RISK MODELING USING DECISION TREE ALGORITHM FOR VOLUNTARY MOTOR INSURANCE

SUPARAT TUBNAKOG

MASTER OF SCIENCE
IN ADVANCED INFORMATION TECHNOLOGY

MAE FAH LUANG UNIVERSITY
2007

RISK MODELING USING DECISION TREE ALGORITHM FOR VOLUNTARY MOTOR INSURANCE

SUPARAT TUBNAKOG

AN INDEPENDENT STUDY SUBMITTED TO
MAE FAH LUANG UNIVERSITY IN PARITIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN ADVANCED INFORMATION TECHNOLOGY

MAE FAH LUANG UNIVERSITY
2007

RISK MODELING USING DECISION TREE ALGORITHM FOR VOLUNTARY MOTOR INSURANCE

SUPARAT TUBNAKOG

THIS INDEPENDENT STUDY HAS BEEN APPROVED
TO BE A PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF THE MASTER OF SCIENCE
IN ADVANCED INFORMATION TECNOLOGY
2007

EXAMINING COMMITTEE

....CHAIRPERSON
(Asst. Prof. Gp. Capt. Dr. Sanlayut Sawangwan)

Yooyativong, T
....MEMBER
(Gp. Capt. Dr. Thongchai Yooyativong)

....MEMBER
(Lecturer Paola Di Maio)

....MEMBER
(Lecturer Vittayasak Rujivarakul)

ACKNOWLEDGEMENT

First and foremost I would like to thank my advisor, Professor Dr.Sanlayut Sawangwan, Dr.Thongchai Yooyativong, Flt.Lt. Dr.Tossapon Boongoen, Ms.Paola Di Maio for valuable guidance and motivation in this project. They always support and encourage me during all situations. My work and study could not progress without them.

I also would like to thank my independent study committee members, Mr.Piyasak Jeatrakul and Mr.Vittayasak Rujivorakul for their valuable advice, comment, and guidance in this thesis.

I would like to thank Ms.Ai-yarest Saelee the manager from The Viriyah Insurance Co., Ltd for providing data, suggestions throughout the project.

I am grateful of my friend Ms.Sakaowrath Chatjaroenchap for recommend assistance.

This study is dedicated to my parents and my friends for their constant support and encouragement throughout the years of my study.

Suparat Tubnakog

Independent Study Title	Risk Modeling Using Decision Tree Algorithm for Voluntary Motor Insurance	
Author	Miss Suparat Tubnakog	
Degree	Master of Science (Advanced Information Technology)	
Supervisory Committee	Gp.Capt.Dr.Thongchai Yooyativong	Chairperson
	Flt. Lt. Dr.Tossapon Boongoen	Member
	Lecturer Miss Paola Di Maio	Member

ABSTRACT

In this paper, we propose a framework based on data mining algorithms for building a Web-page underwriter system. The insurance has developed a risk model system for underwriting that makes lower claims. This paper describes an approach for building risk predictive models using SQL Server 2005 Analysis Services; the decision tree is used for classifying customers into one of pre-defined levels of risk. ID3, Entropy and Bayesian with K2 Method are score-based for split decision tree are used to analyze the data. In order to conduct data mining process we use the industry-standard CRISP-DM methodology (CRISP-DM Group,1996). Implement representing result from model we use OLE DB for Data mining support for data mining APIs on Microsoft platforms.

As a result, the system response risk level of customer could be recommend and support information about level risk of customer for underwriter which would allow the insurance company to avoid customer who high risk.

Keywords : Insurance risk modeling / Decision tree / Entropy

CONTENTS

CONTENTS (Cont.)

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES (Cont.)

Figure **Page**

LIST OF TECHNICAL VOCABULARY

Accuracy

A measure of a predictive model that reflects the proportionate number of times that the model is correct when applied to data.

Application Programming Interface (API)

The formally defined programming language interface between a program (system control program, licensed program) and its user.

Database

A collection of data stored in a standardized format, designed to be shared by multiple users. A collection of tables for a particular business situation.

Data Mining

Searching databases for unknown patterns and information. Tools include statistical analysis, pattern-matching techniques, and data segmentation analysis, classification analysis, association rules, and cluster analysis.

Decision Trees

Logical branching on variables; enumerating cases with outputs.

Embedded Data Mining

An implementation of data mining where the data mining algorithms are embedded into existing data stores and information delivery processes rather than requiring data extraction and new data stores.

Entropy

The measure of information, or more specifically, the degree of randomness in data. Lower scores indicate a greater ability to discriminate outcome categories.

Lift

A number representing the increase in responses from a targeted marketing application using a predictive model over the response rate achieved when no model is used.

Test Set

Portion of the data available that’s used to test the data mining model

Training Set

Portion of the data available that’s used to build the data mining model

CHAPTER I

INTRODUCTION

1.1 Background Information

In general, data mining techniques can be divided into two broad categories discovery data mining and predictive data mining. Discovery data mining is applied to a range of techniques which find patterns inside your data without any prior knowledge of what pattern exist. Predictive data mining is applied to a range of techniques that find relationships between a specific variable called the target variable and the other variables in your data. Decision tree is one technique of predictive mining techniques.

The insurance company has over 80 databases containing details about their products, customers and claims. As the amount of data increases, the company will have huge databases containing mostly historical data. Currently, there is a data warehouse to manage information for effective data access and summarization. However, such summarized data alone is insufficient for assessing customer's risk. This traditional risk management can be terribly enormous without utilizing the true probabilistic nature and relationships of historical data. Thus, it is necessary to implement a mining model to find the pattern of customer's risk so it helps underwriter to evaluate customers.

In this paper, we introduce some representation models, based on decision tree, which are applied to the specific case of classified insurance risk level. These models have been obtained from real data recorded at the insurance company, by using several algorithms for building decision trees. Study of the capabilities decision tree learning algorithms in order to generate useful models for this problem. We have therefore selected a representative subset of the currently available algorithms for learning decision tree and we have carried out a series of experiments to evaluate their behavior from different perspectives.

1.2 Objectives

This project aims to build a risk models based on the voluntary motor insurance historical data that will to maximize the accuracy of their risk assessments for avoid high risk. The objectives for develop risk model are:

1.2.1 Create risk model for predicting level risk of customer or prospects that have high likelihood of creating a loss for the company. Using risk scoring model get the good new customer.

1.2.2 Understand their profile high-risk customers in their high volume data.

1.2.3 Avoiding high-risk customers.

1.3 Scope

This project was to create a risk model using a data mining tool for finding pattern data. This project is a web-based mining system with the following features:

1.3.1 Risk models functionality directly to be applied on-line into the underwriter application. The risk model display graphical user interface is tailored to the insurance for enhanced ease of use. Risk model will help underwriter for determine level of risk for new customers.

1.3.2 Risk model can be building off-line by administrator, but should be rebuilding periodically.

1.3.3 The system includes the following risk voluntary motor insurance report, the information contained customer id and level risk with information support each case that is determine future for each customer.

1.3.4 Users can use all the functionality of the system through a web browser.

1.3.5 Users are separated into Administrators and Staff with corresponding rights and privileges

1.4 Expected Benefits

1.4.1 Underwriter will have facilitated for making decision about risk of customers based on historical data by using risk model that can support decision for underwriter determine their risk.

1.4.2 A risk models that will suggest the most likely risk customers from a list of potential customers.

CHAPTER II

FEASIBILITY STUDY

2.1 Introduction

In this study, companies have extended applied data mining tools to aid in the prediction the probability of future claims for voluntary motor insurance and supporting decision. Decision tree method based on data mining techniques can be used to classification policy risk level. In order to conduct data mining analyses, process of data mining consisting of a sequence of steps. There is a standard process CRISP-DM widely used (S.Daskalaki, I.Kopanas, M.Goudata, and N.Avouris,2003). We use CRISP-DM methodology. In major tasks of CRISP-DM consists of six phases are data preparation, data preprocessing, data mining, interpretation, application and evaluation.

2.2 Problem Statement

The insurance company collects data for each policy they underwrite and claim data, a high volume of data is created through transactions made daily. The problem of insurance need to know “**What policy risk level of new customer?”** for evaluates risk level; they needs to make a decision based statistical probability of insurance claims. For finding pattern risk data so large that it is difficult to analyze data.

2.2.1 The availability of large quantities of insurance data represents both an opportunity and a challenge for data mining. The opportunity exists to use data mining techniques to discover previously unrecognized risk groups and thereby assist actuaries in developing more competitive rating systems that better reflect the true risks of the policies that are underwritten.

2.2.2 Data collects a large amount of information on insured entities Policy information that consists of many variables, involving both categorical and numerical data.

2.2.3 Insurance needs to take decision from information when sizes of data increases, information can be gain easily from them are decreases. From historical data and use mining technique calculations can discover useful, structural pattern.

2.3 Related Research and Projects

There have been the several projects related to data mining. Chickering and Heckerman (D.Chickering and D.Heckerman, 2000) analyze direct mailing promotional strategies that can be used to help increase sales of a product. They describe two methods for partitioning the potential consumers into groups, and show how to perform a simple cost-benefit analysis to decide which, if any, of the groups should be targeted. They consider two decision tree learning algorithms. The first is an "off the shelf" algorithm used to model the probability that groups of consumers will buy the product. The second is a new algorithm that is developed from the first algorithm.

There has been the insurance industry utilizes data mining for marketing, just as retailing do. (T. Hoffman,1999) Farmers Insurance Group decided to give data mining a try. They believe that data mining can address several specific problems that arise and it can provide and effective tool for narrow niche marketing. They have developed a system for underwriting that generates higher revenues and lower claims. They develop system for underwriting that can predict losses for specific line of insurance.

Ling and Li (Ling, Chares X. and Li,Cenghui, 1998) analyze direct marketing as a process of identifying likely buyers of certain products and promoting the product accordingly. They believe that Data Mining can address several specific problems that arise and it can provide an effective tool for direct marketing. They discuss methods of coping with these problems in direct-marketing projects.

2.4 Requirement Specifications for the New System

This project will utilize risk model and deploy to underwriter web. For this implementation will aim to satisfy the following scope

2.4.1 Administrator can builds and process risk model that developed from insurance data set on server with decision tree algorithm.

2.4.2 Administrator can use server functionality that is extraction, transformation and load (ETL) package for data warehousing

2.4.3 Underwriter user can retrieve real-time classified level risk from risk model.

2.4.4 The system will be tested on the Windows 2000/XP platforms

2.4.5 The user interface of underwriter is web-based and can be used in IE (Internet explorer) version 5.0 or upper

2.4.6 Equipment and Software Required

From scope in section 2.4, this project will utilize web and database technologies with embedded data mining. Our implementation will use tools as show in table 2.1.

Table 2.1 Applications require the environment

Operating System	▪ Microsoft Windows® 2000 Server (SP2 or later). ▪ Microsoft Windows XP.
Applications	▪ SQL Server 2005 (Standard, Developer, or Enterprise Edition). ▪ SQL Server 2005 Integration Services ▪ Analysis Services in Microsoft® SQL Server™ 2005 ▪ Microsoft Visual Studio .NET (with Microsoft Visual C#™ components installed).
Accounts	▪ Valid Windows accounts and SQL Server login accounts for any users who will be working.
Services	▪ Microsoft Windows Internet Information Services (IIS)
Connection	▪ LAN connection

(Microsoft Corporation, 2005)

In figure 2.1 shows the component of SQL Server Data mining is part of Analysis Services server. Microsoft® SQL Server™ 2005 provides an integrated environment for creating and working with data mining models as follows:

1. Analysis Services in Microsoft® SQL Server™ 2005 includes nine algorithms :

1) Microsoft Decision Trees

2) Microsoft Clustering

3) Microsoft Naïve Bayes

4) Microsoft Sequence Clustering

5) Microsoft Time Series

6) Microsoft Association

7) Microsoft Neural Network

8) Microsoft Linear Regression

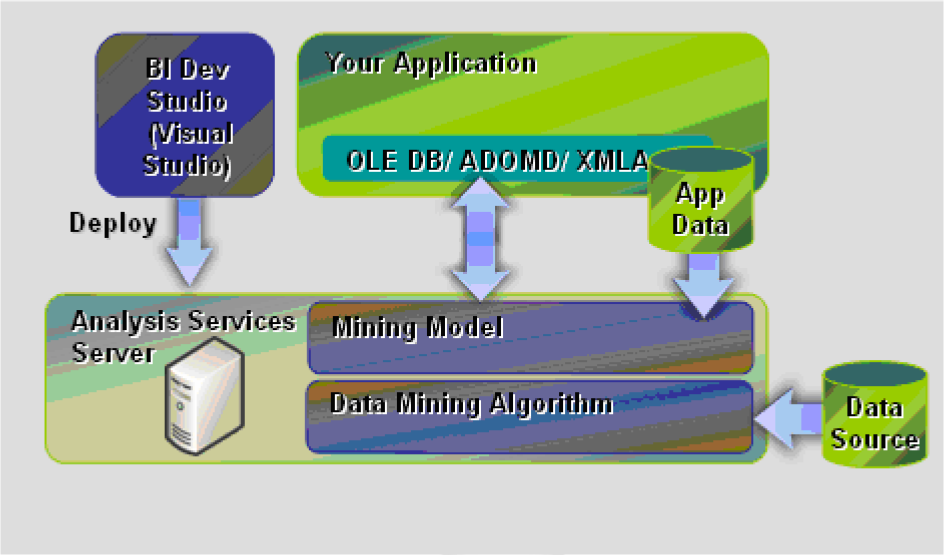
9) Microsoft Logistic Regression

2. SQL Server 2005 Integration Services create data pipeline, organization data, separate data, and fill in missing values based on the predictive analytic of the data. This can be integrating any data source with ETL and data warehousing capabilities.

1) Microsoft visual C# 2005 for develops underwriter application.

2) SQL queries will be used to process and retrieve data and information from the DB server as much as possible, this may limit Thai language support

3) OLE DB for Data mining API or DMX language will be implemented to be representing result from mining model as SQL-style query language for data mining. Embedded data mining is key for wide adoption of the technology Complete SQL language based API



(Microsoft Corporation, 2005)

Figure 2.1 Server for mining architecture

2.5 Implementation Techniques

In this Section, the basic concepts of decision tree are explained with example. Decision trees are one of such methods of automatic data research and a simple successful technique for supervised classification learning. Evaluate model by lift chart and classification matrix. In order to conduct data mining analyses data mining standards are concerned. The last step in the data mining process is to deploy to a production environment the models. Deploy data mining functionality to production environment with standard data mining API.

2.5.1 Decision Tree is used as classifiers. Tree is constructed in a top-down recursive divide-and-conquer manner. At start, all the training examples are at the root. Attributes are categorical. Examples are partitioned recursively based on selected attributes. Test attributes are selected on the basis of a heuristic or statistical measure (C.Apte and S.Weiss, 1997). There are different types of decision tree. Each uses a different rule for deciding splits. The impurity measures are used as following.

1. Entropy used by the C4.5 and C5.0 algorithms. This measure is based on the concept of entropy used in information theory. Entropy is function to measuring the impurity level of

collection of example where the target classification is binary consists of positive example and negative example. Entropy measures the impurity level of collection of examples that the lower entropy value, the better. If the entropy is 0 that is all case belong to the same class for positive cases. If entropy is 1 that is all case is negative case. The range of entropy is 0 that perfectly classified and 1 is totally random.

$$\textbf{Entropy} = -\sum_{i=1}^{n} P_i \log_2 P_i$$

ID3 is based off the Concept Learning System (CLS) algorithm. It uses a greedy concept with Entropy and Information Gain as heuristics criteria in finding the best attribute to split the tree. ID3 (Quinlan, 1986) stand for Iterative Dichotomiser 3 is an algorithm used to generate a decision tree. The ID3 system selects an attribute as a root, with branches for different values of the attribute (D.A.Koonce, C. H.Fang, and S. C. Tsai,1997). All objects in the training set are classified into these branches. If all objects in a branch belong to the same output class, the node is labeled and this branch is terminated. If there are multiple classes on a branch, another attribute is selected as a node, with all possible attribute values branched.

2. Bayesian with K2 is a score-based greedy search algorithm for learning Bayesian networks from data. It was published in Cooper, G. and Herskovits, E. (1992). It is a Bayesian method for the induction of probabilistic networks from data. K2 uses a Bayesian score, *P(Bs, D)*, to rank different structures and it uses a greedy search algorithm to maximize *P(Bs, D)*.

$$g(i,\pi_i) = \prod_{j=1}^{q_i} \frac{(r_i-1)!}{(N_{ij}+r_i-1)!} \prod_{k=1}^{r_i} N_{ijk}!$$

By computing such ratios for pairs of bayes network structures, we can rank order a set of structures by their posterior probabilities. Based on four assumptions, the paper introduces an efficient formula for computing P(Bs,D), let B represent an arbitrary bayes network structure containing just the variables in D

K2 Algorithm (Cooper and Herskovits, 1992) start with a given ordering attributes.

1. Greedily consider adds edges
2. Maximize the network's score
3. Using the number of parents for over fitting
4. Run the algorithm several times with
5. Random ordering of attributes

Stopping and Pruning Rules Decision tree can stop building the tree when**:**

1. The impurity of all nodes is zero. Problem is that this tends to lead to bushy, highly-branching trees, often with one example at each node.

2. No split achieves a significant gain in purity

3. Node size is too small. That is, there are less than a certain number of examples, or proportion of the training set, at each node.

For demonstration risk ID3 with insurance dataset, use ID3 algorithm to generate a tree, with outcome as the dependent variable (target). In table 2.2 data set or examples risks insurance has attribute as follows :

1. Instances of problem with correct answer
2. Each problem is described by set of attributes for example; Policy type, Liability, Car name, Car type, Car's use, Car age, Outcome
3. Each attribute has a definite range of value
    1) Policy = {Type1, Type2, Type3}
    2) Liability = {0-300, 301-500, 501-800, 801-1000}
    3) Car name = {TOYOTA, HONDA, BENZ, NISSAN}
    4) Car type = {Passenger, Bus, Truck, Motorcycle}
    5) Car's use = {Private, Commercial }
    6) Car age = {1 yr, 2 yr,3 yr, 4 yr, 5-7 yr,8-10 yr}
    7) Outcome ={Safety, Loss}

Table 2.2 Example of dataset risk insurance

Case	Policy type	Liability	Carname	Car type	Car's use	Car age	Outcome
1	Type1	0 - 300,000	TOYOTA	Passenger	Private	1 Year	Safety
2	Type1	0 - 300,000	TOYOTA	Passenger	Private	2 Years	Safety
3	Type2	0 - 300,000	HONDA	Passenger	Private	2 Years	Safety
4	Type3	500,001 - 800,000	TOYOTA	Bus	Commercial	5 - 7 Years	Loss
5	Type3	500,001 - 800,000	HONDA	Passenger	Private	4 Years	Safety
6	Type1	800,001 - 1,000,000	BENZ	Passenger	Private	5 - 7 Years	Safety
7	Type1	500,001 - 800,000	BENZ	Truck	Commercial	3 Years	Safety
8	Type1	300,001 - 500,000	TOYOTA	Passenger	Private	2 Years	Loss
9	Type2	0 - 300,000	HONDA	Passenger	Private	4 Years	Loss
10	Type2	300,001 - 500,000	NISSAN	Passenger	Private	8 - 10 Years	Safety
11	Type1	0 - 300,000	HONDA	Truck	Commercial	2 Years	Loss
12	Type1	300,001 - 500,000	NISSAN	Passenger	Commercial	2 Years	Safety
13	Type2	0 - 300,000	TOYOTA	Truck	Commercial	2 Years	Safety
14	Type3	800,001 - 1,000,000	HONDA	Motorcycle	Private	3 Years	Loss
15	Type3	300,001 - 500,000	TOYOTA	Passenger	Private	5 - 7 Years	Loss
16	Type3	300,001 - 500,000	TOYOTA	Passenger	Private	8 - 10 Years	Safety

Process or ID3 algorithm generate tree can be following steps:

Step1 : Calculate entropy before splitting

P(*safety*) = 10/16

P*(loss)* = 6/16

E*(before)* = -10/16 log2(10/16) -6/16log2(6/16)

= 0.95

Step2 : Test and calculate the entropy after splitting of each attribute

Testing "Policy type" features and calculate weight each attribute.

Type1 P*(safety)* = 5/7

P*(loss)* = 2/7

E*(Type1)* = 0.86

Type2 P*(safety)* = 3/4

P*(loss)* = 1/4

E*(Type2)* = 0.81

Type3 P*(safety)* = 2/5

$$\text{P}(loss) = 3/5$$

$$\text{E}(Type3) = 0.97$$

$$\text{W1} = 10/16,\ \text{W2} = 6/16$$

$$\text{E}(after) = \text{W1} \times \text{E}(Type1) + \text{W2} \times \text{E}(Type2) + \text{W2} \times \text{E}(Type3)$$

$$= 1.194$$

Table 2.3 Select attribute with highest IG

Attribute	E (after)	IG
policy type	1.19483	-
carname	0.9763	-
car's use	0.95514	-
liability	1.92057	-
car type	0.66137	0.309583055
car's age	1.6683	0.977053453

The highest IG

Step3 : pick the attribute with the highest IG to split the tree (to generate nodes on the next level)**.** From table attribute "car's age" is the highest IG. So attribute "car's age" is selected to split to root of tree show in figure 1.

Step4 : Continue step (1) to (3) recursively to the next level of the tree until the leaf is pure (E = 0), or no more candidate attribute to work on. The final decision tree with ID3 approach shows in figure 2.2.

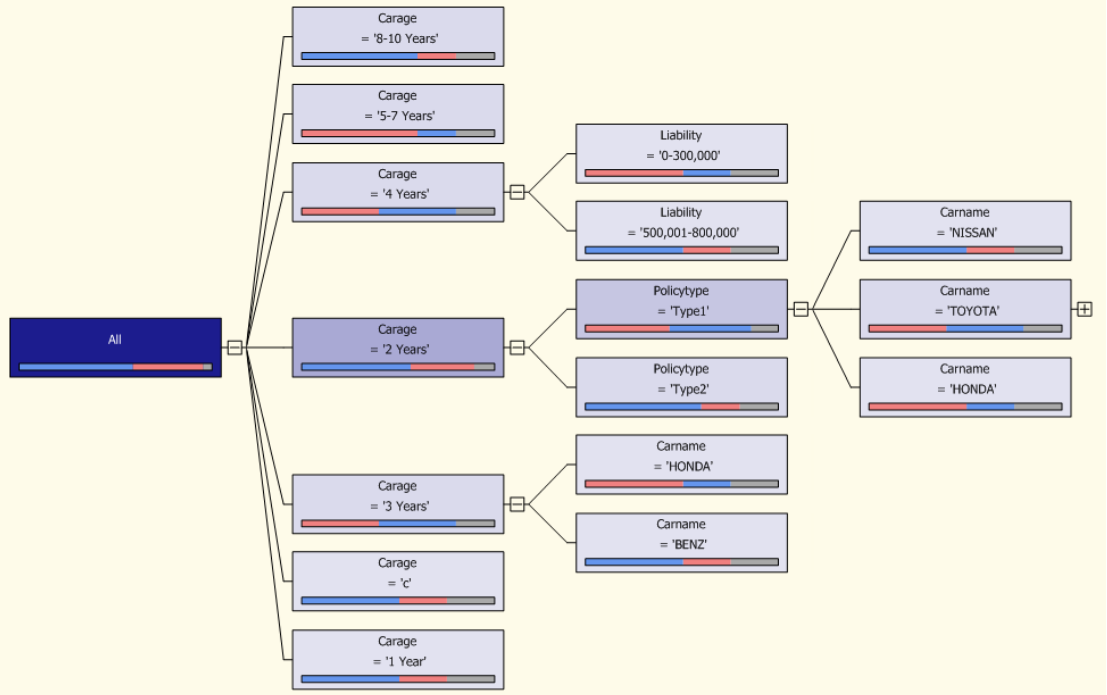


Figure 2.2 Final decision tree with ID3 approach

At each node the tree splits shows values all states with color probabilities. For safety level show pink color, blue color show as loss level and gray color shows as missing. The splits are chosen so that the sub nodes are more homogeneous. In this case we are trying to classified risk level include safety level, medium level and high level. The final nodes are called leaves. By tracing the splits or decisions that lead to each leaf we obtain classification rules. For example in table2.4 rule as follow:

60% of Customers who have car's age are 8-10 years is classified to safety level, without loss of claim.

