer Portfolio Analysis, *ICDM*. LNAI 6171,pp.292-307.

Young Moon Chae, Seung Hee Ho, Kyoung Won Cho, Dong Ha Lee & Sun Ha Ji,(2001),Data mining approach to policy analysis in a health insurance domain, *International Journal of Medical Informatics*, vol. 62, pp.103–111.

Young Moon Chaea, Dong Ha Leeb, Hwa Young Kima, Tae Soo Kima & Tae Min Songc(2004),Mining time dependency patterns in a Health Insurance Domain, *International journal of Medical informatics*,Vol.99.pp.1545-1555.