60% of Customer who have car's age are 5-7 years is classified to safety level, with 20% of this segment is classified to loss level and rest is missing.

The result rules from decision tree returns the data shown in Table 2.4. Each row in result table contains statistics, support and the probability. The support is the number of cases similar in the dataset. For example Customer who has age of car between 8 and 10 years and has same no claim history. The probability indicates how likely it is for it to outcome.

Table 2.4 Rule of example insurance risk model

Rule	Condition	Outcome	Support	Probability
1	If carage = 8-10 years	Safety Loss Missing	2 0 0	60% 20% 20%
2	if carage = 5-7 years	Loss Safety Missing	2 0 0	60% 20% 20%
3	If carage = 4 years and Liability 0-300,000	Loss Missing Safety	1 0 1	40% 20% 40%
4	If carage = 4 years and Liability not 0-300,000	Safety Loss Missing	1 0 0	50% 25% 25%
5	if carage = 2 years and Policy type = Type1 and carname = Nissan	Safety Loss Missing	1 0 0	50% 25% 25%
6	If Carage = '2 Years' and Policytype = 'Type1' and Carname = 'TOYOTA'	Safety Loss Missing	1 1 0	40% 40% 20%
7	If Carage = '3 Years' and Carname not = 'BENZ'	Loss Safety Missing	1 0 0	50% 25% 25%
8	If Carage = '3 Years' and Carname = 'BENZ'	Loss Missing Safety	0 0 1	25% 25% 50%
9	If Carage = '1 Year'	Safety Loss Missing	1 0 0	50% 25% 25%

Advantage building decision tree with learning algorithm we obtain smaller tree and fewer paths as show in figure 1. Building a decision tree without learning algorithm and six attributes and each possible value as follow:

1. Policy are 3 possible values
2. Liability are 4 possible values
3. Car name are 4 possible values
4. Car type are 4 possible values
5. Car's use are 2 possible values
6. Car age are 6 possible values
7. The tree will have 3 × 4 × 4 × 4 × 2 × 6 paths

2.5.2 Evaluate method objective is to identify the performance of model. Performance of models can be evaluated using well known standard methods. These methods can use to evaluation approaches.

1. Confusion matrix (Kohavi and Provost, 1998) contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. From figure 2.3 shows a confusion matrix for a given data mining model. The table can be as follows:

1) TN (True Negatives) is the number of prospects predicted as being non-responders which are actually non-responses.

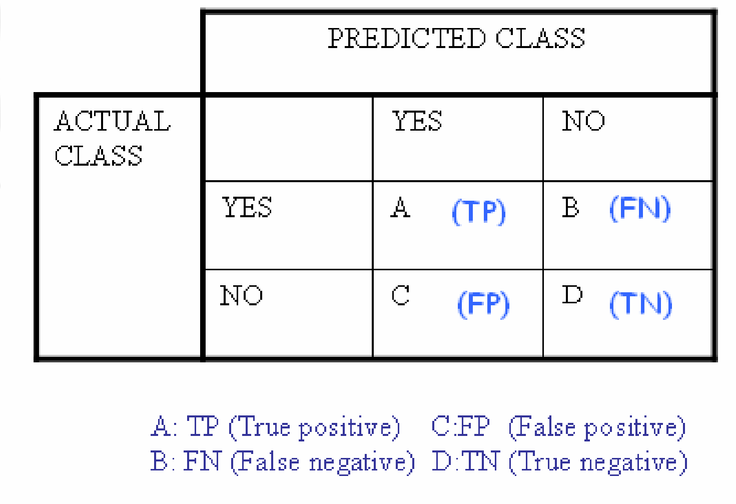
2) FN (False Negatives) is the number of prospects predicted as being non-responders which are actually responders.

3) FP (False Positives) is the number of prospects predicted as being responders, which are actually non-responders.

4) TP (True Positives) is the number of prospects predicted as being responders, which are actually responders.

Most widely-used metric is accuracy defined by the following formula:

$$\textbf{Accuracy} \qquad \frac{TP + TN}{TP + TN + FP + FN}$$



		PREDICTED CLASS	
ACTUAL CLASS		YES	NO
	YES	A (TP)	B (FN)
	NO	C (FP)	D (TN)

A: TP (True positive) C:FP (False positive)
B: FN (False negative) D:TN (True negative)

Figure 2.3 Metrics for Performance Evaluation

The accuracy determined using equation 1 may not be an adequate performance measure when the number of negative cases is much greater than the number of positive cases (Kubat et al., 1998). Suppose there are 1000 cases, 995 of which are positive cases and 5 of which are negative cases. If the system classifies them all as negative, the accuracy would be 99.5%, even though the classifier missed all positive cases.

2. Lift charts is a measure of the effectiveness of predictive model calculated as the ratio between the results obtained with and without the predictive model**.**

1) Cumulative gains and lift charts are visual aids for measuring model performance

2) Both charts consist of a life curve and a baseline

3) The greater the area between the life curve and the baseline the better the model

4) Generating a lift chart

a) Given a scheme that output probability, sort the instances in descending order according to the predicted probability

b) Select the instance subset starting from the one with the highest predicted probability

c) In the lift chart x-axis is sample size and y-axis is number of true positives

2.5.3 Process of data mining standards

In order to conduct data mining analyses, a general process is useful. This section describes data mining standards are concerned with process and API standards issues as shown in table 2.5.

Table 2.5 Process and APIs of data mining standards

Areas	Data Mining Standard	Description
Process Standards	-CRISP-DM methodology	● Industry mining tool and application neutral standard for business ● Hierarchical breakdown process ● provides framework for applying KDD consistently
	-Fayyad's KDD	The main focus lies in the data-mining phase. Vertical Solutions ● Develop by academy ● Iterative process ● KDD process highly domain dependent
Standard APIs	-Java API (JSR-73) -Microsoft OLE-DB	Java API for Data Mining applications Microsoft API for Data Mining

For develop data mining projects typically employ a formal and iterative process which guides step-by-step. Process method can be use as follow:

1. Generic Model of data mining project consists of three phases

1) The initial exploration

2) Model building or pattern identification with validation

3) Deployment

2. Fayyad’s KDD, Fayyad considers knowledge discovery in databases (KDD) and data-mining as two different concepts. He defines KDD as the general process of identifying valid, novel, potentially useful, and understandable structures in data (Fayyad, 1996). More specifically, it is the overall process of selecting and preparing data, selecting projections, selecting data-mining methods, extracting patterns, evaluating patterns as potential knowledge,

and consolidating this knowledge (see Figure 2.4). Nine-step model by Fayyad and colleagues as follows :

1) Developing and Understanding of the Application Domain.

It includes learning the relevant prior knowledge, and the goals of the end-user of the discovered knowledge.

2) Creating a Target Data Set

It selects a subset of variables (attributes) and data points (examples), which will be used to perform discovery tasks. It usually includes querying the existing data to select the desired subset.

3) Data Cleaning and Preprocessing

It consists of removing outliers, dealing with noise and missing values in the data, and accounting for time sequence information and known changes.

4) Data Reduction and Projection

It consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data.

5) Choosing the Data Mining Task

It matches the goals defined in step 1 with a particular DM method, such as classification, regression, clustering, etc.

6) Choosing the Data Mining Algorithm

It selects methods for searching patterns in the data, and decides which models and parameters of the used methods may be appropriate.

7) Data Mining

It generates patterns in a particular representational form, such as classification rules, decision trees, regression models, trends, etc.

8) Interpreting Mined Patterns

It usually involves visualization of the extracted patterns and models, and visualization of the data based on the extracted models.

9) Consolidating Discovered Knowledge

It consists of incorporating the discovered knowledge into the performance system, and documenting and reporting it to the interested parties. It also may include checking and resolving potential conflicts with previously believed knowledge.

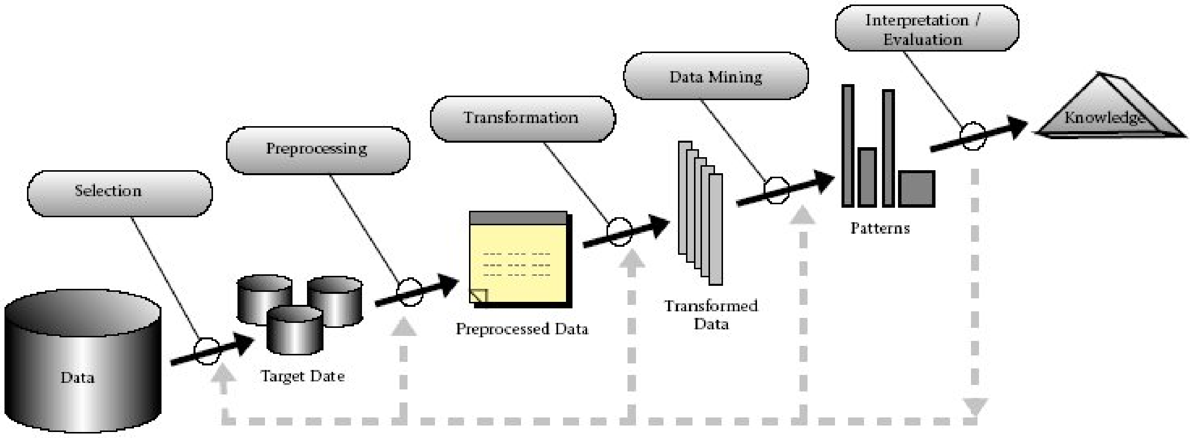


Figure 2.4 The KDD phases.

The KDD process is most of the time an iterative process that contains significant loops between phases until proceeding to a higher phase. The main focus lies in the data-mining phase. All other steps are equally important in order to gain appropriate and beneficial results.

3. CRISP-DM is the Cross-Industrial Standard Process for data mining proposed by a European consortium of companies in the mid-1990s (Chapman,P. et al.,2000). CRISP is based on business and data understanding, then data preparation and modeling, and then on to evaluation and deployment. It was conceived in late 1996 by DailerChrysler, SPSS and NCR. Cross-industry Standard Process of Data mining (CRISP-DM) model consists of six phases (CRISP-DM Group,1996).

1) Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

2) Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

3) Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

4) Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

5) Evaluation

At this stage in the project you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

6) Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions will need to be carried out in order to actually make use of the created models.

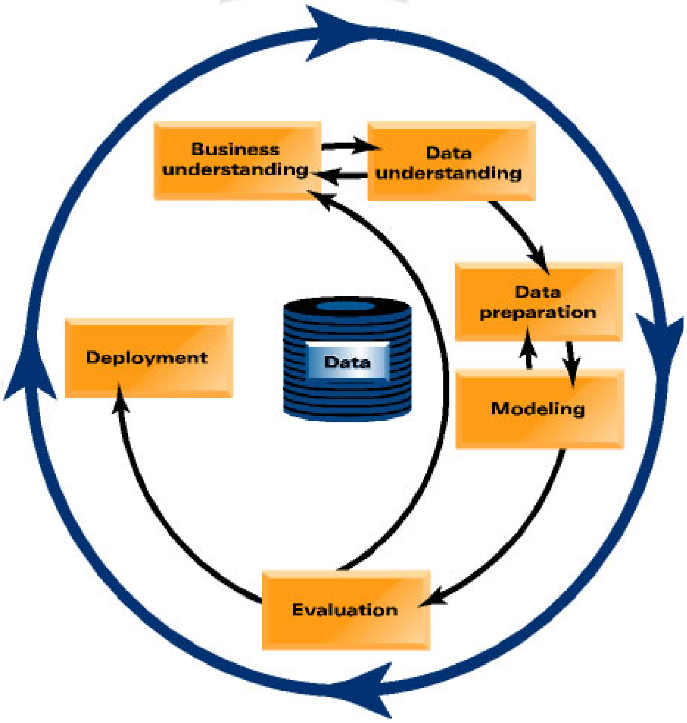


Figure 2.5 The CRISP-DM phases. (CRISP-DM Group,1996)

CRISP–DM is one method for fitting the data mining process into the overall business or research plan of action.

Aim :

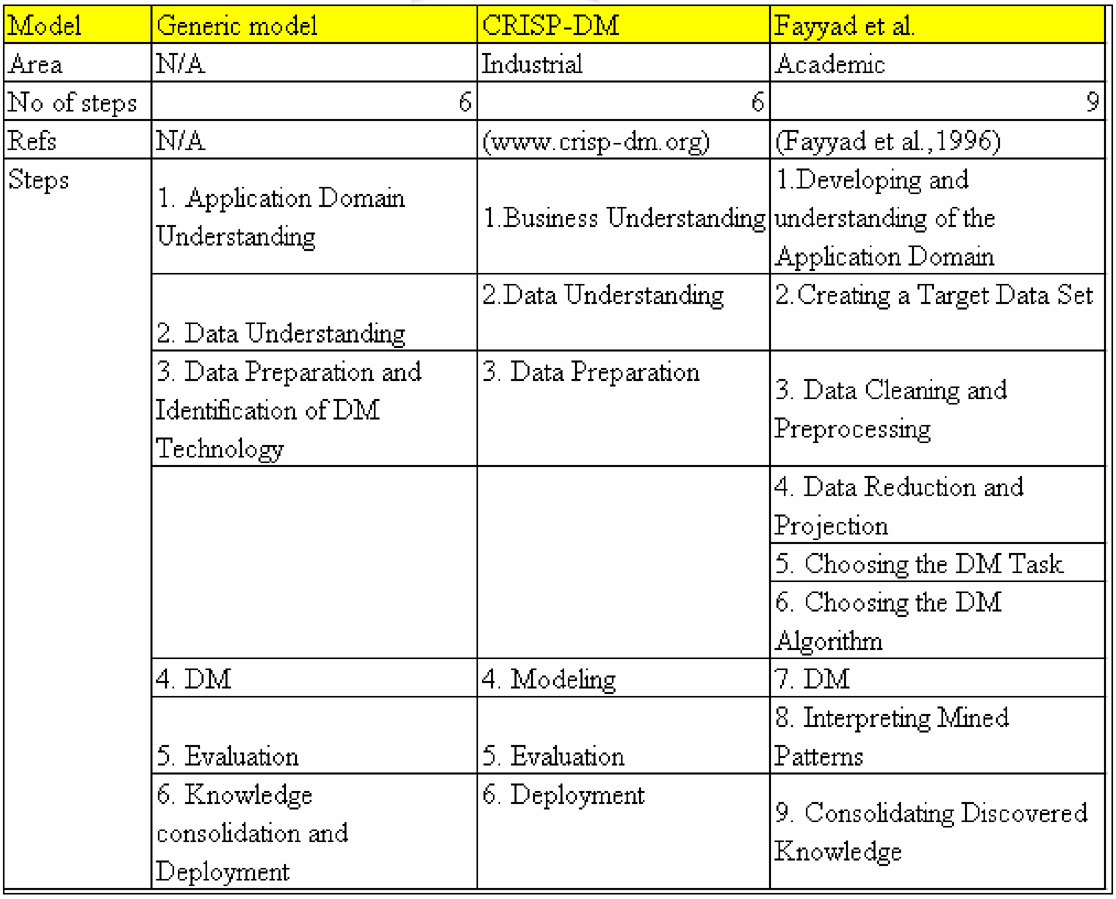
1. To develop an industry, tool and application neutral process for conducting Knowledge Discovery

2. Define tasks, outputs from these tasks, terminology and mining problem type characterization

3. Acknowledges the strong iterative nature of the process with loops between several of the steps

4. Designed to provide guidance to data mining beginners and to provide a generic process model.

5. It focus on business issues



Model	Generic model	CRISP-DM	Fayyad et al.
Area	N/A	Industrial	Academic
No of steps	6	6	9
Refs	N/A	(www.crisp-dm.org)	(Fayyad et al.,1996)
Steps	1. Application Domain Understanding	1.Business Understanding	1.Developing and understanding of the Application Domain
	2. Data Understanding	2.Data Understanding	2.Creating a Target Data Set
	3. Data Preparation and Identification of DM Technology	3. Data Preparation	3. Data Cleaning and Preprocessing
			4. Data Reduction and Projection
			5. Choosing the DM Task
			6. Choosing the DM Algorithm
	4. DM	4. Modeling	7. DM
	5. Evaluation	5. Evaluation	8. Interpreting Mined Patterns
	6. Knowledge consolidation and Deployment	6. Deployment	9. Consolidating Discovered Knowledge

Figure 2.6 Comparison of process data mining

2.5.4 Standard API perform to deploy model in to production environment

1. Java Data Mining API

1) Support for data mining APIs on J2EE platforms

2) To provide for data mining systems what JDBC did for relational databases

3) Build, manage, and score models programmatically

4) Data Mining clients can be coded against a single API that is independent of the underlying data mining system / vendor

2. OLE DB for Data Mining

1) OLE DB for DM was initialized by Microsoft

2) Works within SQL Server database suite

3) Using OLE DB for DM use a SQL-style query language for data mining

4) Limited to analytics provided by vendor

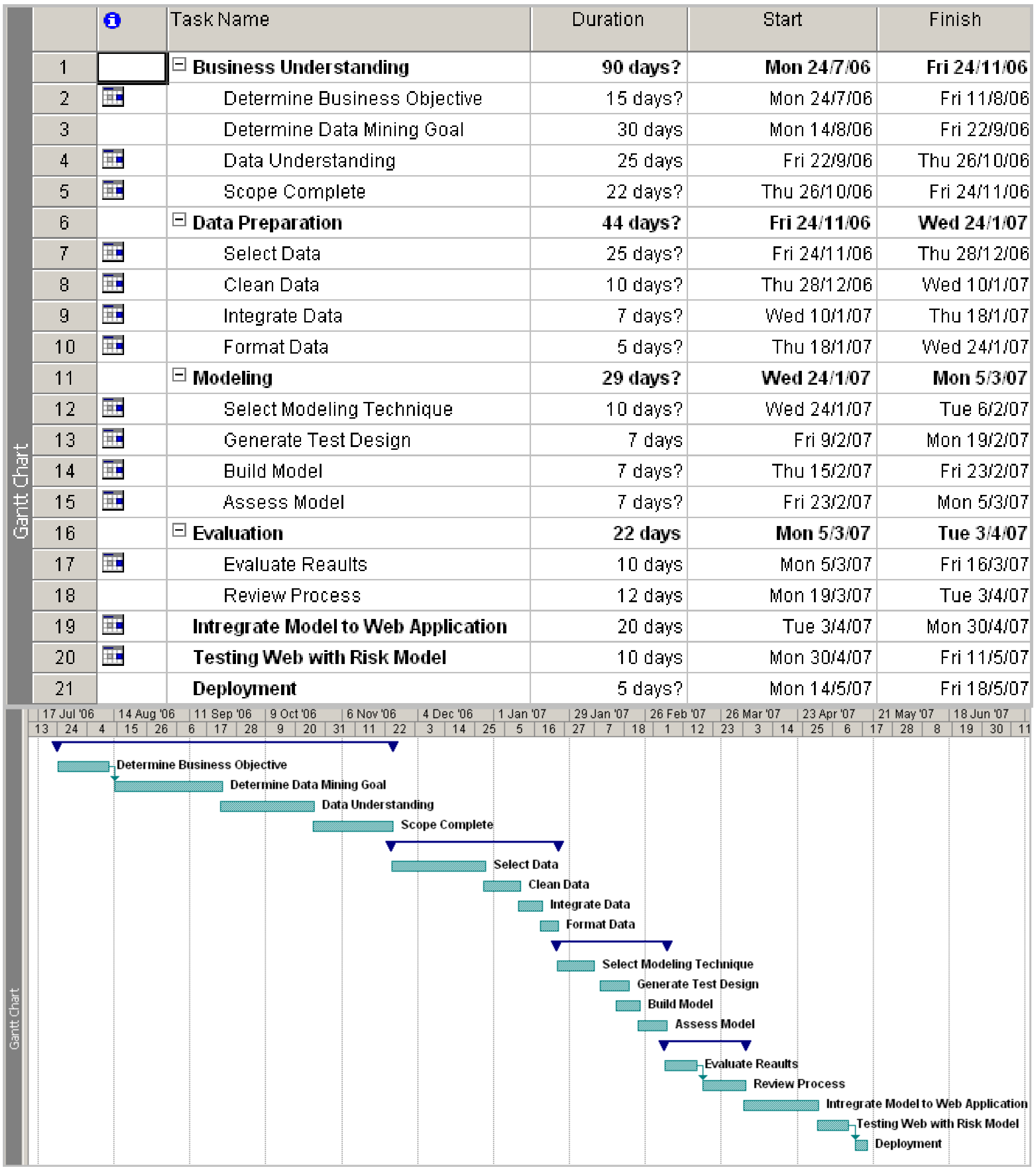
2.6 Deliverables

2.6.1 CD containing source code

2.6.2 User Manual /System Administrator Manual/Documentation

2.6.3 Web-based Deployment include insurance scores

2.7 Implementation Plan



	ℹ	Task Name	Duration	Start	Finish
1		⊟ **Business Understanding**	**90 days?**	**Mon 24/7/06**	**Fri 24/11/06**
2	▦	Determine Business Objective	15 days?	Mon 24/7/06	Fri 11/8/06
3		Determine Data Mining Goal	30 days	Mon 14/8/06	Fri 22/9/06
4	▦	Data Understanding	25 days	Fri 22/9/06	Thu 26/10/06
5	▦	Scope Complete	22 days?	Thu 26/10/06	Fri 24/11/06
6		⊟ **Data Preparation**	**44 days?**	**Fri 24/11/06**	**Wed 24/1/07**
7	▦	Select Data	25 days?	Fri 24/11/06	Thu 28/12/06
8	▦	Clean Data	10 days?	Thu 28/12/06	Wed 10/1/07
9	▦	Integrate Data	7 days?	Wed 10/1/07	Thu 18/1/07
10	▦	Format Data	5 days?	Thu 18/1/07	Wed 24/1/07
11		⊟ **Modeling**	**29 days?**	**Wed 24/1/07**	**Mon 5/3/07**
12	▦	Select Modeling Technique	10 days?	Wed 24/1/07	Tue 6/2/07
13	▦	Generate Test Design	7 days	Fri 9/2/07	Mon 19/2/07
14	▦	Build Model	7 days?	Thu 15/2/07	Fri 23/2/07
15	▦	Assess Model	7 days?	Fri 23/2/07	Mon 5/3/07
16		⊟ **Evaluation**	**22 days**	**Mon 5/3/07**	**Tue 3/4/07**
17	▦	Evaluate Reaults	10 days	Mon 5/3/07	Fri 16/3/07
18		Review Process	12 days	Mon 19/3/07	Tue 3/4/07
19	▦	**Intregrate Model to Web Application**	20 days	Tue 3/4/07	Mon 30/4/07
20	▦	**Testing Web with Risk Model**	10 days	Mon 30/4/07	Fri 11/5/07
21		**Deployment**	5 days?	Mon 14/5/07	Fri 18/5/07

Figure 2.7 The implementation plan

CHAPTER III

SYSTEM ANALYSIS AND DESIGN

In this section we use iterative approach to insurance risk within the Cross-industry Standard Process of Data mining (CRISP-DM) model which is the traditional data mining stage. Existing data warehouse of insurance company have large amounts of insurance data potential to build data set. The dataset used contains information about policies and insurance claims on those policies. Decision trees can be constructed to identify and describe areas of high risk which are then evaluated

3.1 Analysis of the Existing System

Generally, insurance underwriters evaluate the risk of the exist customer. They decide level risk with claim history of customers as shown in figure 3.1. They use history claim in database system to measuring risk and determining the premium that needs to be charged to insure. Each insurance company has its own set of underwriting guidelines to help the underwriter determine risk. This method can be used with new customer. For this method they evaluate risk with its own guidelines without information support decision.

Currently underwriter system implement by client server system computing. This creates an additional advantage to this architecture: All the data are stored on the servers. Servers can better control access and resources, to guarantee that only those clients with the appropriate permissions may access and change data. In client side, they used PC based. It difficult to deployment application for every client must have same version on underwriter application. It can be possible to various version of application in client.

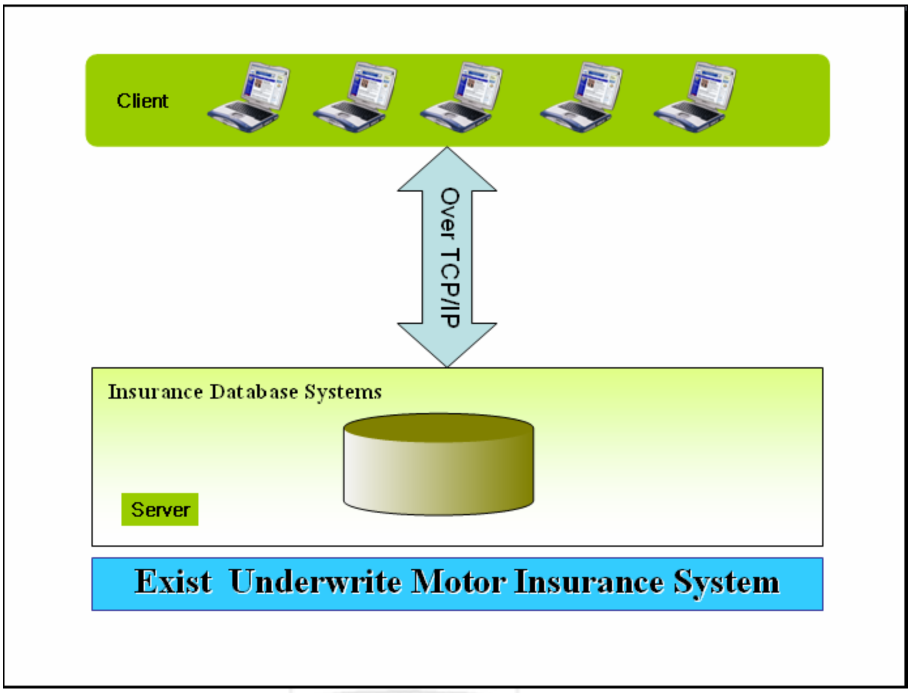

Client
Over TCP/IP
Insurance Database Systems
Server
Exist Underwrite Motor Insurance System

Figure 3.1 Exist underwrite motor insurance system

3.2 User Requirement Analysis

The business purpose is to identify the characteristics of automobile insurance claims that are more likely to be claimed. This requirement needs to know patterns of data about claim, “**What policy risk level of new customer?”** for evaluate risk level. They needs to make a decision based statistical probability of insurance claims.

In figure 3.2 currently the insurance data and claim are stored in databases. Each day, such data from all the branches are replicated into databases at the central branch. The amount of insurance data is increasing every day. OLAP helps organizations to find out the measures like sales drop, productivity, determine real-time product sales to make vital pricing and distribution decisions etc. Simply, OLAP tell us 'What has happened'. OLAP (Online Analytical Processing) provides a very good view of what is happening, but can not predict what will happen in the future or why it is happening. On the other hand, Data Mining can also be used to predict 'What will happen in the future' with the help of data patterns available within the organization.

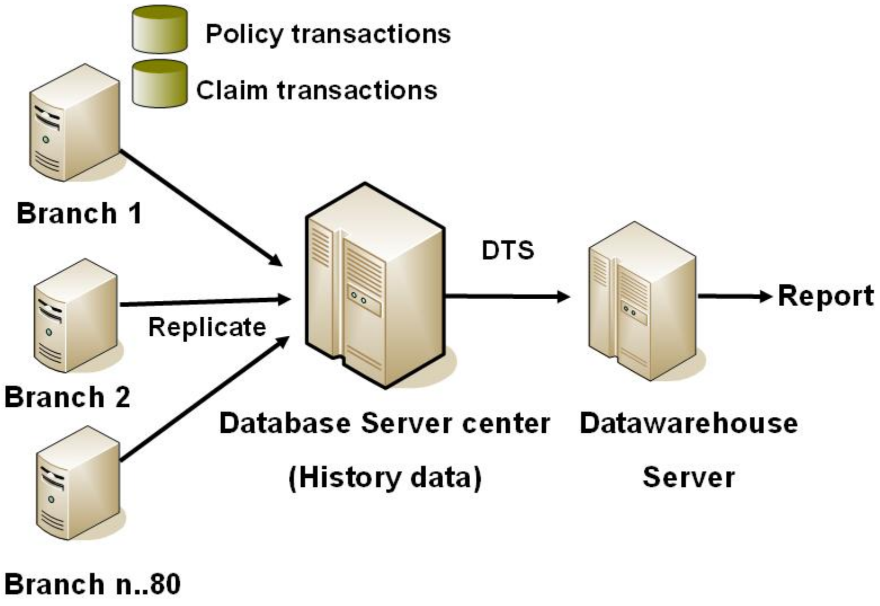


Figure 3.2 Data warehouse insurance systems

From Table 3.1 OLAP report can tell big picture of information but it doesn't tell about what's characteristic of claim data. The number count policy is large number if employee makes report display attribute of data that it difficult to understand data.

Table 3.1 OLAP Report of Case claim by year 2002-2003

Year	Count Policy	Premium	Count Claim	Claim
2003	354,207	5,547,345,030.208	258,655	4,347,414,728.03
2002	378,475	5,050,466,495.20	223,590	3,918,769,363.68

The insurance company collects data for each policy they underwrite and claim data. By this way we can use these data to build risk model by using technique data mining to extract pattern. It measures the level of risk of being claimed. With much information at hand, insurers can evaluate risk of customer better than no information to support decision. Risk pattern can be applied to underwriter system for classify a new customer into safe or risky group based on historical data alone.

Underwriter required system that can tell information about new customer who buys new policy compare with historical data is safety or risk because new customer underwriter haven't data to assets risk or safely. To identify the characteristics of motor insurance claims that is more likely to be claimed. They want to improve your assessment with mining technology to new applicant.

3.3 New System Design

In this section we are design system for implement data mining task include data extraction, transformation and loading (ETL) process. The design of schemas used in Analysis Services for train data set in mining process. Deploy risk model to client front-end systems with OLE-DM API.

3.3.1 Architecture Mining Service

In Figure 3.3, shows the client and server components of the architecture mining service. For building risk model we use Microsoft analysis services tools for data mining to create predictions. Client side connects to Analysis Services Server through DM provider by using ADOMD.NET (ActiveX Data Objects for .NET) that access to analysis services data object with C# language. It's enables to define data mining function display mining results on the client. The component of server includes two functions: Ms Analysis Server and Ms Integration Service.

1. Microsoft SQL Server 2005 Analysis Services provides tools for data mining. It can run data through and algorithm that generates a mathematical model of the data, a process that is call training the model. Mining bases are stored on the server, which allows access from different clients. It works in client-server architecture, allowing clients to connect by using IIS through HTTP through the Internet. The component of Analysis Services that perform the following :

Step 1: Creating risk insurance Data source

Step 2: Building Model

Step 3: Creating a Microsoft Decision tree

Step 4: Specific algorithm parameter

Step 5: Testing the Accuracy of the Mining Models

Step 6: Viewing the Lift Chart

Step 7: Evaluate model by Classification Matrix

Step 8: Package for risk mining

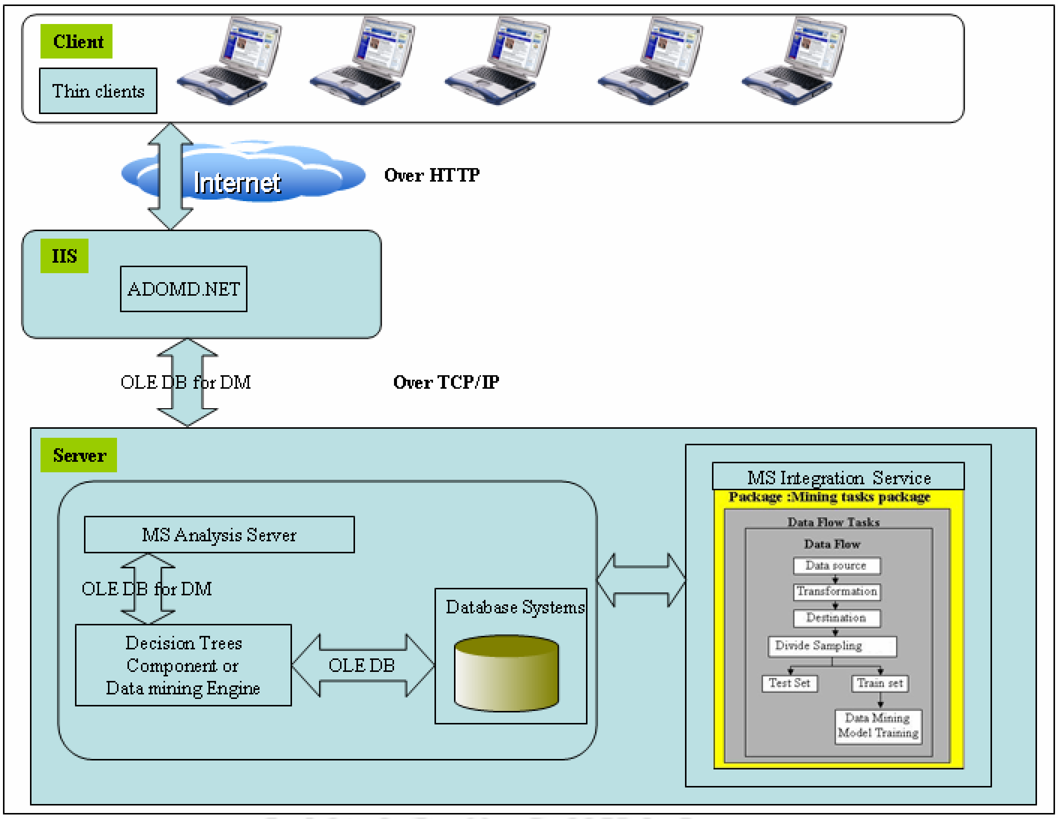


Figure 3.3 Architecture mining service

2. Integration Service is a platform for build data integration including extraction, transformation and loading operation. This function we can build package; tasks for contain process of mining task that performs the aggregation and extracted from destination data. Package is set of mining functions that control the execution of mining runs and results. Sequences of functions and mining operations can be defined and executed by using the user interface. This package that it is facilitate for administrator privilege to build risk model. Administration need to repeat process data mining task such as extract, transform, and load (ETL) processing for data. In figure 3.4 package control flow consists of two tasks. The first task is creates the input table to data. It’s consisting of loading data. The second task is a data flow that executes the data flow that performs the aggregation data and extracted from the destination data. The data flow identifies right format of data use for process model. The third tasks are divide data into training set and test set and process mining model.

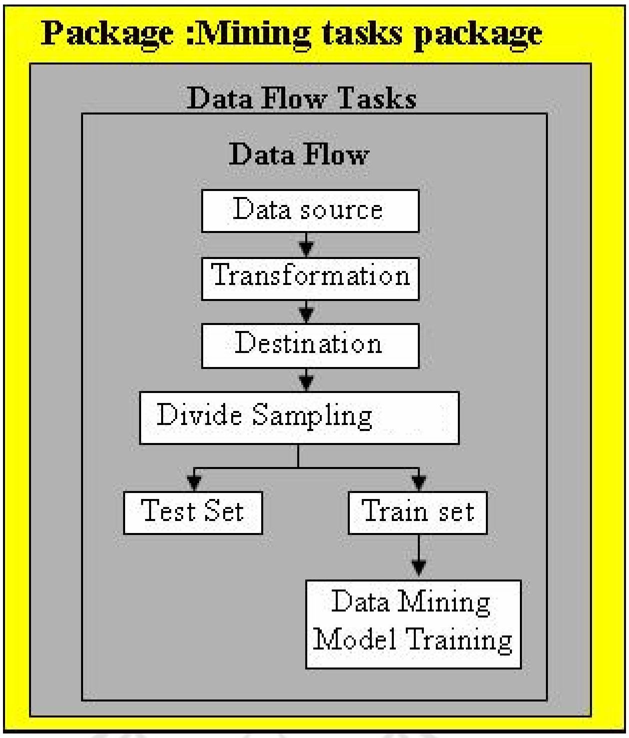


Figure 3.4 Package workflow for mining tasks

In table 3.2 show components in package. It's consisting of transformation task, divide sampling and execute model.

Table 3.2 Component in package

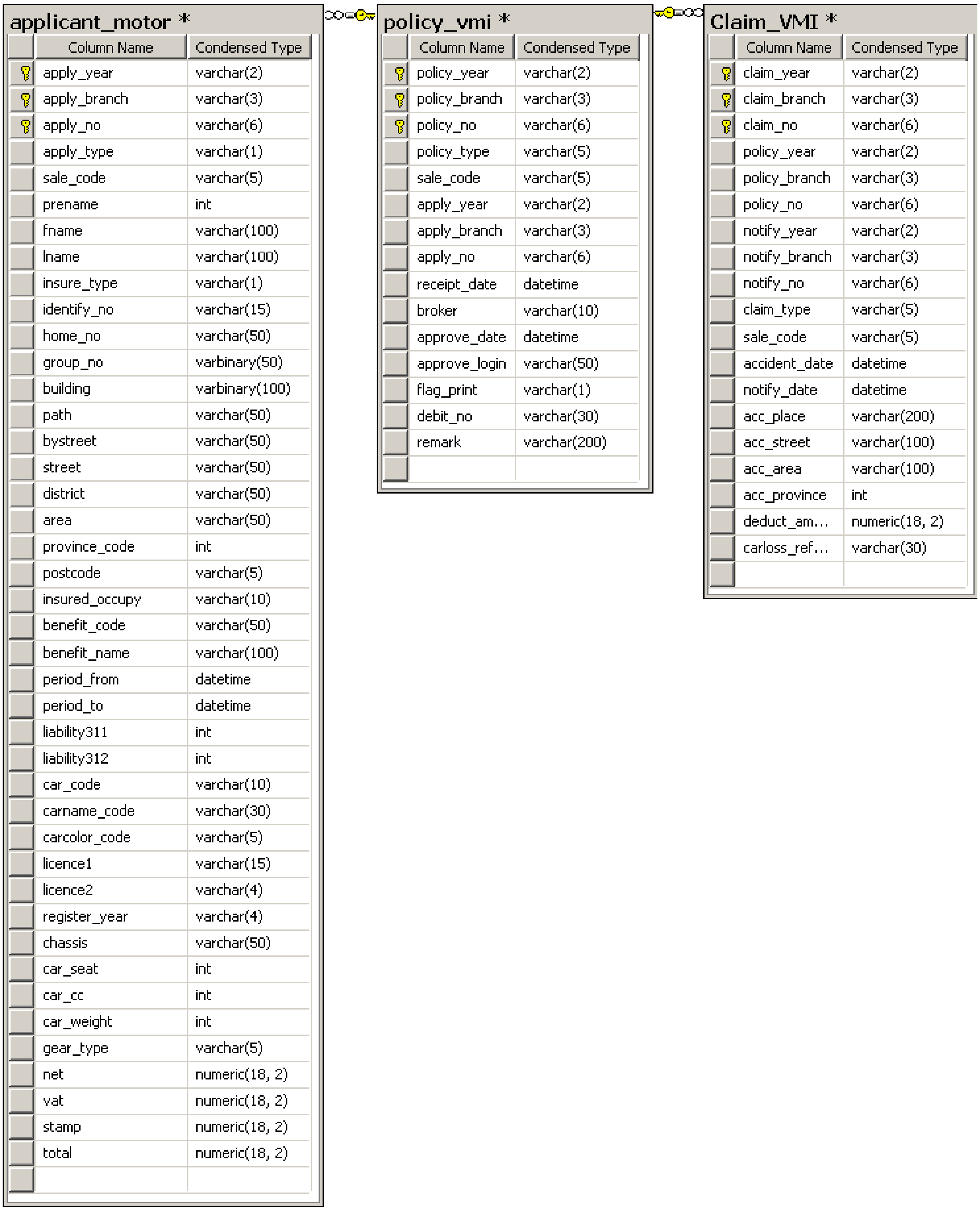
Element	**Purpose**
Transformation	Transform data into right format such as calculate risk level from pure premium and loss ratio.
Divide Sampling	Divide data to table training set and test set
Data Mining Model Training	Process training set with decision tree algorithm to build model

3.3.2 Database design for risk mining. In study we use policy and claim database include applicant motor table, policy vmi and claim transaction table for build insurance risk model as show in figure 3.5.

1. Applicant motor table contain detail of motor and customer data including some key elements such as insurer_type, address, car detail, address, premium, etc.

2. Policy_vmi table: contain detail of policy including some key elements such as policy no, apply no, debit_no, etc.

3. Claim Transaction table: contains records of claim history. It may include policy no, claim no, claim amount.



applicant_motor *

	Column Name	Condensed Type
🔑	apply_year	varchar(2)
🔑	apply_branch	varchar(3)
🔑	apply_no	varchar(6)
	apply_type	varchar(1)
	sale_code	varchar(5)
	prename	int
	fname	varchar(100)
	lname	varchar(100)
	insure_type	varchar(1)
	identify_no	varchar(15)
	home_no	varchar(50)
	group_no	varbinary(50)
	building	varbinary(100)
	path	varchar(50)
	bystreet	varchar(50)
	street	varchar(50)
	district	varchar(50)
	area	varchar(50)
	province_code	int
	postcode	varchar(5)
	insured_occupy	varchar(10)
	benefit_code	varchar(50)
	benefit_name	varchar(100)
	period_from	datetime
	period_to	datetime
	liability311	int
	liability312	int
	car_code	varchar(10)
	carname_code	varchar(30)
	carcolor_code	varchar(5)
	licence1	varchar(15)
	licence2	varchar(4)
	register_year	varchar(4)
	chassis	varchar(50)
	car_seat	int
	car_cc	int
	car_weight	int
	gear_type	varchar(5)
	net	numeric(18, 2)
	vat	numeric(18, 2)
	stamp	numeric(18, 2)
	total	numeric(18, 2)

policy_vmi *

	Column Name	Condensed Type
🔑	policy_year	varchar(2)
🔑	policy_branch	varchar(3)
🔑	policy_no	varchar(6)
	policy_type	varchar(5)
	sale_code	varchar(5)
	apply_year	varchar(2)
	apply_branch	varchar(3)
	apply_no	varchar(6)
	receipt_date	datetime
	broker	varchar(10)
	approve_date	datetime
	approve_login	varchar(50)
	flag_print	varchar(1)
	debit_no	varchar(30)
	remark	varchar(200)

Claim_VMI *

	Column Name	Condensed Type
🔑	claim_year	varchar(2)
🔑	claim_branch	varchar(3)
🔑	claim_no	varchar(6)
	policy_year	varchar(2)
	policy_branch	varchar(3)
	policy_no	varchar(6)
	notify_year	varchar(2)
	notify_branch	varchar(3)
	notify_no	varchar(6)
	claim_type	varchar(5)
	sale_code	varchar(5)
	accident_date	datetime
	notify_date	datetime
	acc_place	varchar(200)
	acc_street	varchar(100)
	acc_area	varchar(100)
	acc_province	int
	deduct_am...	numeric(18, 2)
	carloss_ref...	varchar(30)

Figure 3.5 Schema of policy and claim data

3.3.3 Design and deployment risk model of client systems. In figure 3.6 Deployment insurance risks scoring by Web Application Architecture is accessed with a web browser over a network such as the intranet. User use Web Browser to connect to the Intranet, requesting as ASP page. First ASP page user must authenticate with username and password and connects to the database. Risk model can be embedding to underwriter and generating a risk report. Predict shows the probability of an outcome, which guides the action on that customer

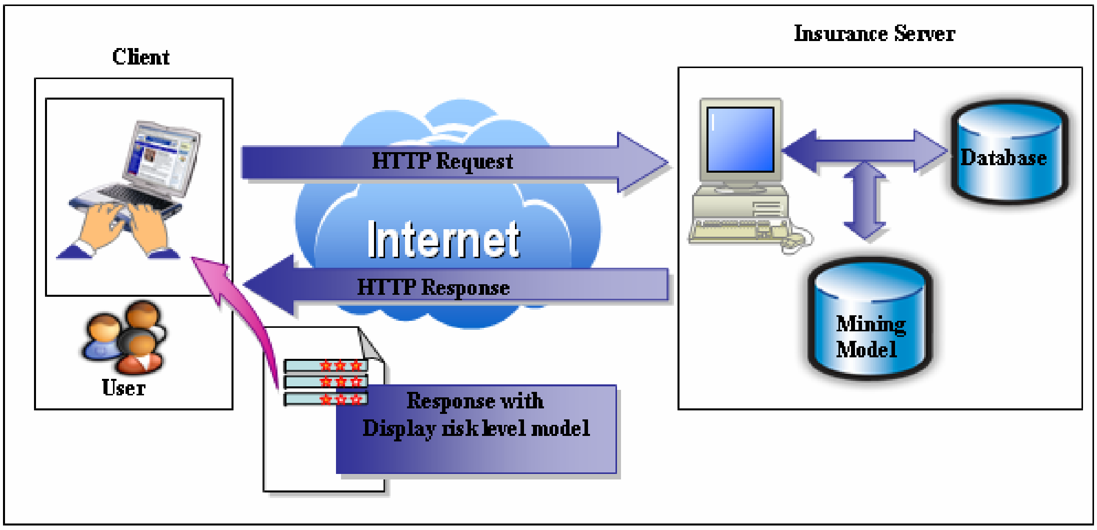


Figure 3.6 The interaction of user and the underwriter system via HTTP

In figure 3.7 shows process of transaction data after input about car detail, insurance detail and customer detail. Next step is preparing data and converts data into appropriate format such as clean values, filling missing values. Test data set with mining model to make a prediction. The results of the prediction are displayed in the Web page as risk scoring. A data prediction can be executed in real time.

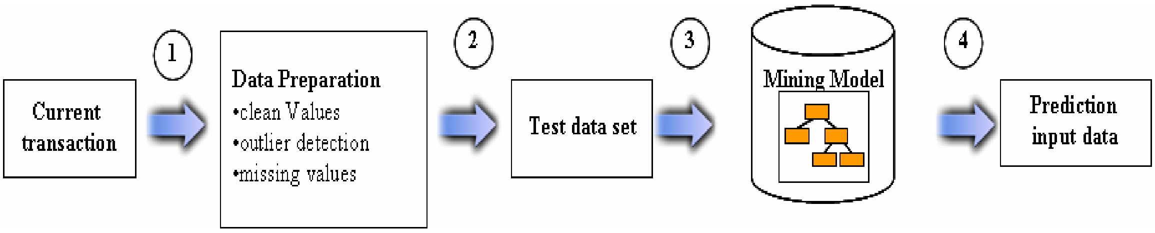


Figure 3.7 Process diagram of risk model for underwriter system

3.3.4 Deploy model by using OLE-DM API, in this section, we shows DMX connect and query risk model for display risk level. . The code presented here is written in C# and is embedded in the ASP.NET page.

Listing 3.1 Codes for connects to analysis services through ADOMD.NET.

```
using System;
using Microsoft.AnalysisServices.AdomdClient;
using System.Data;

AdomdConnection con;

static bool Create ADOMDConnection()
{
  con = new AdomdConnection (“Data Source=.; Catalog=MiningRisk; Integrated Security=True”);
  try
  {
     con.Open();
  }
  catch (System.Exception e)
  {
     Log(e.Message);
     return false;
  }
   Return true;
}
```

Use the risk model to make predictions. Predictions are made with a SELECT statement that joins the model's set of all possible cases with another set of actual cases. In listing 3.2 show prediction query statement.

Listing 3.2 Example prediction query statement

```
public string GenerateDMX()
{
  static string DMX= "SELECT flattened " +
      "PredictHistogram([46vibvmi Trainset].[Class Code])"+
      "From  [46vibvmi Trainset] " +
      "NATURAL PREDICTION JOIN "+
      "(SELECT "+ls_carkind +" AS [Car Kind], "+
       ls_caruse+" AS [Car Use], " +
       ls_carage_key+ " AS [Carage Key],"+
       ls_carname_key+ " AS [Carname Key],"+
       ls_carcode_key+ " AS [Car CodeKey],"+
       ls_liability_key+ " AS [LiabilityKey],"+
       ls_poltype_key+ "AS [Poltype Key]) AS t";
  return DMX;}
```

In listing 3.3 the ExecuteAndFetchSQL function executes the prediction query through AdomdCommand object and fetches the query result using DataReader object. Each predicted item in the record set is displayed in a table of the recommendation in web page.

This example is a prediction query, which predicts for the given customer whether he will be safety or risk in the motor insurance with probability, support and each possible predict level of cases.

Listing 3.3 Example prediction query statement

```
public bool ExecuteAndFetchSQL (string strCommand)
{
    AdomdCommand cmd = (AdomdCommand)
    Con.CreateCommand();
    cmd.Text = strCommand;
    IDataReader dr = null;
    try
    {
       dr = cmd.ExecuteReader();
    }
   catch (Exception e)

    {
       Log(e.Mssage);
       Return false;
     }
}
  …
//display the result in the web page
While(dr.Read())
{
    string val = dr.GetString(0);
    SuggestedList.Items.Add(val);
}
…
return true;
}
```

3.3.5 Risk mining permissions

To process model that contained inside the mining structure is not user-accessible. Only Administrator can read, write, browse and process mining structure. User can read perform prediction queries on the mining model. The types of permissions that can be assigned are described in Table 3.3

Table 3.3 Risk mining permissions

Permission	Effect
Read	User can perform prediction queries on the mining model
Write	User can update model
Browse	User can access the model
Process mining	User can process the mining structure

CHAPTER IV

SYSTEM FUNCTIONALITY

4.1 Introduction

This project designed to support accuracy of risk assessments. Risk model help underwriter avoid high risk customer with information support their evaluate risk. We implements risk insurance result into the insurance's business process. It enables the underwriter to act faster and more precise and therefore optimizes the company's profitability. The result from decision tree returns the data contains statistics; support and the probability of each customer depend on history data.

4.1.1 Scope of system functionality :

1. Data mining tasks for build model include of

1) Data transformation and data cleaning

2) Data aggregation

3) Build model

2. Prediction function with risk model in underwriter application

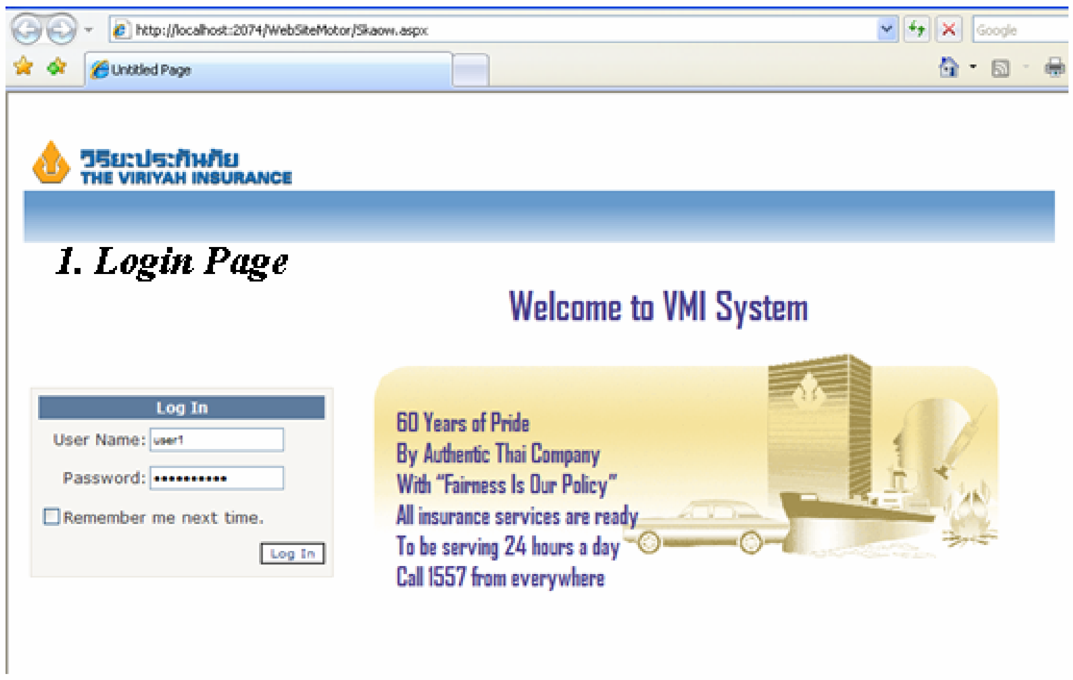
1) Log in - First page user must authenticate with username and password and connects to the database.

2) Risk model function for classify risk level. User entry data about car detail within underwriter application. The results of prediction are displayed as risk level.

3) Risk model generate report

4.1.2 Webpage layout

The main reason of design web page layout is the common way based on simple, easy to understand. In figure 4.1 Login page user use Web Browser to connect to the Intranet, requesting as ASP page. First ASP page user must authenticate with username and password and connects to the database.



http://localhost:2074/WebSiteMotor/Skaow.aspx

Untitled Page

วิริยะประกันภัย THE VIRIYAH INSURANCE

1. Login Page

Welcome to VMI System

Log In

User Name: user1

Password: ••••••••••

□ Remember me next time.

Log In

60 Years of Pride
By Authentic Thai Company
With "Fairness Is Our Policy"
All insurance services are ready
To be serving 24 hours a day
Call 1557 from everywhere

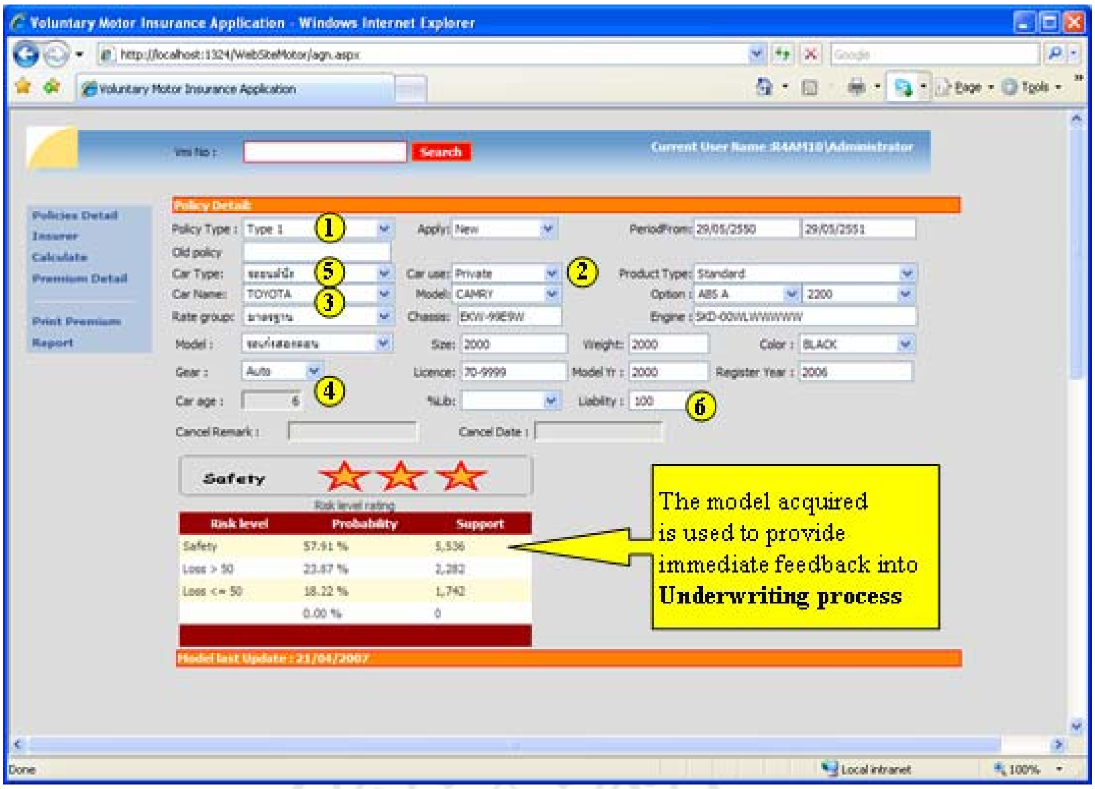
Figure 4.1 Web page design – login Page

4.1.3 Webpage layout – motor insurance Page

In figure 4.2 Motor insurance Page - Here is a main process page, page contain information about car detail. In this page underwriter user can perform prediction queries on the risk mining model by input data as shown in table 4.1.

Table 4.1 Fields description on entry motor insurance page

Field Name	Description	Example
Car Detail		
Policy Type	Select the type of policy: value can be Type1, Type 2, Type 3	Type 1
Car use	Type of use car value can be private or public	private
Car name	Name of car	TOYOTA
Car age	Age of car	6
Car type	Type of car	passenger
Liability	Amount of Liability	0-300,000



Voluntary Motor Insurance Application - Windows Internet Explorer

http://localhost:1324/WebSiteMotor/agn.aspx

Vmi No : Search   Current User Name :R4AM10\Administrator

Policy Detail

Policy Type : Type 1   Apply: New   PeriodFrom: 29/05/2550 29/05/2551

Old policy

Car Type: รถยนต์นั่ง   Car use: Private   Product Type: Standard

Car Name: TOYOTA   Model: CAMRY   Option : ABS A 2200

Rate group: มาตรฐาน   Chassis: EKW-99E9W   Engine : SKD-00WLWWWWW

Model : รถเก๋งสองตอน   Size: 2000   Weight: 2000   Color : BLACK

Gear : Auto   Licence: 70-9999   Model Yr : 2000   Register Year : 2006

Car age : 6   %Lib: Liability : 100

Cancel Remark :   Cancel Date :

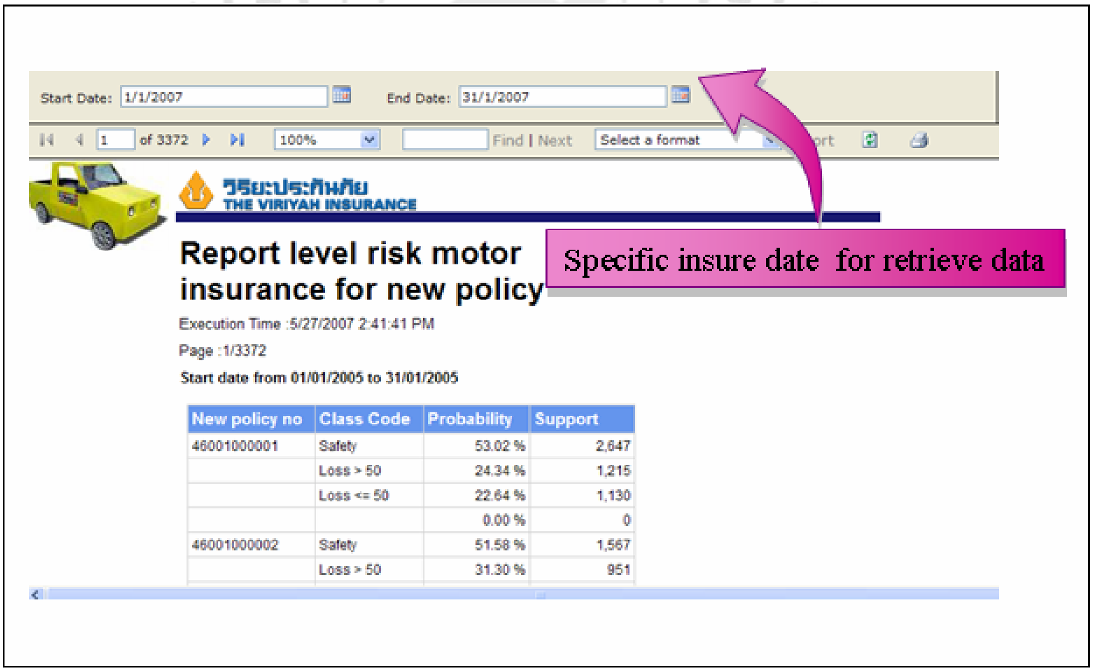
Risk level rating

Risk level	Probability	Support
Safety	57.91 %	5,536
Loss > 50	23.87 %	2,282
Loss <= 50	18.22 %	1,742
	0.00 %	0

Model last Update : 21/04/2007

The model acquired is used to provide immediate feedback into **Underwriting process**

Figure 4.2 Web page design - motor insurance page



Start Date: 1/1/2007   End Date: 31/1/2007

Specific insure date for retrieve data

Report level risk motor insurance for new policy

Execution Time :5/27/2007 2:41:41 PM

Page :1/3372

Start date from 01/01/2005 to 31/01/2005

New policy no	Class Code	Probability	Support
46001000001	Safety	53.02 %	2,647
	Loss > 50	24.34 %	1,215
	Loss <= 50	22.64 %	1,130
		0.00 %	0
46001000002	Safety	51.58 %	1,567
	Loss > 50	31.30 %	951

Figure 4.3 Web page design – report level risk motor insurance

4.2 System Architecture

In the architecture shown in Figure 4.4, we use client server architecture. In server side is used to execution of mining tasks. In client side contacts servers by DM provider to access to risk model for browsing or retrieving data to show risk level on client side.

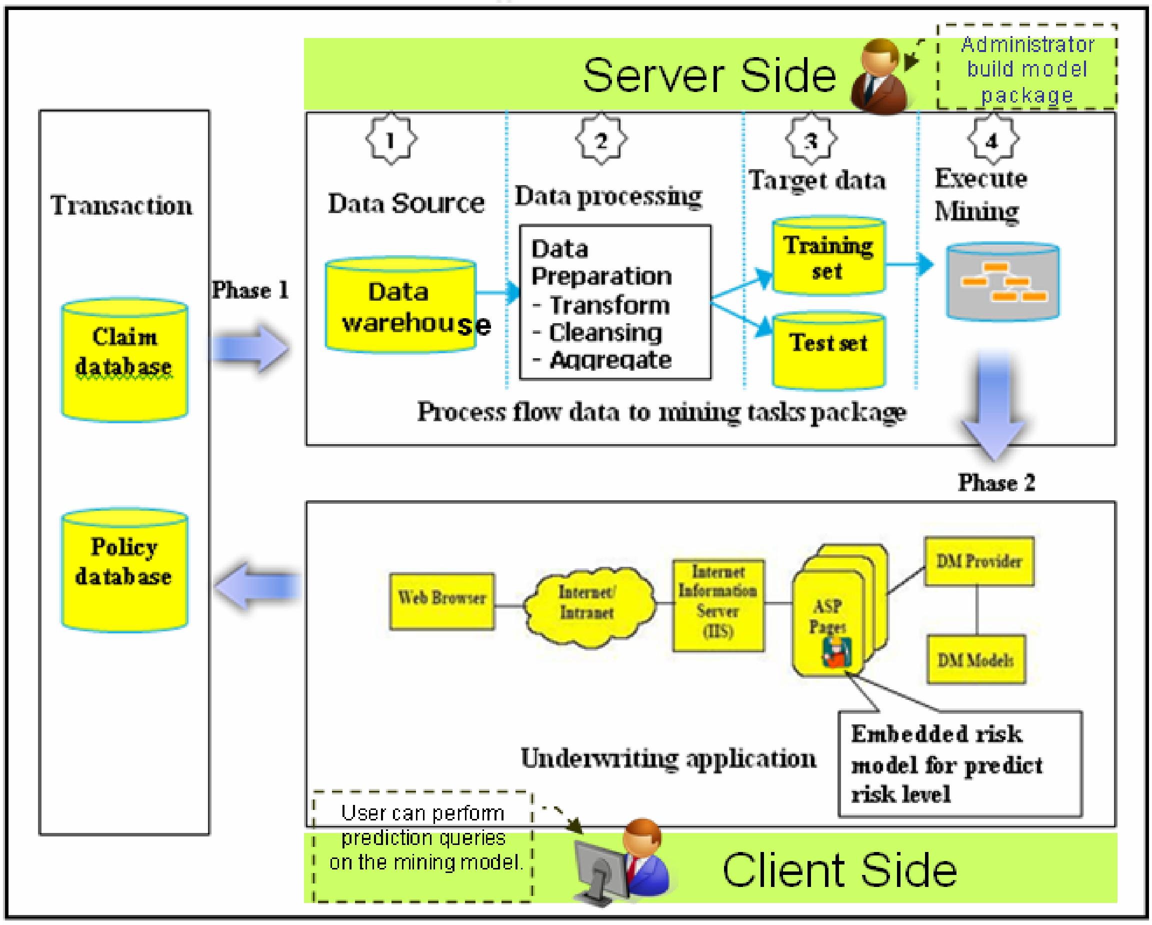


Figure 4.4 Architecture of a web-based underwriter application

4.2.1 Mining process task, in section for conduct data mining analyses in study risk insurance we use CRISP-DM methodology for guideline this study. The deployment phase can be as embedded risk model to underwriter application. Using OLE DB for Data Mining is APIs use a SQL-style query language for data mining. In server side contain package which collect data mining task such as data transformation, data cleaning data and aggregation data. This step is processed by administrator. Step of data mining process as follows:

1. Data Collection, for study we use raw data were extracted from the motor insurance data warehouse for all premium-paying policy holders with claim data. For data selection we select as following:

1) Customer data consist of address of customer such as province (Bangkok, Phuket) and occupation such as management, student, engineer, etc.

2) Vehicle data consist of age of car , vehicle manufacturer, car use such as private or public, car type such as passenger, truck, bus, trailer, ambulance, caravan, motorcycle and car gear

3) Policy data consist of policy no, policy type, liability and premium net

4) Claim data consist of amount of claim and count of claim

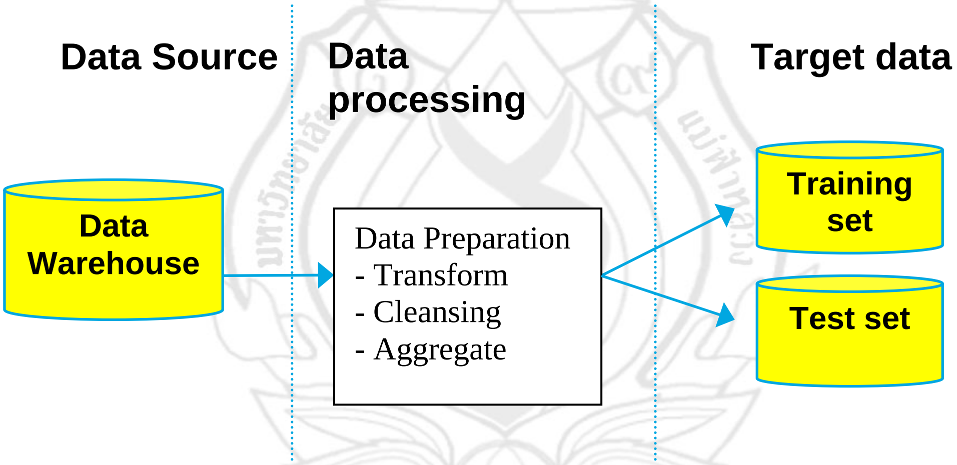


Figure 4.5 Data preparation process

2. Data processing, in figure 4.5 second tasks is data processing that is one of the more important tasks in data mining. It is also one of the most times consuming ones. Data should be converted into appropriate formats. Data preprocessing can be separated into two major tasks are data cleaning and data transformation.

1) Data cleaning purpose is to handling missing values or incomplete data, noisy and inconsistent data.

In case of miss value and incomplete data comes from no available data valued when collected or problems with human, software and hardware. Some methods that can be used to deal with missing data are :

a) Leave as it is

b) Ignore the instance with the missing value(s), i.e. remove the instance

c) Manually enter the missing value, assign a default value depending on the most implicit meaning

For example in variable occupation is empty or define to 'other' that not meaning. In this case replacement was not performed on the data set; in figure 4.6 most of the variables occupation were missing and define value to 'other' than 69% of their values (missing value/ all case = (433,211/627,341)*100 = 69 %) . So missing over 50% of its values was excluded in the variable selection process.

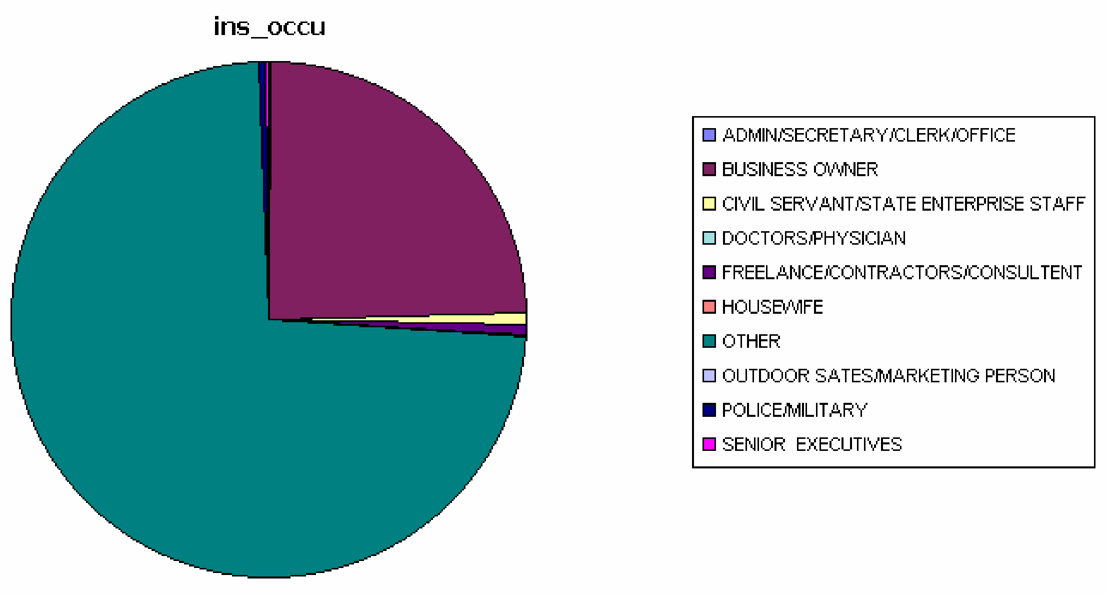


Figure 4.6 Variable occupations with incomplete data

Data noisy that data containing errors or outliers e.g., car's age = "-96" year. Noisy data comes from the process of data such as collection and entry. Most of the variables were noisy less than 5% of their values. In this case we recalculate car's age and assign correct value to variable car's age.

Data Inconsistent that containing discrepancies in codes or names e.g., Register year of car is empty or Register year =“-1997”. Inconsistent data comes from different data sources and functional dependency violation. Most of the variables were Inconsistent less than 5% of their values. In this case we manually entry correct value.

2) Data Transformation that aggregation or summarization loss ratio variable is added to the expense ratio to determine the company's combined ratio The combined ratio is a reflection of the company's overall underwriting profitability.

Net earned premium = earned premium– commission – Costs of contract

Claims = amount of claims – costs of claims

Loss ratio = Claims/Premium

Loss ratio variable should be able to rate three levels as follow:

a) Safety is customers without past claim records are classified as "safety".

b) Middle is customer have past claim history and loss between 1 to 50 % or very few claims are classified as medium risk level that little significant impact on pricing.

c) High is customer have past claim history and loss greater than 50% or very high pay per claim are classified as high risk level that primary impact on pricing.

The final set of variables used in the modeling is shown in Table 4.2.

Table 4.2 Variable selection use for build risk model

Variables	Represent
Pol No	Policy No
Poltype Key	Policy type
Carage Key	Car's age
Carcode Key	Car code
Carname Key	Car's name
Lib Key	Liability insurance
Car Gear	Car's Gear
Car Kind	Car Kind
Car Use	Car Use
Ins Occu	Insurer's occupation
Ins Province	Place of residence
Class Code	Level risk by rang of losses

3. Model is developed in two phases: training and testing. For training refers to building a new model by using historical data. Testing refers to trying out the model one new, to determine its accuracy and performance.

4. Executing Mining that select modeling Decision tree technique for building an insurance risk model the algorithm we using entropy-based and using Bayesian with K2 prior method to compare performance of model. In building a model, the algorithm examines how each variable in the dataset affects the result of the predicted attribute. Each node is creating based on the predicted attribute as compared to the input attributes. Develop initial risk model, based on outcomes of past historical data

4.2.2 Evaluating the risk model that has process of performance evaluation as shown in figure 4.7 is explained as follows :

1. Split data 900,510 instances into train and test sets
2. Training set has 600,340 instances
3. Testing set has 300,170 instances
4. Build a model on a training set by using decision tree technique
5. Evaluate on test set

Model evaluation accuracy by estimate the accuracy of the model based on set of testing data. The known class of testing data is compared with the classified result from the model. Accuracy rate is the percentage of testing data that are correctly classified by the model. There are two methods can use to evaluation approaches are Confusion matrix and Lift charts method

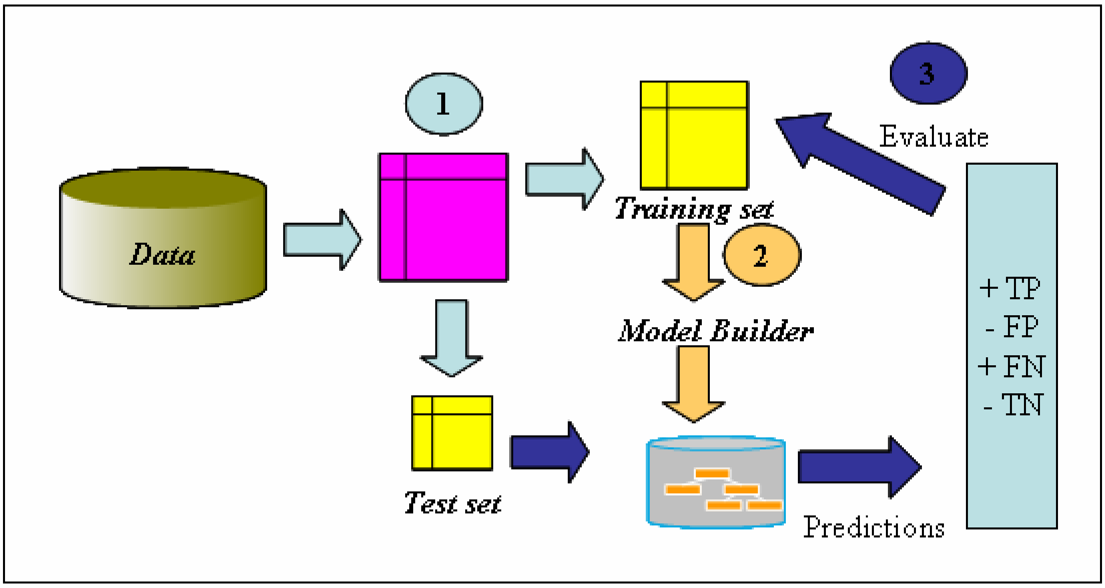


Figure 4.7 Process evaluate model

1. Confusion matrix method, after running decision tree with the ID3 algorithms generates grid from the test data. It considers a three class problem with the classes Safety, Medium, High. A predictive model may result in the following confusion matrix when tested on independent data. The results show in table 4.3 is interpreted as follows:

1) 195,742 instances of class Safety were correctly classified,

2) 47,241 instances of class Safety were misclassified as Medium

3) 57,187 instances of class Safety were misclassified as High

4) 0 instances of class Medium and High were misclassified

From table 4.3 see that there were actually predict correct 195,742 safety instances. There were predicting incorrect 47,241 medium instances and 57,187 high instances. Accuracy is the overall correctness of the model is 65.21 %. By accuracy of risk model is 65.21%, generally data set of insurance claims are very low events. Decision trees develop predictive models by use frequent value. The risk model is possible predict risky customer in to safety level. This type of models will have low accuracy in predicting risky customer to risk level. Apply risk model to use can be done for only suggestion risk information and generate possible risk value for each segment.

Table 4.3 Classification matrix for risk model on test data

Predicted	**Safety** **(Actual)**	**Medium** **(Actual)**	**High** **(Actual)**
Safety	**195,742**	47,241	57,187
Medium	0	0	**0**
High	**0**	0	0

Accuracy
$$\frac{TP + TN}{TP + TN + FP + FN}$$

As above formula accuracy is the overall correctness of the model and is calculated as the sum of correct classifications divided by the total number of classifications as follow :

TP = 195,742

TN = 0

FP = 0

FN = 47,241 + 57,187 = 104,428

$$\text{Accuracy} = \frac{195{,}742}{300{,}170} \text{ x } 100$$
$$= 65.21\ \%$$

2. Create the lift Chart is measure the effectiveness of risk model calculated probability for each case ration between the results obtained with and without the model. According to the Lift Chart shown in figure 4.8, the X-Axis shows the percentage of prospects we are looking at, and the Y-Axis shows the percentage of respondents we achieve.

1) To plot the chart: Calculate the points on the lift curve by determining the ratio between the result predicted by our model and the result using no model.

2) For the random sampling case, sampling 50% of the prospects gives us 50% of the respondents. An Ideal model would predict everything correctly. That is, the percent of correct prediction is the same as the percentage of predictions covered at any point.

3) For prospects ranked by the ideal model, choosing the top 50% of prospects gives us coverage 65.21% of the instances. Amount of instances in test data set model can find an answer might be correct 65.21 % of instances out of 100.

Response Rate = Number of correct / Total Number of instances

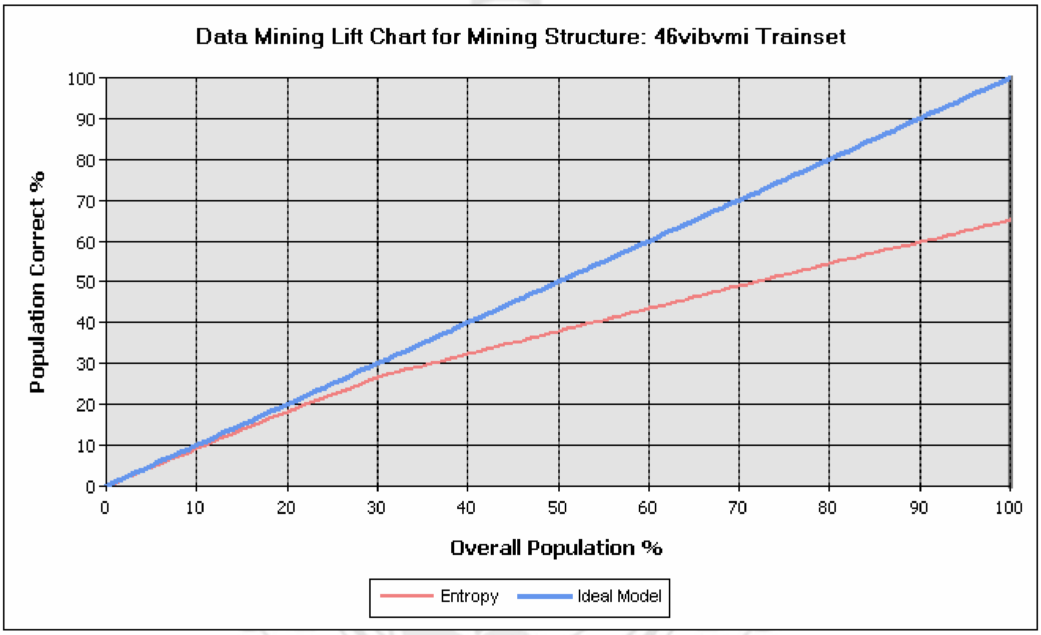


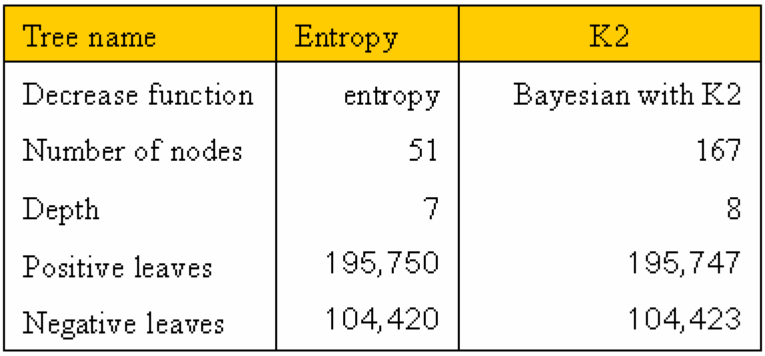
Figure 4.8 Lift chart for motor insure predictions

4.2.3 Comparison splitting with Entropy and Bayesian with K2, section show decision trees were generated using different selection measures for the data set we will refer to the entropy and Bayesian with K2 to calculate the impact of each attribute, and then select the most significant attributes. After running the decision trees algorithms, we obtained two models, In order to summarize the differences method to split trees. The result in this table represents the number of nodes by each method. In Table 4.4 shows the results as follows :

1. The number of nodes is number of rules. For each node in the tree, read the rule from the root to that node. We will arrive at a set of rules. There are 51 rules form the entropy and 167 rules from the Bayesian with K2.

2. Positive leaves are the number of correct predicted. The entropy can determine number of positive 195,750 cases. The Bayesian with K2 can determine number of positive 195,747 cases.

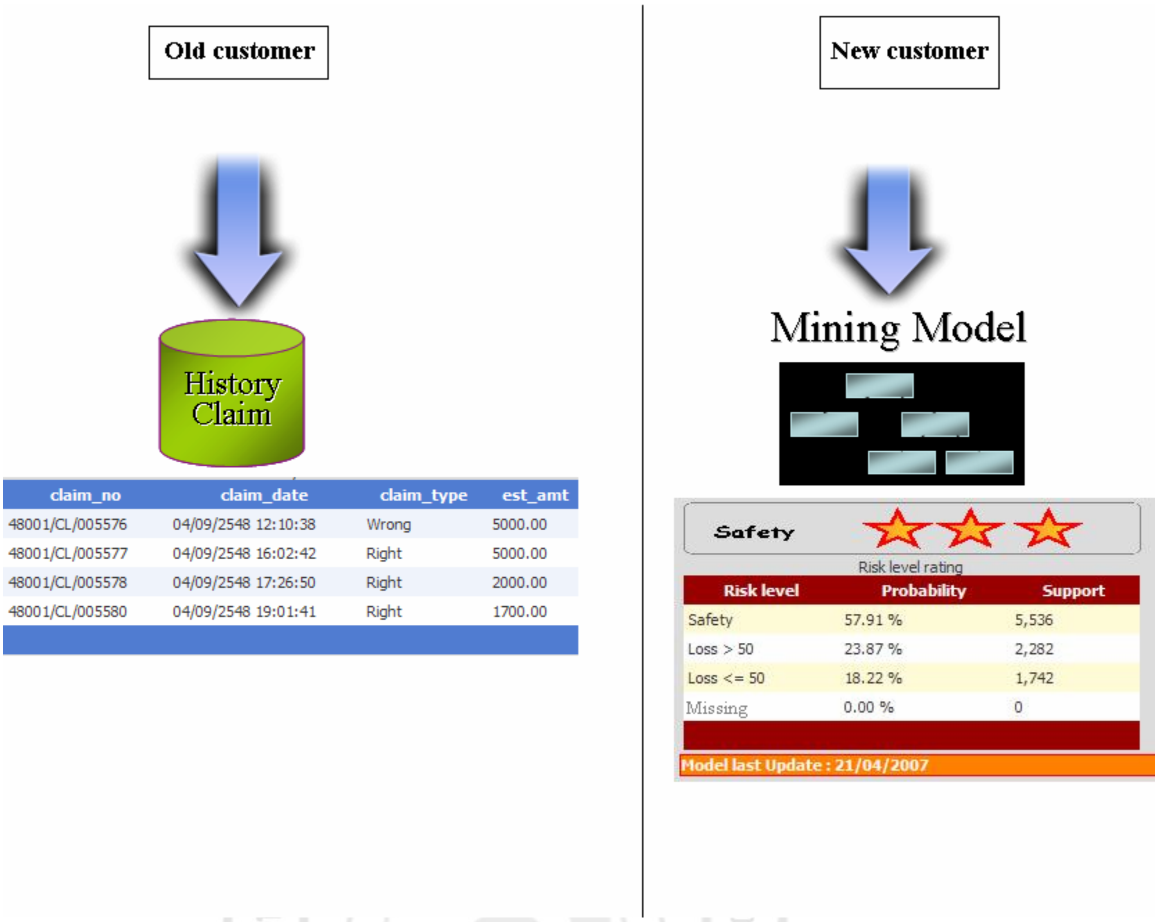
Table 4.4 Using alternative decrease functions



Tree name	Entropy	K2
Decrease function	entropy	Bayesian with K2
Number of nodes	51	167
Depth	7	8
Positive leaves	195,750	195,747
Negative leaves	104,420	104,423

From insurance data set with two search methods, entropy and Bayesian with K2 is achieved with relatively little change in accuracy. For smaller tree of entropy can searching faster than Bayesian with K2.

4.2.4 Deployment insurance risk scoring by Web Application Architecture is accessed with a web browser over a network such as the intranet. User immediately sees a list of possible level risk suggestions below the web page. The recommendations are based on other customers' insurance patterns in similar cases. These patterns are stored in either a decision trees model. In figure 4.9 risk method for assessment risk, risk model is used to determine new customer that assign to which level of risk group. For exist customer used their past history to evaluate risk. Risk modeling is predictors classifications of future loss in motor insurance depend on history past claim, premium.



claim_no	claim_date	claim_type	est_amt
48001/CL/005576	04/09/2548 12:10:38	Wrong	5000.00
48001/CL/005577	04/09/2548 16:02:42	Right	5000.00
48001/CL/005578	04/09/2548 17:26:50	Right	2000.00
48001/CL/005580	04/09/2548 19:01:41	Right	1700.00

Risk level	Probability	Support
Safety	57.91 %	5,536
Loss > 50	23.87 %	2,282
Loss <= 50	18.22 %	1,742
Missing	0.00 %	0

Model last Update : 21/04/2007

Figure 4.9 Risk methods for assessment risk

4.2.5 Prediction Function for risk model function return prediction results all states and their probabilities. In table 4.5 show the query result using risk model. Actuaries develop risk models by large populations of policies into predictive risk groups, each with its own distinct risk characteristics. For example in table 4.5, see case 2 outcome is high risk level, risk model classify Predict risk is safety with probability is 61% and support 5,512 cases. By this way risk model function can be use to classify customer with level risk with each probabilities of level risk. Each row in the table may contain statistics, support and the probability as follow.

1. Risk level is classifying group that comprise three groups such as Safety, Medium and High.

2. Probability is the ratio of the number of favorable cases to the number of all cases.

3. Support is number of case to support each case.

Table 4.5 Query result of risk model

Case	Policy Type	Liability	Car name	Car type	Car's use	Car age (Year)	Outcome	Predict	Support	Probability (%)
1	Type2	0-300	HONDA	Passenger	Private	2	Safety	Safety	14,642	0.92
								Medium	350	0.02
								High	755	0.04
2	Type1	NA	BENZ	Passenger	Private	5-7	High	Safety	5,512	0.61
								Medium	1,917	0.21
								High	1,595	0.17
3	Type1	501-800	BENZ	Passenger	Private	10-15	Safety	Safety	4,311	0.53
								Medium	2,520	0.31
								High	1,242	0.15
4	Type1	301-500	HONDA	Passenger	Private	2	Safety	Safety	7,191	0.56
								Medium	2,545	0.19
								High	3,097	0.24
5	Type1	501-801	HONDA	Passenger	Private	3	Medium	Safety	2,235	0.50
								Medium	1,038	0.24
								High	1,135	0.26
6	Type1	301-500	ISUZU	Truck	Public	3	Safety	Safety	7,905	0.57
								Medium	3,058	0.22
								High	2,787	0.20
7	Type1	301-500	TOYOTA	Bus	Public	5-7	Safety	Safety	732	0.58
								Medium	286	0.22
								High	240	0.19

4.3 Test Plan

In Test Phase, We use unit testing to validate each unit of function. Test plan for program testing as follow:

1. Test web page – Review the content, layout, and screen design, user interface, format and input field as following :

1) Login Page

2) Motor insurance Page

3) Web page design – Report level risk motor insurance

2. Test mining package task

1) Data transformation tasks

2) Data aggregation task

3) Processing data mining

4.3.1 Test risk model embedding to client side to generating risk and display. For test risk model can be :

1. Test user use function is accessed with web browser connect to the intranet, with user must authenticate.

2. Test risk model can be making a classification risk level. The results of the prediction are displayed in the web page as risk scoring. In figure 4.10 is the underwriter application with a risk model feature integrated. This module test by input insurance data as follows:

1) Coverage policy type 1

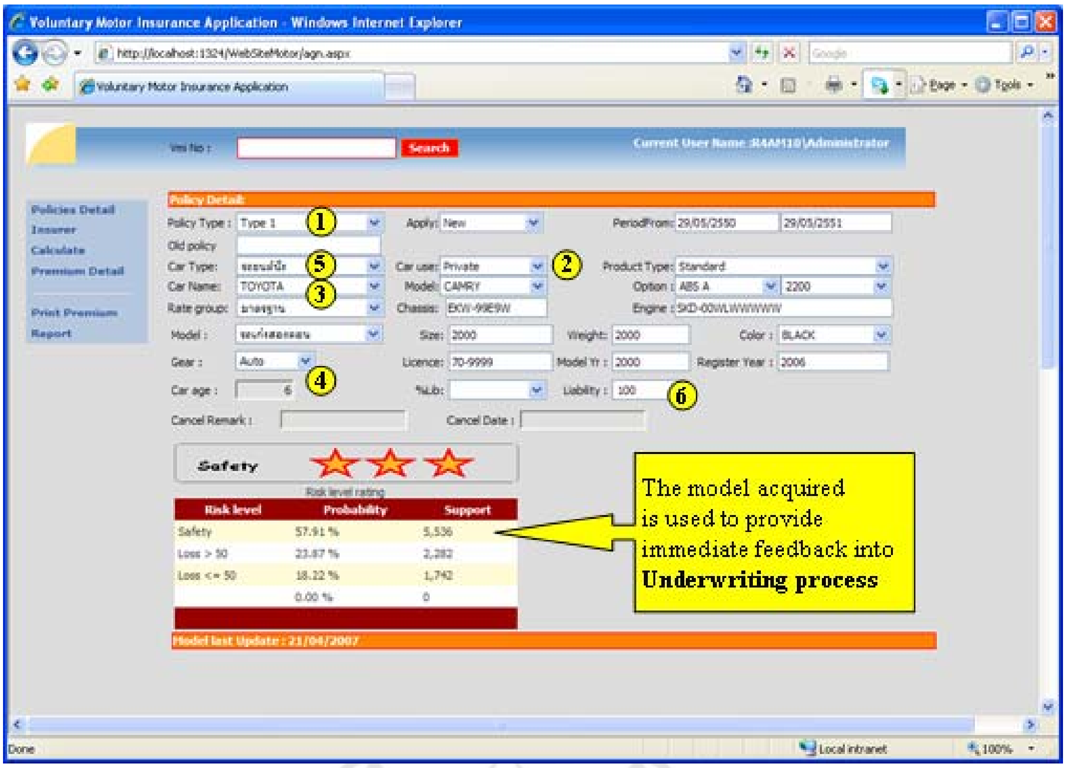
2) Car's use is private

3) Car's name is TOYOTA

4) Car's age is 6 years

5) Car code is passenger

6) Liability between 0-300,000


Voluntary Motor Insurance Application - Windows Internet Explorer
http://localhost:1324/WebSiteMotor/agn.aspx
Voluntary Motor Insurance Application
Page
Tools
Search
Policies Detail
Insurer
Calculate
Premium Detail
Print Premium
Report
1
2
3
4
5
6
Done
Local intranet
100%

Vmi No :

Current User Name :R4AM10\Administrator

Policy Detail

Policy Type : Type 1
Apply: New
PeriodFrom: 29/05/2550 29/05/2551
Old policy
Car Type: รถยนต์นั่ง
Car use: Private
Product Type: Standard
Car Name: TOYOTA
Model: CAMRY
Option : ABS A 2200
Rate group: มาตรฐาน
Chassis: EKW-99E9W
Engine : SKD-00WLWWWWW
Model : [illegible]
Size: 2000
Weight: 2000
Color : BLACK
Gear : Auto
Licence: 70-9999
Model Yr : 2000
Register Year : 2006
Car age : 6
%Lib:
Liability : 100
Cancel Remark :
Cancel Date :

Safety ☆☆☆

Risk level rating

Risk level	Probability	Support
Safety	57.91 %	5,536
Loss > 50	23.87 %	2,282
Loss <= 50	18.22 %	1,742
	0.00 %	0

Model last Update : 21/04/2007

The model acquired is used to provide immediate feedback into **Underwriting process**

Figure 4.10 Voluntary motor insurance web site

3. Test risk model can be generating risk report from risk model with statistic and support information.

4.3.2 Test mining package at server side, In figure 4.11 is Job2MiningRisk package that collect and preprocessing data for mining risk insurance. Test package at server side must have administrative permission. For test this function can be as follows:

1. Test step of data processing into data set for mining structure. It can be convert, aggregation data to appropriate data for build model as show in table 4.6.

2. Test step of data split into traing set and test set. Requiement output tables for this step is vib_training table and vib_testset table as show in table 4.7. To see the execution results of the package, run the following Transact-SQL query.

```
Select * from vib_training
Select * from vib_testset
```

3. Test step of executing mining. It can be develop initial risk model.

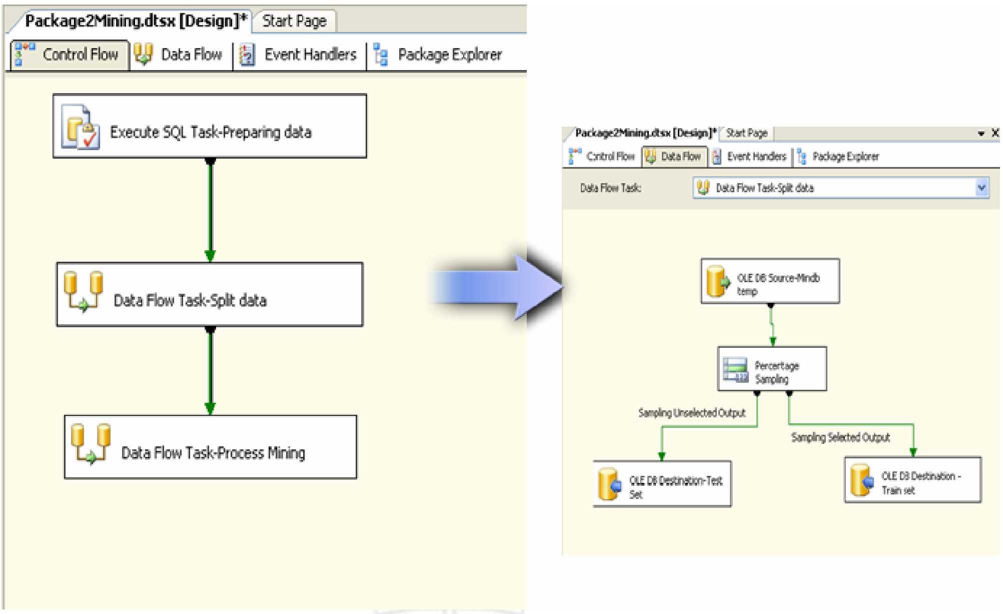

Package2Mining.dtsx [Design]*
Start Page
Control Flow
Data Flow
Event Handlers
Package Explorer
Execute SQL Task-Preparing data
Data Flow Task-Split data
Data Flow Task-Process Mining
Package2Mining.dtsx [Design]*
Start Page
Control Flow
Data Flow
Event Handlers
Package Explorer
Data Flow Task:
Data Flow Task-Split data
OLE DB Source-Mindb temp
Percentage Sampling
Sampling Unselected Output
Sampling Selected Output
OLE DB Destination-Test Set
OLE DB Destination - Train set

Figure 4.11 Package for implementing processes of data mining tasks

Table 4.6 Data set for mining structure

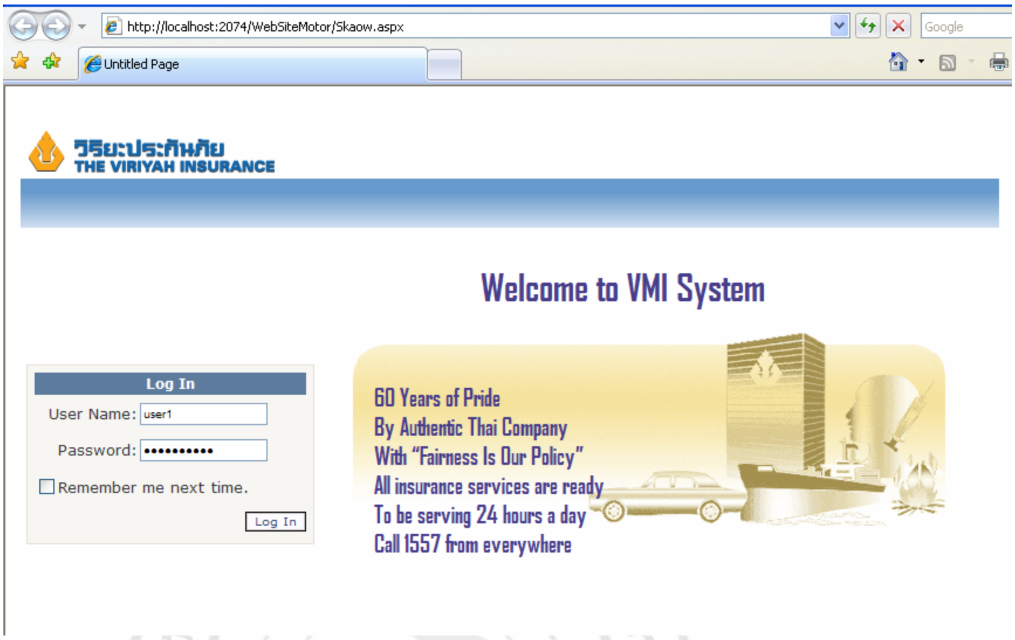
Variable name	Value range
Poltype Key	{1,2,3}
Carage Key	Eg., 2 years, 3 years.
Car Kind	E.g. Bus, Passenger
Carname Key	E.g. Masda, Volvo
Lib Key	Eg.0-300,301-500
Car Use	Private, Public
Ins Occu	Insurer's occupation
Ins Province	Place of residence
Class Code	{Safe, Medium, High}

Table 4.7 Requirement output tables for package

Table	**Description**
Vib_Trining	Contains records for which there was variable available that's used to build the data mining model.
Vib_Testset	Contains records for which there was variable available that's used to test the data mining model.

4.4 Test Result

4.4.1 Module: Test insurance risk scoring test by underwriter user. User use function is accessed with web browser connect to the intranet, with user must authenticate.



http://localhost:2074/WebSiteMotor/Skaow.aspx

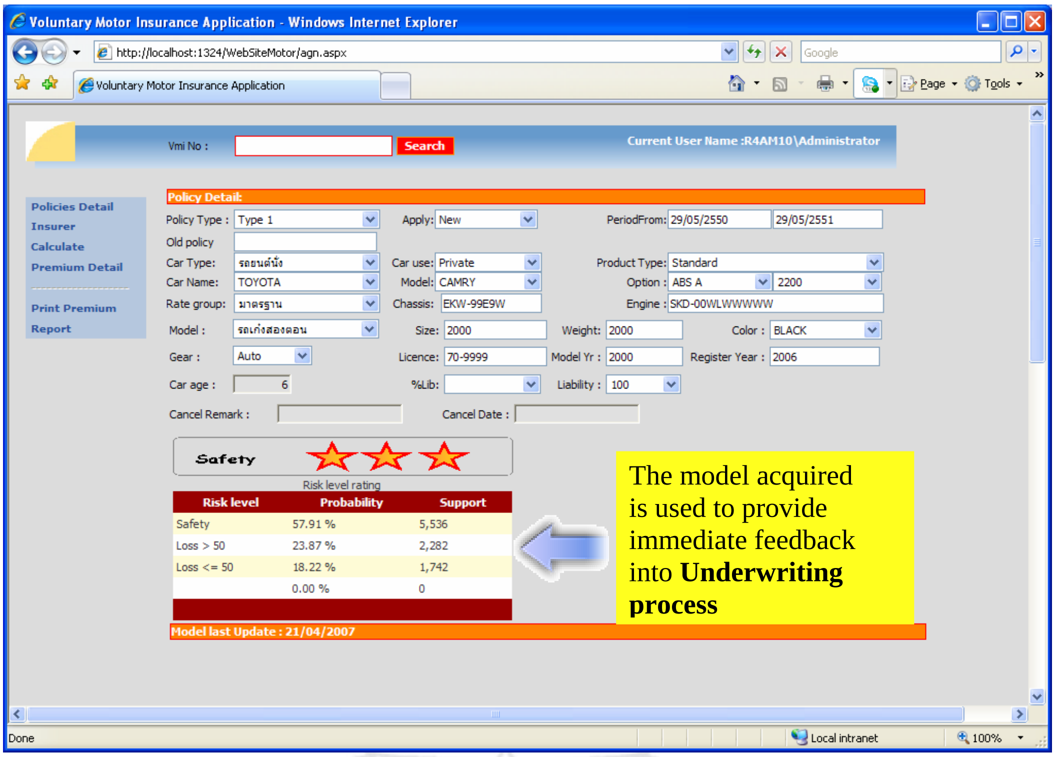
User Name: user1

Password: ••••••••••

Google
Untitled Page
วิริยะประกันภัย
THE VIRIYAH INSURANCE
Welcome to VMI System
Log In
☐ Remember me next time.
Log In
60 Years of Pride
By Authentic Thai Company
With "Fairness Is Our Policy"
All insurance services are ready
To be serving 24 hours a day
Call 1557 from everywhere

Figure 4.12 A snapshot of the underwriter system interface- user login page.

4.4.2 Test risk model with underwriter application. In figure 4.13 display result of case new customer, risk model can display risk level as table with three levels. For example 57.91% of customers who have age of car 6 years and use car as private and policy coverage all risk and liability between 0 and 300,000 is classified to safety level with medium risk is 23.87% and high risk is 18.22%. In case of customer who renewal policy we used claim history to determine risk of customer show in figure 4.14


Voluntary Motor Insurance Application - Windows Internet Explorer
http://localhost:1324/WebSiteMotor/agn.aspx
Google
Voluntary Motor Insurance Application
Page
Tools
Search
Current User Name :R4AM10\Administrator
Policies Detail
Insurer
Calculate
Premium Detail
Print Premium
Report
Done
Local intranet
100%

Vmi No :

Policy Detail:

Policy Type : Type 1 | Apply: New | PeriodFrom: 29/05/2550 | 29/05/2551

Old policy

Car Type: รถยนต์นั่ง | Car use: Private | Product Type: Standard

Car Name: TOYOTA | Model: CAMRY | Option : ABS A | 2200

Rate group: มาตรฐาน | Chassis: EKW-99E9W | Engine : SKD-00WLWWWWW

Model : รถเก๋งสองตอน | Size: 2000 | Weight: 2000 | Color : BLACK

Gear : Auto | Licence: 70-9999 | Model Yr : 2000 | Register Year : 2006

Car age : 6 | %Lib: | Liability : 100

Cancel Remark : | Cancel Date :

Safety ★ ★ ★

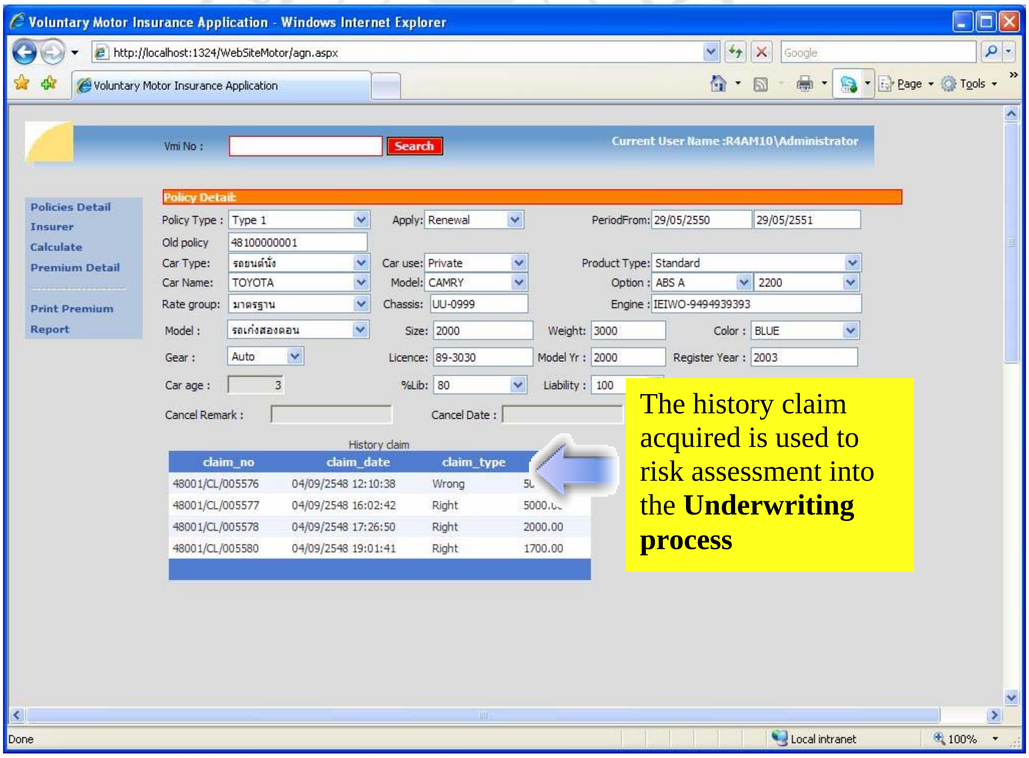
Risk level rating

Risk level	Probability	Support
Safety	57.91 %	5,536
Loss > 50	23.87 %	2,282
Loss <= 50	18.22 %	1,742
	0.00 %	0

Model last Update : 21/04/2007

The model acquired is used to provide immediate feedback into **Underwriting process**

Figure 4.13 Web page design – entry page using risk model


Voluntary Motor Insurance Application - Windows Internet Explorer
http://localhost:1324/WebSiteMotor/agn.aspx
Google
Voluntary Motor Insurance Application
Page
Tools
Search
Current User Name :R4AM10\Administrator
Policies Detail
Insurer
Calculate
Premium Detail
Print Premium
Report
Done
Local intranet
100%

Vmi No :

Policy Detail:

Policy Type : Type 1 | Apply: Renewal | PeriodFrom: 29/05/2550 | 29/05/2551

Old policy 48100000001

Car Type: รถยนต์นั่ง | Car use: Private | Product Type: Standard

Car Name: TOYOTA | Model: CAMRY | Option : ABS A | 2200

Rate group: มาตรฐาน | Chassis: UU-0999 | Engine : IEIWO-9494939393

Model : รถเก๋งสองตอน | Size: 2000 | Weight: 3000 | Color : BLUE

Gear : Auto | Licence: 89-3030 | Model Yr : 2000 | Register Year : 2003

Car age : 3 | %Lib: 80 | Liability : 100

Cancel Remark : | Cancel Date :

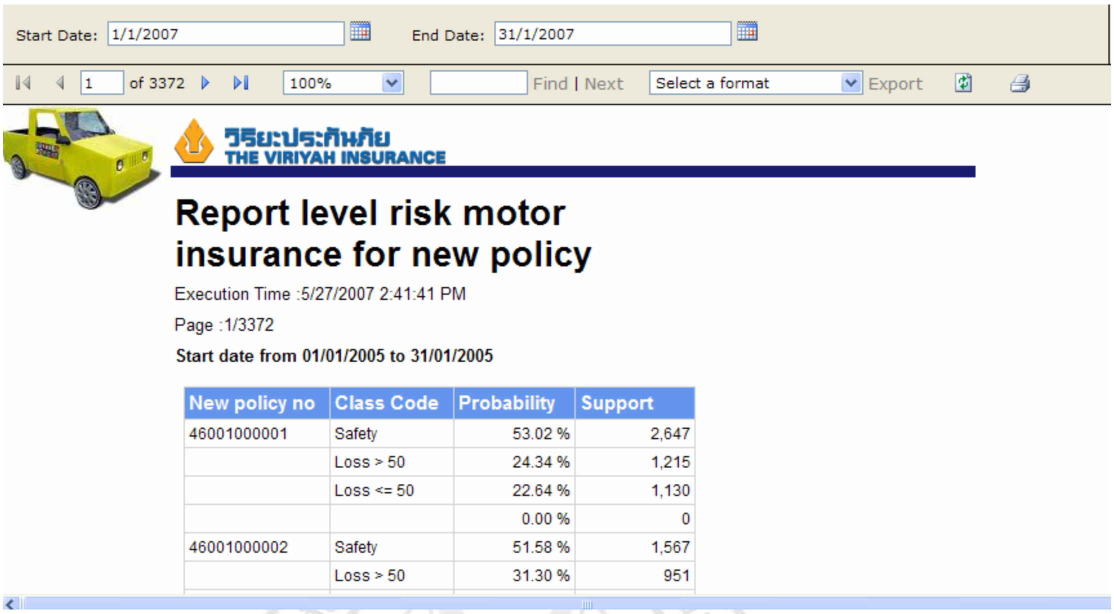
History claim

claim_no	claim_date	claim_type	
48001/CL/005576	04/09/2548 12:10:38	Wrong	5
48001/CL/005577	04/09/2548 16:02:42	Right	5000.
48001/CL/005578	04/09/2548 17:26:50	Right	2000.00
48001/CL/005580	04/09/2548 19:01:41	Right	1700.00

The history claim acquired is used to risk assessment into the **Underwriting process**

Figure 4.14 Web page design – entry page using history data

4.4.3 Test risk report, from figure 4.15 shows the result of risk mining model of new customer that generate report to predict levels of risk for class code include percent of probability and number of support each cases of new.



Start Date: 1/1/2007 End Date: 31/1/2007

1 of 3372 100% Select a format

วิริยะประกันภัย
THE VIRIYAH INSURANCE

Report level risk motor insurance for new policy

Execution Time :5/27/2007 2:41:41 PM

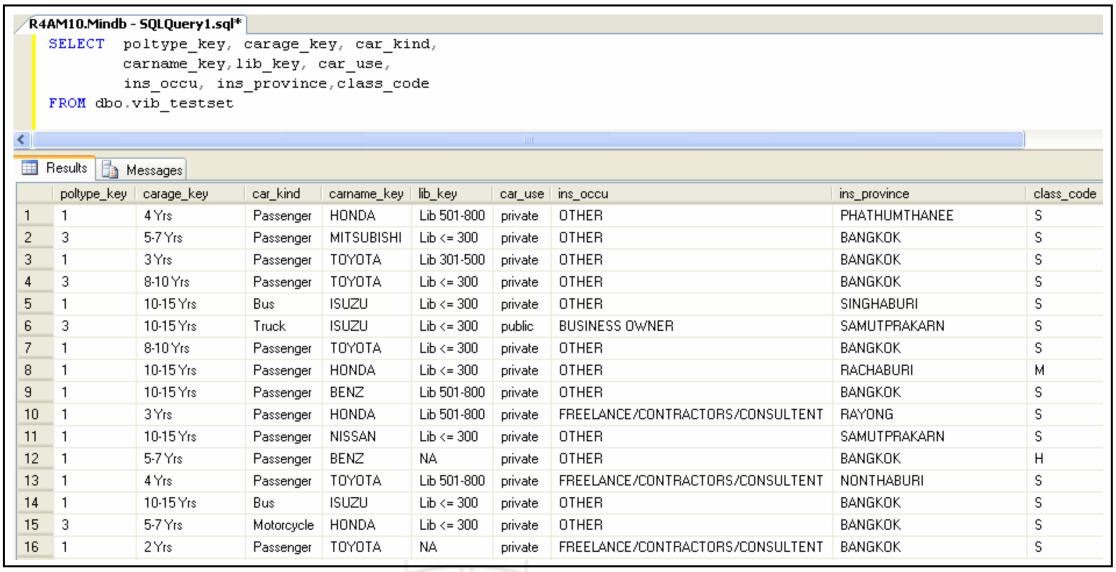
Page :1/3372

Start date from 01/01/2005 to 31/01/2005

New policy no	Class Code	Probability	Support
46001000001	Safety	53.02 %	2,647
	Loss > 50	24.34 %	1,215
	Loss <= 50	22.64 %	1,130
		0.00 %	0
46001000002	Safety	51.58 %	1,567
	Loss > 50	31.30 %	951

Figure 4.15 Report level risk motor insurance for new policy

4.4.4 Test package collect preprocessing data, divide data to training and test table for build mode and execute model by using decision tree algorithm. This module can test by administrator. The result of data preprocessing and data divide can see in figure 4.16. It shows structure of mining model correct with show in table 4.8. The vib_tesset table contains records which there were available use in data mining model. Result step of executing mining model.



R4AM10.Mindb - SQLQuery1.sql*

```
SELECT  poltype_key, carage_key, car_kind,
        carname_key,lib_key, car_use,
        ins_occu, ins_province,class_code
FROM dbo.vib_testset
```

	poltype_key	carage_key	car_kind	carname_key	lib_key	car_use	ins_occu	ins_province	class_code
1	1	4 Yrs	Passenger	HONDA	Lib 501-800	private	OTHER	PHATHUMTHANEE	S
2	3	5-7 Yrs	Passenger	MITSUBISHI	Lib <= 300	private	OTHER	BANGKOK	S
3	1	3 Yrs	Passenger	TOYOTA	Lib 301-500	private	OTHER	BANGKOK	S
4	3	8-10 Yrs	Passenger	TOYOTA	Lib <= 300	private	OTHER	BANGKOK	S
5	1	10-15 Yrs	Bus	ISUZU	Lib <= 300	private	OTHER	SINGHABURI	S
6	3	10-15 Yrs	Truck	ISUZU	Lib <= 300	public	BUSINESS OWNER	SAMUTPRAKARN	S
7	1	8-10 Yrs	Passenger	TOYOTA	Lib <= 300	private	OTHER	BANGKOK	S
8	1	10-15 Yrs	Passenger	HONDA	Lib <= 300	private	OTHER	RACHABURI	M
9	1	10-15 Yrs	Passenger	BENZ	Lib 501-800	private	OTHER	BANGKOK	S
10	1	3 Yrs	Passenger	HONDA	Lib 501-800	private	FREELANCE/CONTRACTORS/CONSULTENT	RAYONG	S
11	1	10-15 Yrs	Passenger	NISSAN	Lib <= 300	private	OTHER	SAMUTPRAKARN	S
12	1	5-7 Yrs	Passenger	BENZ	NA	private	OTHER	BANGKOK	H
13	1	4 Yrs	Passenger	TOYOTA	Lib 501-800	private	FREELANCE/CONTRACTORS/CONSULTENT	NONTHABURI	S
14	1	10-15 Yrs	Bus	ISUZU	Lib <= 300	private	OTHER	BANGKOK	S
15	3	5-7 Yrs	Motorcycle	HONDA	Lib <= 300	private	OTHER	BANGKOK	S
16	1	2 Yrs	Passenger	TOYOTA	NA	private	FREELANCE/CONTRACTORS/CONSULTENT	BANGKOK	S

Figure 4.16 Result of preprocess and split data with package

In table 4.8 display test result of mining package and risk model within underwriter application.

Table 4.8 Test result for insurance risk modeling

TEST REPORT	
Project : Risk modeling for voluntary motor insurance with decision tree algorithm	
Test author: Suparut Tubnakog	
Test date: 15-31 May 2007	
Task Name	**Results**
Module: Test a package **Test by: Administrator**	
1. Load data from data source	Pass
2. Check mapping column correctly	Pass
3. Check data to divide to training set and test set	Pass
4. Check transform data correctly	Pass
5. Process data mining model	Pass
Module: Test insurance risk scoring **Test by: Underwriter user**	
Test risk model function for underwriter application can predict new data	Pass
Test report shows the result of risk mining model of new customer	Pass

CHAPTER V

SUMMARY AND SUGGESTIONS

5.1 Introduction

Today’s business environment is more competitive than ever. This advantage is often the result of better information technology providing the basis for improved business decisions. The problem of how to make such business decisions is therefore crucial. One answer is through the better analysis of data. Data mining is a methodology to assess the value of the data and to leverage that value as an asset to provide large returns on the analytic investment.

The problem that often confronts researchers new to the field is that there are a variety of data mining techniques available—which one to choose? All these tools give you answers. Some are more difficult to use than others, and they differ in other, superficial ways, but most importantly, the underlying algorithms used differ and the nature of these algorithms is directly related to the quality of the results obtained and ease of use.

Data mining technology is used increasingly by many companies to analyze large databases in order to discover previously unknown and actionable information. It’s not a perfect result depend on data, to gain small advantage. It offers the opportunity to apply technology to improve process of business.

Many traditional reporting and query tools and statistical analysis systems use the term "data mining" in their product descriptions.

Decision tree methods are techniques for partitioning a training file into a tree representation. The starting node is called the root node. Depending upon the results of a test this node is then partitioned into two or more sub-sets. Each node is then further partitioned until a tree is built. This tree can be mapped into a set of rules.

5.2 Project Summary

This project aims to build a risk models based on the voluntary motor insurance historical data that will to maximize the accuracy of their risk assessments for avoid high risk customer. The insurance company wants to identify level risk insure that might indicate whether those customer are likely to be claimed in the future. The problem of insurance need to know "What risk level of new customer". The insurance database stores policy and claim data that describes customers. By using the Microsoft Decision Tree algorithm to analyze this information, underwriter can use model that predicts whether a particular customer will be claimed, based on the historical data about the insurance data, such as car detail and claim history. In this study we have followed the CRISP-DM process and highlighted its application to the domain of motor vehicle insurance risk assessment that model will solve, to deploying the model into working environment.

In this report, we present a model that classifying level risk insurance from insurance data warehouse using ID3 techniques. This classification method attempts to predict claim probability or claim amounts for new insurance applications based on historical insurance data. Data may hold hidden patterns that can be discovered, the historical information is also useful to business. In data set we must cleansing and transform data into appropriate format for use with ID3 method.

Method for measuring a tree split, consisting of Entropy and Bayesian K2, performed the best for select attributes. Based on calculated accuracy of two methods is closely prediction. The entropy have smaller tree. Thus entropy can searching faster than Bayesian with K2 based on data set.

In evaluate phase, to ensure performance model to assess the model performance against a data set that has the same characteristics, the model should be executed against the test data in test mode. Accuracy of risk model with Entropy method is 65.12%, generally data set of insurance claims are very low events. Decision trees develop predictive models by use frequent value. The risk model is possible predict risky customer in to safety level. As the result of risk model have low accuracy in predicting risky customer to risk level. Apply risk model to use can be done for only suggestion risk information and generate possible risk value for each segment. For example 92% of Customer who have car's age are 2 years and policy type1 and type of car is passenger and car use for private car is classified to safety level, with 0.02% of this segment is classified to Medium level and rest is High level risk.

Deploy phase, after the mining models exist in a production environment, we use the models to create predictions for classify level risk, which underwriter can then use to make insurance business decisions. DMX language can use to create prediction queries, and Prediction Query Builder to help you build the queries. Embed data mining functionality directly into an underwriter application. Create a report that lets users directly query against an insurance risk mining model. Updating the model is part of the deployment strategy; administrator must reprocess the models, there by improving their effectiveness.

For this project, the result from insurance risk model used for identify level risk insure that might indicate whether those customer are likely to be claimed in the future. Underwriter can make a decision based statistical probability of their insurance data.

5.3 Problems Encountered and Solutions

The most important step in developing a model is to clear objective and develop a process to achieve that goal. Understanding the business applications of data mining is necessary to be exposed. Data mining requires understanding of data and business problem. This problem can be efficiently solved by subject-area experts joining the team.

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems. A large quantity of database and information consists of many variables, involving both categorical and numerical data. The problem of data as follows:

5.3.1 Missing value and incomplete data. In case of missing value and incomplete data may come from no available data valued when collected or problem with human, software. Manually enter the missing value can be done in case missing less. In case missing over 50% we were excluded in the variable selection process.

5.3.2 Data noisy that containing error or outliers. Noisy data comes from the process of data such as collection and entry. Few noisy data of their values. Calculate or manually assign correct value use to fix this problem.

5.3.3 Data Inconsistent may come from different data sources. Little of inconsist manually assign correct value can be use.

In a typical data mining project, the most resource-consuming step in data preparation. Creating and tuning mining models may represent only 20 percent of the total project effort. However, before creating these models, your data needs to be in the right format. Data preparation consists of multiple step, including data gathering, cleaning, and transformation. Data should be converted into appropriate formats. To solve the problem, we use extract, transform and load (ETL) software, which includes reading data from its source, cleaning it up and formatting it, and then load to target data source. It helps reducing the time for generating data from data warehouses to target data source for mining model.

5.4 Suggestions for Further Development

One of key to success of model is the appropriate selection of data for the development and validation of a model. Validation is an important step in the data mining process. Knowing how well your mining models perform against real data is important before you deploy the models into a production environment. Also, you may have created several models and will have to decide which model will perform the best. There several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data.

In this work, we specify decision tree algorithm method only to build risk model. Better approaches are to predict risk level of customer using another algorithm to build model. We expect the performance to be further improved by considering multiple techniques and data set. In addition, the decision tree can be more robust by including a complete set of the available concepts in the database, and by learning through a larger sample data set.

Further build insurance risk by alternatively, based on the type of claim can be further divided, “wrong”, “right”. With this information as the target of prediction, classification techniques can predict insurance risks level of new applications and find relationships in their data.

REFERENCES

Apte, C. and Weiss, S. Data mining with decision trees and decision rules. In: Future Generation Computer Systems. [Amsterdam, The Nethelands]: Noerth - Holland; 1997. p.13.

Chapman, P., et al. "CRISP-DM 1.0 : Step by Step Data Mining Guide," CRISPDM Consortium [online]. 2000 [cited 2007] Available from : http://www.crisp-dm.org

Chickering, D. and Heckerman, D. A Decision Theoretic Approach toTargeted Advertising. n.l.: UAI; 2000.

Cooper, G.F., and Herskovits, E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning. 1992; 9(4) : 309-347.

CRISP-DM Group. Cross-industry Standard Process of Data mining [online] 1996 [cited 2007] Available from: http://www.crisp-dm.org/

Daskalaki, S. Kopanas, I. Goudara,M. and Avouris, N. Data mining for Decision Support on Customer Insolvency in the Telecommunications Business. European Journal of Operational Research. 2003; 145: 239-255.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. Knowledge Discovery and Data Mining. Dordrecht ,Boston : Kluwer Academic Publishers; 1996.

Hoffman, T. Finding a Rich Niche. Computer World. 1999; 33(6): 44.

Kohavi, R. and Provost,F. On applied research in machine learning. Machine learning. 1998; 30: 127-132.

Koonce, D. A., Fang, C. H. and Tsai, S.C. A Data Mining Tool for Learning from Manufacturing Systems. Computers & Industrial Engineering. 1997; 33(1-2): 27-30.

Kubat, M., Holte, R.C., and Matwin, S. Machine learning for the detection of oil spills in satellite radar images. Machine Learn. 1998; 30: 195-215.

Levinsohn, A. Modern Miners Plumb for Gold. ABA Banking Journal. 1998; 90(12) : 52-55.

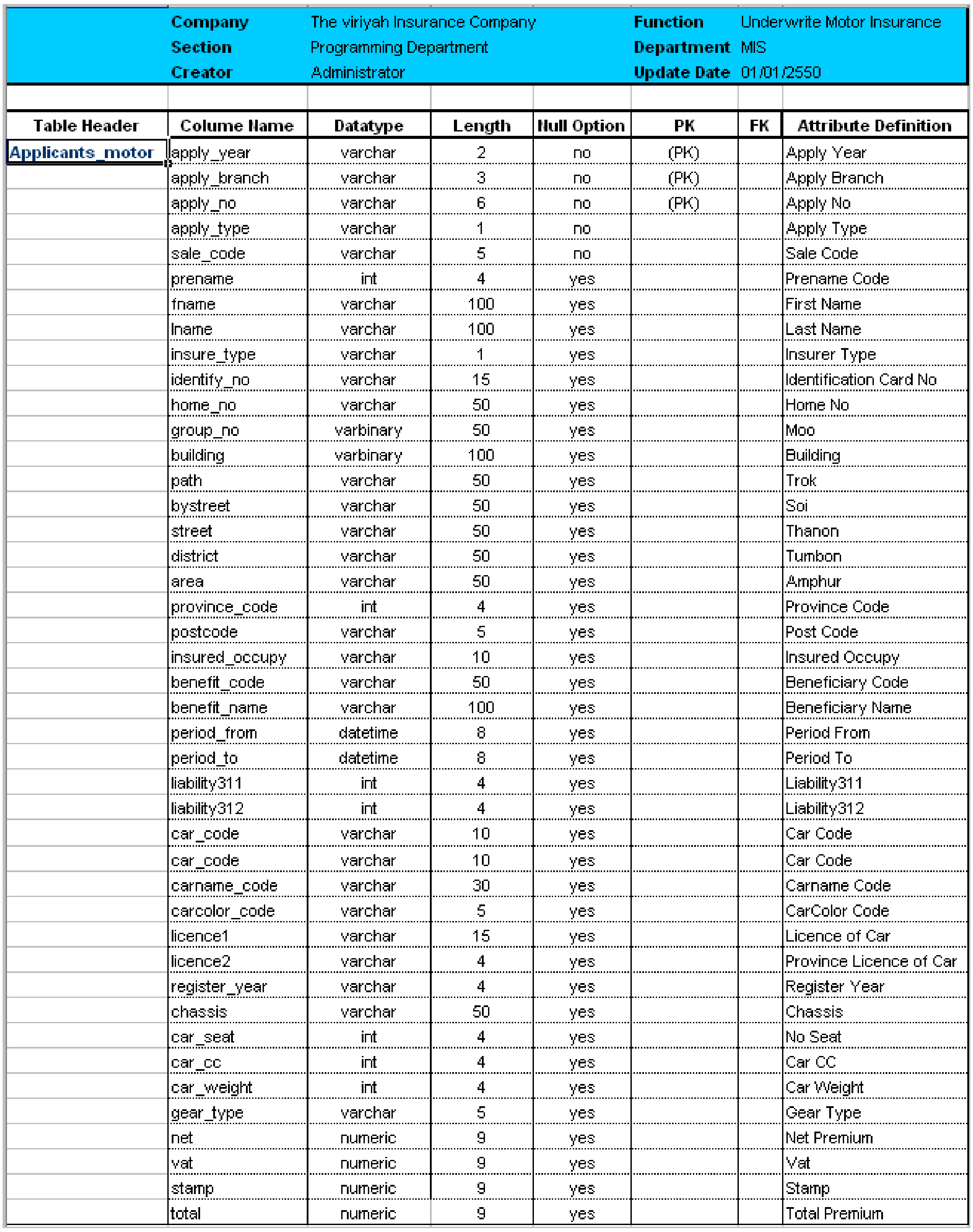
Ling, Chares X. and Li,Chenghui. Data Mining for Direct Marketing : Problems and Solutions. [Thesis] The University of Western Ontario. KDD; 1998.

Quinlan, J. R. Induction of decision trees. Machine Learning. 1986; 1: 81-106.

APPENDIX A

DATA DICTIONARY

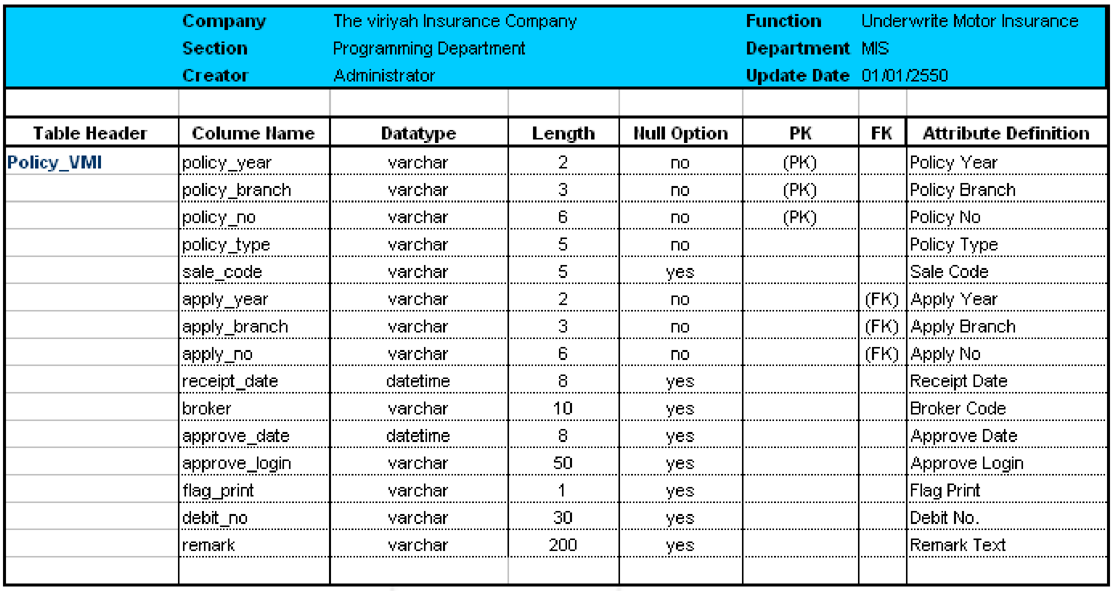
Data Dictionary



	Company	The viriyah Insurance Company	**Function**	Underwrite Motor Insurance
	Section	Programming Department	**Department**	MIS
	Creator	Administrator	**Update Date**	01/01/2550

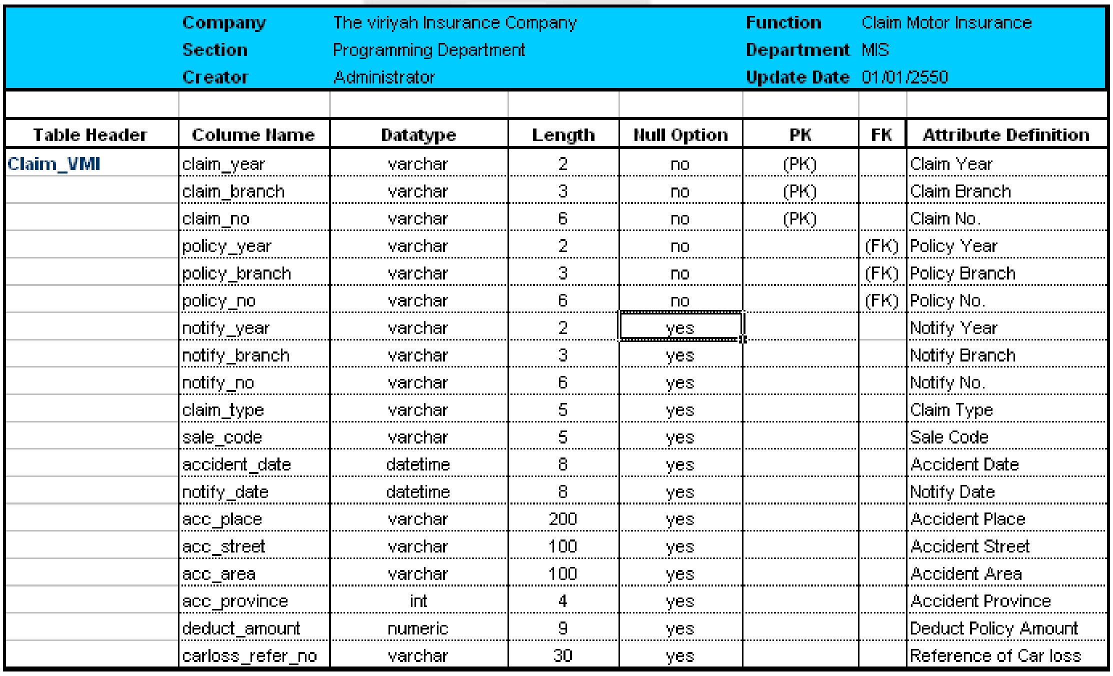
Table Header	Colume Name	Datatype	Length	Null Option	PK	FK	Attribute Definition
Applicants_motor	apply_year	varchar	2	no	(PK)		Apply Year
	apply_branch	varchar	3	no	(PK)		Apply Branch
	apply_no	varchar	6	no	(PK)		Apply No
	apply_type	varchar	1	no			Apply Type
	sale_code	varchar	5	no			Sale Code
	prename	int	4	yes			Prename Code
	fname	varchar	100	yes			First Name
	lname	varchar	100	yes			Last Name
	insure_type	varchar	1	yes			Insurer Type
	identify_no	varchar	15	yes			Identification Card No
	home_no	varchar	50	yes			Home No
	group_no	varbinary	50	yes			Moo
	building	varbinary	100	yes			Building
	path	varchar	50	yes			Trok
	bystreet	varchar	50	yes			Soi
	street	varchar	50	yes			Thanon
	district	varchar	50	yes			Tumbon
	area	varchar	50	yes			Amphur
	province_code	int	4	yes			Province Code
	postcode	varchar	5	yes			Post Code
	insured_occupy	varchar	10	yes			Insured Occupy
	benefit_code	varchar	50	yes			Beneficiary Code
	benefit_name	varchar	100	yes			Beneficiary Name
	period_from	datetime	8	yes			Period From
	period_to	datetime	8	yes			Period To
	liability311	int	4	yes			Liability311
	liability312	int	4	yes			Liability312
	car_code	varchar	10	yes			Car Code
	car_code	varchar	10	yes			Car Code
	carname_code	varchar	30	yes			Carname Code
	carcolor_code	varchar	5	yes			CarColor Code
	licence1	varchar	15	yes			Licence of Car
	licence2	varchar	4	yes			Province Licence of Car
	register_year	varchar	4	yes			Register Year
	chassis	varchar	50	yes			Chassis
	car_seat	int	4	yes			No Seat
	car_cc	int	4	yes			Car CC
	car_weight	int	4	yes			Car Weight
	gear_type	varchar	5	yes			Gear Type
	net	numeric	9	yes			Net Premium
	vat	numeric	9	yes			Vat
	stamp	numeric	9	yes			Stamp
	total	numeric	9	yes			Total Premium

Data Dictionary



Company	The viriyah Insurance Company	Function	Underwrite Motor Insurance
Section	Programming Department	Department	MIS
Creator	Administrator	Update Date	01/01/2550

Table Header	Colume Name	Datatype	Length	Null Option	PK	FK	Attribute Definition
Policy_VMI	policy_year	varchar	2	no	(PK)		Policy Year
	policy_branch	varchar	3	no	(PK)		Policy Branch
	policy_no	varchar	6	no	(PK)		Policy No
	policy_type	varchar	5	no			Policy Type
	sale_code	varchar	5	yes			Sale Code
	apply_year	varchar	2	no		(FK)	Apply Year
	apply_branch	varchar	3	no		(FK)	Apply Branch
	apply_no	varchar	6	no		(FK)	Apply No
	receipt_date	datetime	8	yes			Receipt Date
	broker	varchar	10	yes			Broker Code
	approve_date	datetime	8	yes			Approve Date
	approve_login	varchar	50	yes			Approve Login
	flag_print	varchar	1	yes			Flag Print
	debit_no	varchar	30	yes			Debit No.
	remark	varchar	200	yes			Remark Text



Company	The viriyah Insurance Company	Function	Claim Motor Insurance
Section	Programming Department	Department	MIS
Creator	Administrator	Update Date	01/01/2550

Table Header	Colume Name	Datatype	Length	Null Option	PK	FK	Attribute Definition
Claim_VMI	claim_year	varchar	2	no	(PK)		Claim Year
	claim_branch	varchar	3	no	(PK)		Claim Branch
	claim_no	varchar	6	no	(PK)		Claim No.
	policy_year	varchar	2	no		(FK)	Policy Year
	policy_branch	varchar	3	no		(FK)	Policy Branch
	policy_no	varchar	6	no		(FK)	Policy No.
	notify_year	varchar	2	yes			Notify Year
	notify_branch	varchar	3	yes			Notify Branch
	notify_no	varchar	6	yes			Notify No.
	claim_type	varchar	5	yes			Claim Type
	sale_code	varchar	5	yes			Sale Code
	accident_date	datetime	8	yes			Accident Date
	notify_date	datetime	8	yes			Notify Date
	acc_place	varchar	200	yes			Accident Place
	acc_street	varchar	100	yes			Accident Street
	acc_area	varchar	100	yes			Accident Area
	acc_province	int	4	yes			Accident Province
	deduct_amount	numeric	9	yes			Deduct Policy Amount
	carloss_refer_no	varchar	30	yes			Reference of Car loss

Description of insurance data set

Variables	Represent	Value range
Pol No	Policy No	Number of policies
Poltype Key	Policy coverage	Policy type 1 Policy type2 Policy type 3
Carage Key	Age of the car	New license 2 Years 3 Years 4 Years 5 - 7 Years 8 - 10 Years 10 - 15 Years 16 - 20 Years > 20 Years NA
Carcode Key	Car use	Passenger Public
Carname Key	Car's name	BENZ BMW FORD HONDA ISUZU MITSUBISHI NISSAN TOYOTA OTHER
Deductible	Deductibles	Yes No
Discount	Premium discount	Yes No
Extra Accessories	Car's special decorate	Yes No
Flag Define Drv	Policy define driver	Yes No
Lib Key	Liability insurance	0 - 300,000 300,001 - 500,000 500,001 - 800,000 800,001 - 1,000,000 > 1,000,000

Description of insurance data (Continue)

Variables	Represent	Value range
Car Gear	Car's Gear	Auto Manual
Car Kind	Car Kind	Public bus Caravan Bus Trailer car Passenger Truck Motorcycle Miscellaneous
Car Use	Car Use	Public Private
Ins Occu	Insurer's occupation	SENIOR EXECUTIVES MIDDLE MANAGEMENT CIVIL SERVANT/STATE ENTERPRISE STAFF POLICE/MILITARY POLITICIAN DOCTORS/PHYSICIAN NURSE/OTHER MEDICAL EMGINEER/TECHNICIAN ARCHITECT/DESIFNER/ARTISTS ACTORS/SINGERS/ENTERTAINMENT OUTDOOR SATES/MARKETING PERSON BUSINESS OWNER SELF-EMPLOYED TRADES VEHICLE FREELANCE/CONTRACTORS/CONSULTANT CHAUFFEUR/DRIVER FARMERS HOUSEWIFE OUTDOOR JOURNALIST/MEDIA PHOTOGRAPHE TEACHER AND ALL EDUCATORS ACCOUNTANT/CASHIER ADMIN/SECRETARY/CLERK/OFFICE UNEMPLOYED OTHER
Ins Province	Place of residence	74 Province in Thailand
Class Code	Level risk rang of losses	Safety Loss 1-50 Loss > 50

APPENDIX B

INSTALLING SQL SERVER 2005

Installing SQL Server 2005

The SQL Server 2005 Installation Wizard is Microsoft Windows Installer-based, and provides a single feature tree for installation of all Microsoft SQL Server 2005 components :

1. SQL Server Database Engine
2. Analysis Services
3. Reporting Services
4. Notification Services
5. Integration Services
6. Management Tools
7. Documentation and Samples

Prerequisites The Setup process involves the following steps.

Step 1: Prepare Your Computer to Install SQL Server 2005

Hardware Requirements

Monitor

SQL Server graphical tools require VGA or higher resolution: at least 1,024x768 pixel resolution.

(a) Pointing Device

A compatible pointing device is required.

(b) CD or DVD Drive

A CD or DVD drive, as appropriate, is required for installation from CD or DVD media.

(c) Network Software Requirements

Windows 2003, Windows XP, and Windows 2000 have built-in network software.

Hardware and Software Requirements

This table shows hardware requirements for installing and running SQL Server 2005.

SQL Server 2005	Processor type	Processor speed	Memory (RAM)
SQL Server 2005 Enterprise Edition [4] SQL Server 2005 Developer Edition SQL Server 2005 Standard Edition	Pentium III-compatible processor or higher required	Minimum: 600 MHz Recommended: 1 GHz or higher	Minimum: 512 MB Recommended: 1 GB or more Maximum: OS maximum
SQL Server 2005 Workgroup Edition	Pentium III-compatible processor or higher required	Minimum: 600 MHz Recommended: 1 GHz or higher	Minimum: 512 MB Recommended: 1 GB or more Maximum: OS maximum
SQL Server 2005 Express Edition	Pentium III-compatible processor or higher required	Minimum: 600 MHz Recommended: 1 GHz or higher	Minimum: 192 MB Recommended: 512 MB or more Maximum: OS maximum

Hard Disk Space Requirements

Actual hard disk space requirements depend on your system configuration and the applications and features you choose to install. The following table provides disk space requirements for SQL Server 2005 components.

Feature	Disk space requirement
Database Engine and data files, Replication, and Full-Text Search	150 MB
Analysis Services and data files	35 KB
Reporting Services and Report Manager	40 MB
Notification Services engine components, client components, and rules components	5 MB
Integration Services	9 MB
Client Components	12 MB
Management Tools	70 MB
Development Tools	20 MB
SQL Server Books Online and SQL Server Mobile Books Online	15 MB
Samples and sample databases	390 MB

Software Requirements

SQL Server Setup installs the following software components required by the product:

1. Microsoft Windows .NET Framework 2.0
2. Microsoft SQL Server Native Client
3. Microsoft SQL Server Setup support files

Security Considerations for a SQL Server Installation

Enhance Physical Security

Physical and logical isolation make up the foundation of SQL Server security. To enhance the physical security of the SQL Server installation, do the following tasks:

1. Place the server in a room inaccessible to unauthorized persons.
2. Place computers that host a database in a physically protected location, ideally a locked computer room with monitored flood detection and fire detection or suppression systems.
3. Install databases in the secure zone of the corporate intranet and never directly connected to the Internet.
4. Back up all data regularly and store copies in a secure off-site location.

(d) Use Firewalls

Firewalls are integral to securing the SQL Server installation. Firewalls will be most effective if you follow these guidelines:

1. Put a firewall between the server and the Internet.
2. Divide the network into security zones separated by firewalls. Block all traffic, and then selectively admit only what is required.
3. Always block packets addressed to TCP port 1433 (monitored by the default instance) and UDP port 1434 (monitored by one of the instances on the computer) on your perimeter firewall. If named instances are listening on additional ports, block those too.
4. In a multitier environment, use multiple firewalls to create screened subnets.
5. When you are installing the server inside a Windows domain, configure interior firewalls to permit Windows Authentication.
6. Open ports used by Kerberos or NTLM authentication.

Service accounts

1. Run SQL Server services with the lowest possible privileges.
2. Associate SQL Server services with Windows accounts.

Authentication mode

1. Require Windows Authentication for connections to SQL Server.

Strong passwords

1. Always assign a strong password to the sa account, even when using Windows Authentication.
2. Always use strong passwords for all SQL Server accounts.

(e) Internet Requirements

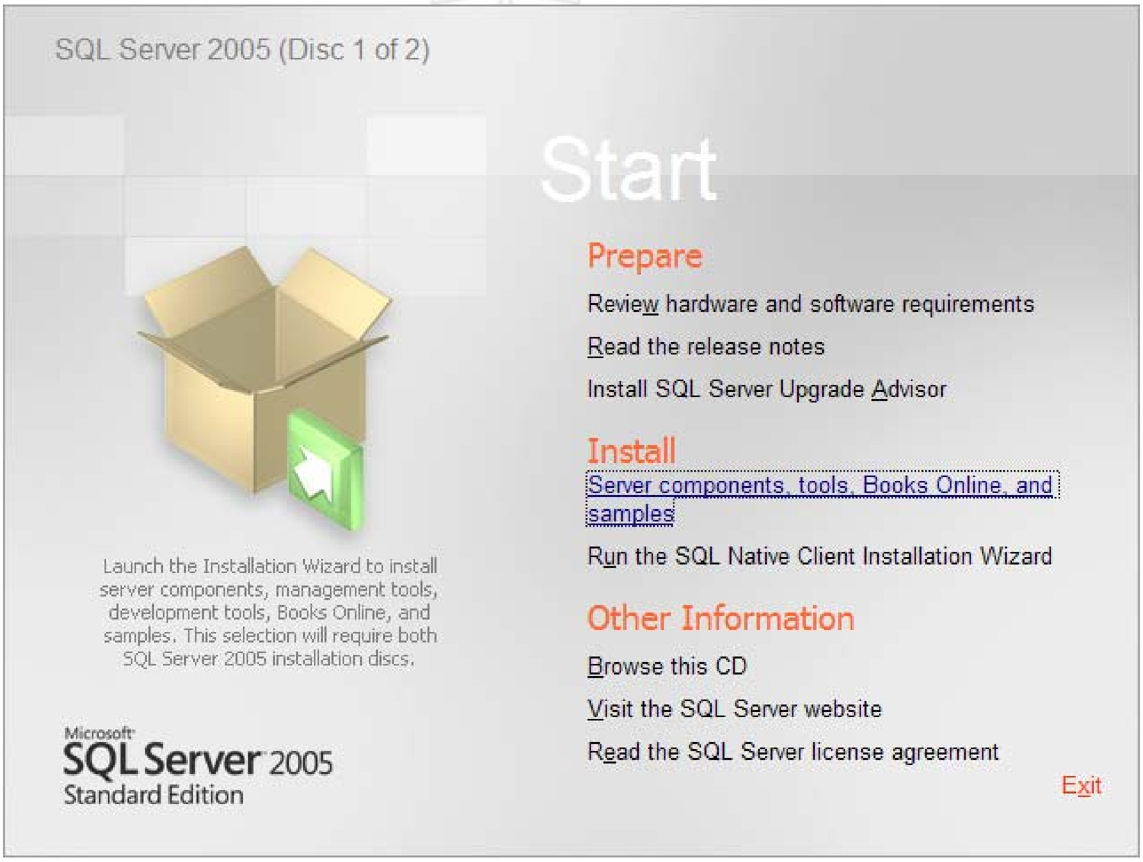
Component	**Requirement**
Internet software	Microsoft Internet Explorer 6.0 SP1 or later is required for all installations of SQL Server 2005, as it is required for Microsoft Management Console (MMC) and HTML Help. A minimal installation of Internet Explorer is sufficient, and Internet Explorer is not required to be the default browser. However, if you are installing client components only and you will not connect to a server that requires encryption, Internet Explorer 4.01 with Service Pack 2 is sufficient.
Internet Information Services (IIS)	IIS 5.0 or later is required for Microsoft SQL Server 2005 Reporting Services (SSRS) installations. For more information on how to install IIS, see How to: Install Microsoft Internet Information Services.
ASP.NET 2.0	ASP.NET 2.0 is required for Reporting Services. When installing Reporting Services, SQL Server Setup will enable ASP.NET if it is not already enabled.

Step 2: To Install SQL Server 2005

To install SQL Server 2005, run Setup using the SQL Server 2005 Installation Wizard or install from the command prompt. You can also add components to an instance of SQL Server 2005, or upgrade to SQL Server 2005 from a previous SQL Server version.

1. **To begin the installation process, insert the SQL Server 2005 DVD into your DVD drive.**

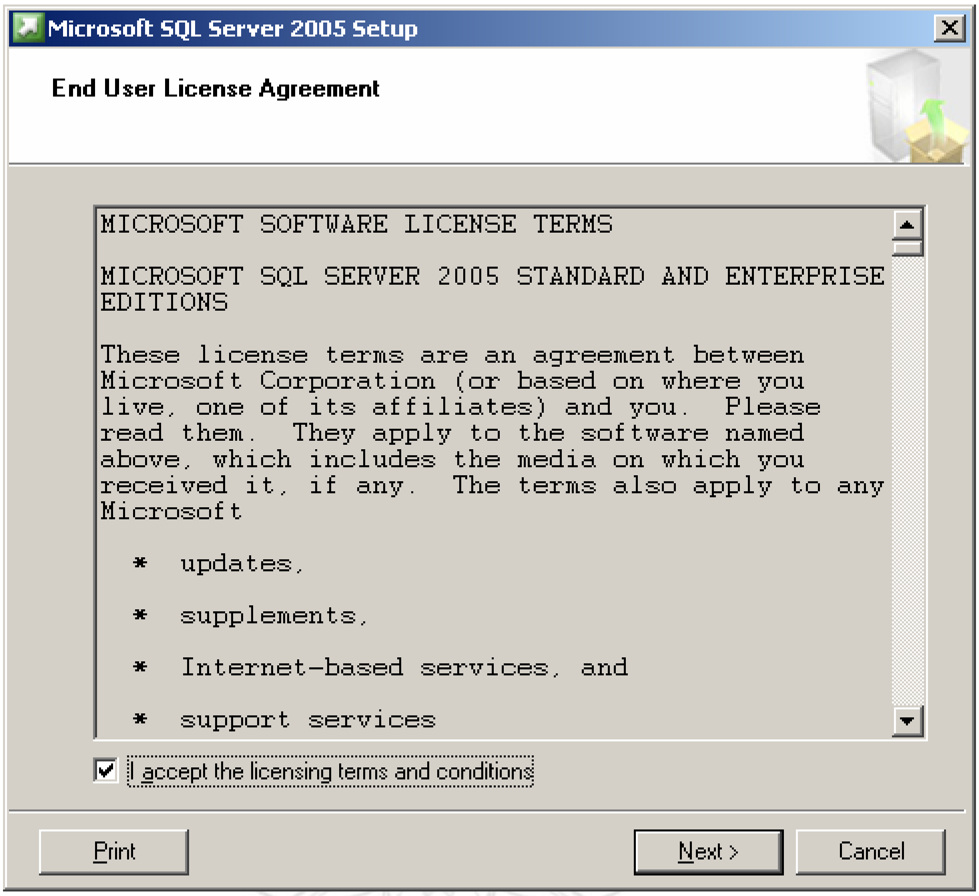
   If the autorun feature on your DVD drive does not launch the installation program, navigate to the root of the DVD.


SQL Server 2005 (Disc 1 of 2)
Start
Prepare
Review hardware and software requirements
Read the release notes
Install SQL Server Upgrade Advisor
Install
Server components, tools, Books Online, and samples
Run the SQL Native Client Installation Wizard
Other Information
Browse this CD
Visit the SQL Server website
Read the SQL Server license agreement
Exit
Microsoft SQL Server 2005
Standard Edition

Launch the Installation Wizard to install server components, management tools, development tools, Books Online, and samples. This selection will require both SQL Server 2005 installation discs.

2. **From the autorun dialog, click Install the SQL Server Installation Wizard.**

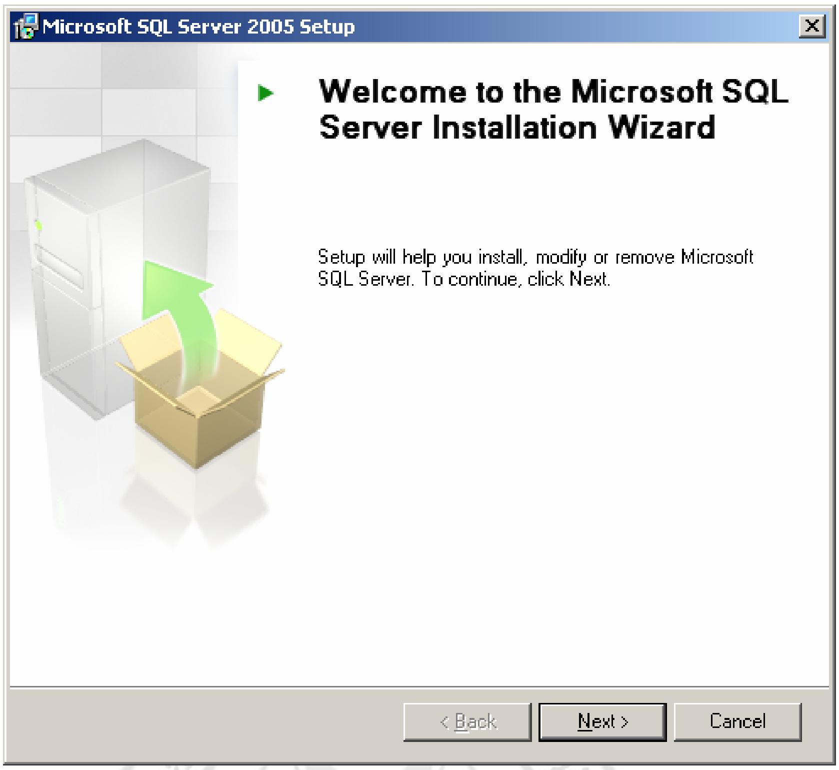
3. **On the End User License Agreement page, read the license agreement, and then select the check box to accept the licensing terms and conditions. Accepting the license agreement activates the Next button. To continue, click Next. To end Setup, click Cancel.**



```
MICROSOFT SOFTWARE LICENSE TERMS

MICROSOFT SQL SERVER 2005 STANDARD AND ENTERPRISE
EDITIONS

These license terms are an agreement between
Microsoft Corporation (or based on where you
live, one of its affiliates) and you.  Please
read them.  They apply to the software named
above, which includes the media on which you
received it, if any.  The terms also apply to any
Microsoft

  *  updates,

  *  supplements,

  *  Internet-based services, and

  *  support services
```

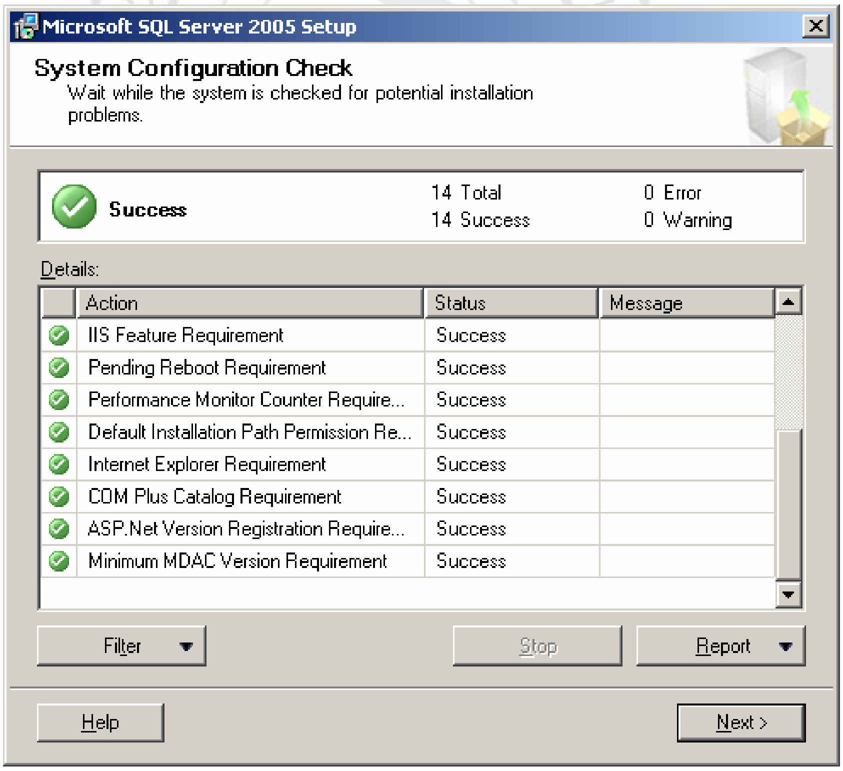
☑ I accept the licensing terms and conditions

4. On the Welcome page of the SQL Server Installation Wizard, click Next to continue.



Welcome to the Microsoft SQL Server Installation Wizard

Setup will help you install, modify or remove Microsoft SQL Server. To continue, click Next.



System Configuration Check

Wait while the system is checked for potential installation problems.

Success

14 Total

14 Success

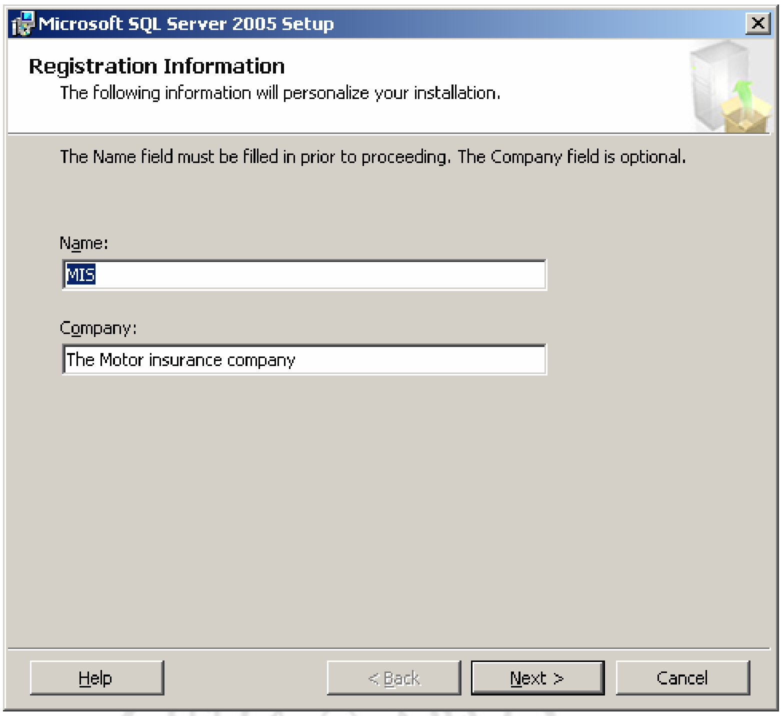
0 Error

0 Warning

Details:

	Action	Status	Message
✓	IIS Feature Requirement	Success	
✓	Pending Reboot Requirement	Success	
✓	Performance Monitor Counter Require...	Success	
✓	Default Installation Path Permission Re...	Success	
✓	Internet Explorer Requirement	Success	
✓	COM Plus Catalog Requirement	Success	
✓	ASP.Net Version Registration Require...	Success	
✓	Minimum MDAC Version Requirement	Success	

5. **On the Registration Information page, enter information in the Name and Company text boxes. To continue, click Next.**



Microsoft SQL Server 2005 Setup

Registration Information

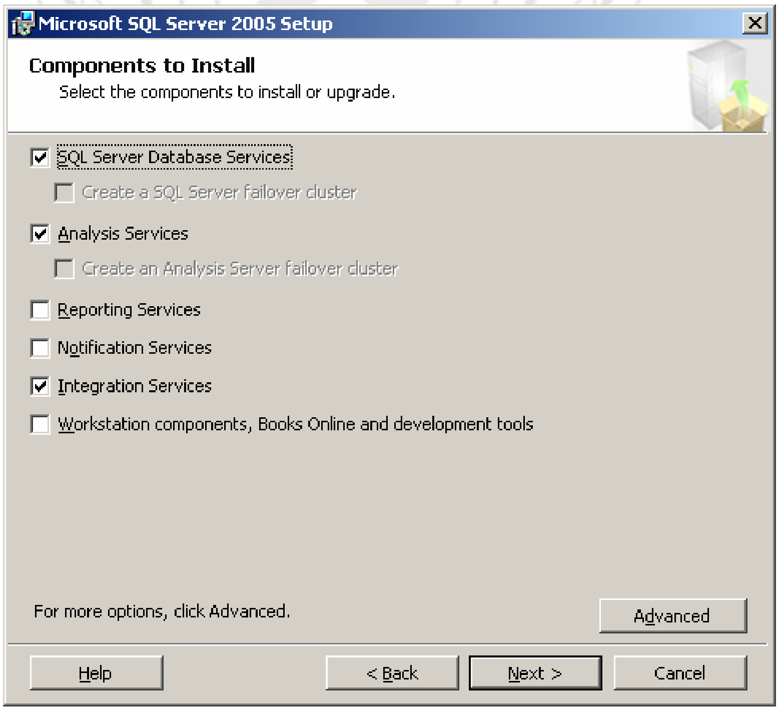
The following information will personalize your installation.

The Name field must be filled in prior to proceeding. The Company field is optional.

Name: MIS

Company: The Motor insurance company

6. **On the Components to Install page, select the components for your installation. A description for each component group appears in the Components to be Installed pane when you select it.**



Microsoft SQL Server 2005 Setup

Components to Install

Select the components to install or upgrade.

☑ SQL Server Database Services

□ Create a SQL Server failover cluster

☑ Analysis Services

□ Create an Analysis Server failover cluster

□ Reporting Services

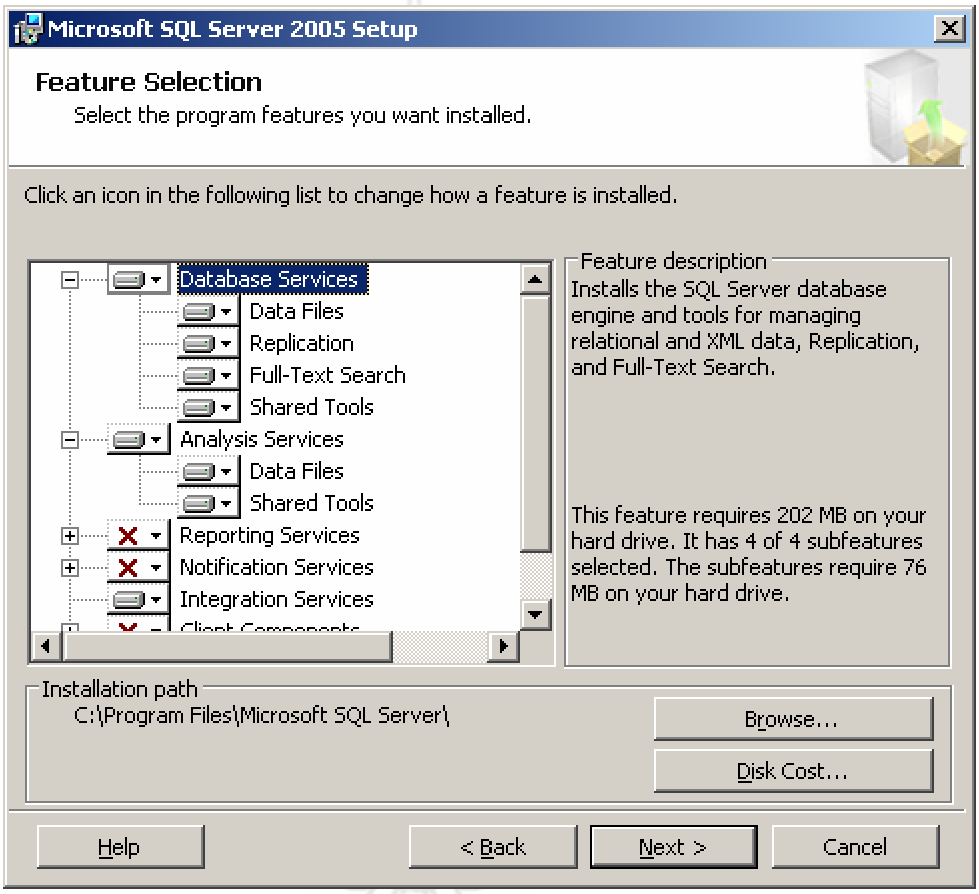
□ Notification Services

☑ Integration Services

□ Workstation components, Books Online and development tools

For more options, click Advanced.

7. If you clicked **Advanced** on the previous page, the **Feature Selection** page is displayed. On the **Feature Selection** page, select the program features to install using the drop-down boxes. To install components to a custom directory, select the feature and then click **Browse**. For more information about the functionality of this page, click **Help**. To continue when your feature selections are complete, click **Next.**



Microsoft SQL Server 2005 Setup

Feature Selection

Select the program features you want installed.

Click an icon in the following list to change how a feature is installed.

Feature description

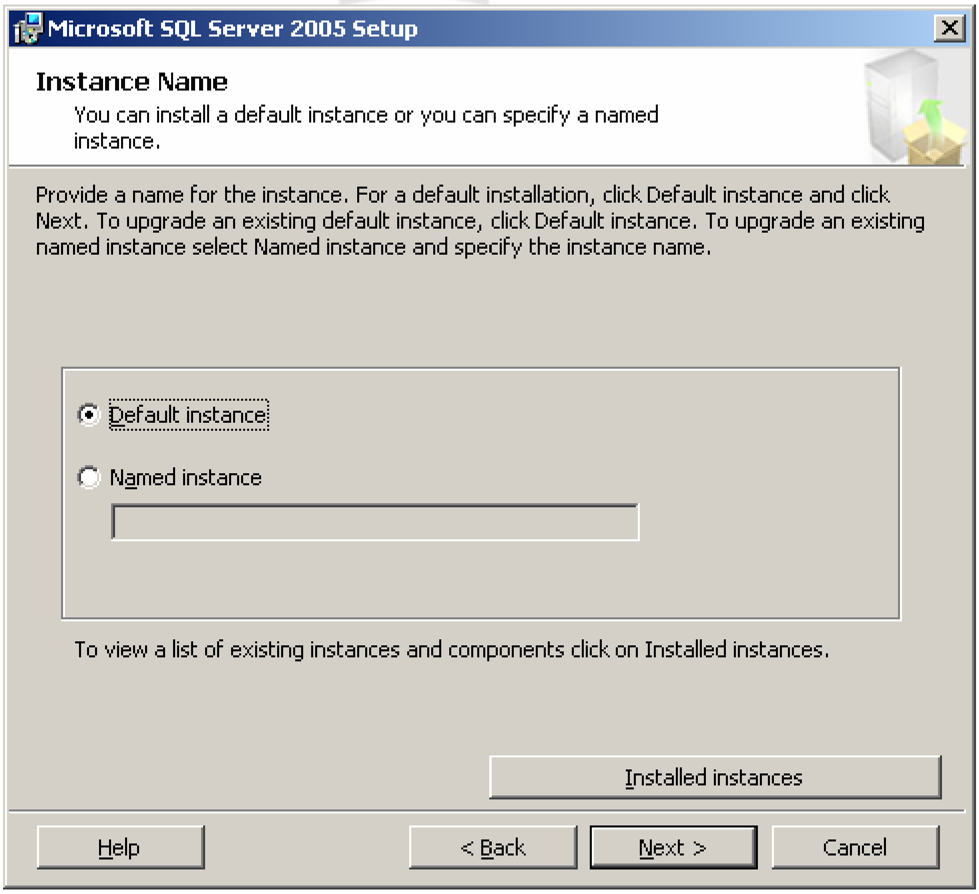
Installs the SQL Server database engine and tools for managing relational and XML data, Replication, and Full-Text Search.

This feature requires 202 MB on your hard drive. It has 4 of 4 subfeatures selected. The subfeatures require 76 MB on your hard drive.

Installation path

C:\Program Files\Microsoft SQL Server\

8. On the Instance Name page, select a default or named instance for your installation. If a default or named instance is already installed, and you select the existing instance for your installation, Setup upgrades it and provides you the option to install additional components. To install a new default instance, there must not be a default instance on the computer. To install a new named instance, click Named Instance and then type a unique instance name in the space provided. For more information about instance naming rules, click Help at the bottom of the page, or see the Instance Name topic in SQL Server 2005 Books Online.



Microsoft SQL Server 2005 Setup

Instance Name

You can install a default instance or you can specify a named instance.

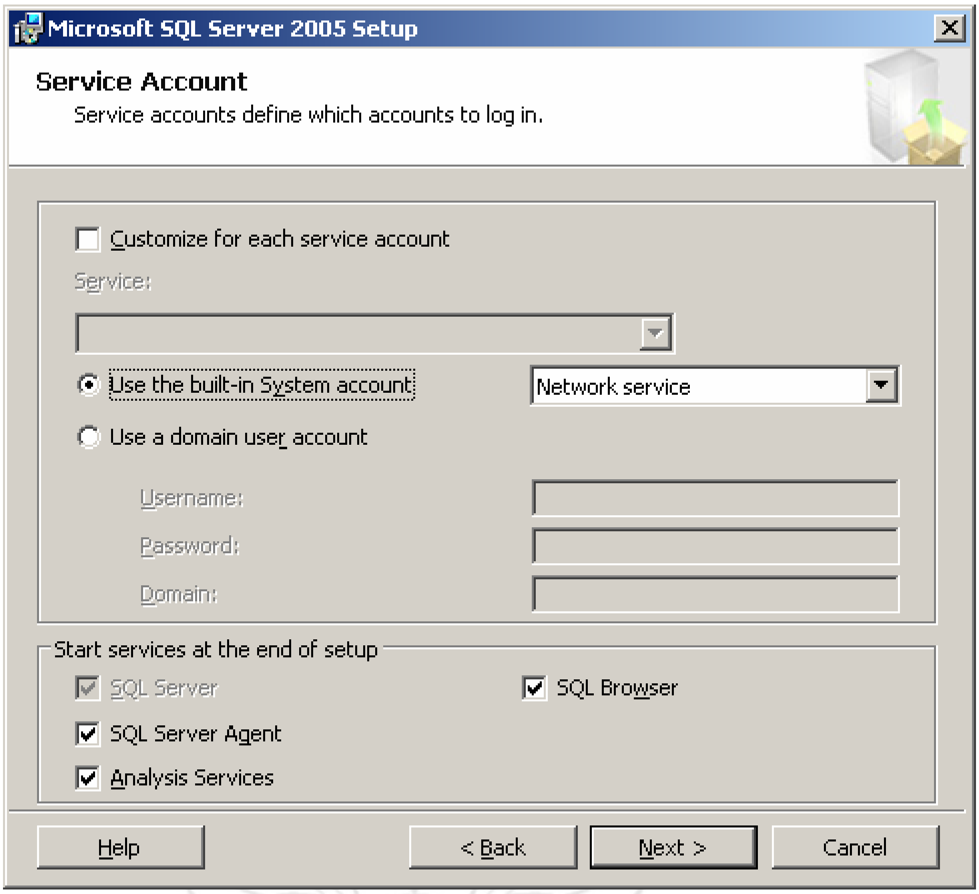
Provide a name for the instance. For a default installation, click Default instance and click Next. To upgrade an existing default instance, click Default instance. To upgrade an existing named instance select Named instance and specify the instance name.

● Default instance

○ Named instance

To view a list of existing instances and components click on Installed instances.

Use the **Service Account** page of the Microsoft SQL Server Installation Wizard to assign a login account to the SQL Server services. The actual services configured on this page depend on the features you have selected to install.



Microsoft SQL Server 2005 Setup

Service Account

Service accounts define which accounts to log in.

□ Customize for each service account

Service:

● Use the built-in System account: Network service

○ Use a domain user account

Username:

Password:

Domain:

Start services at the end of setup

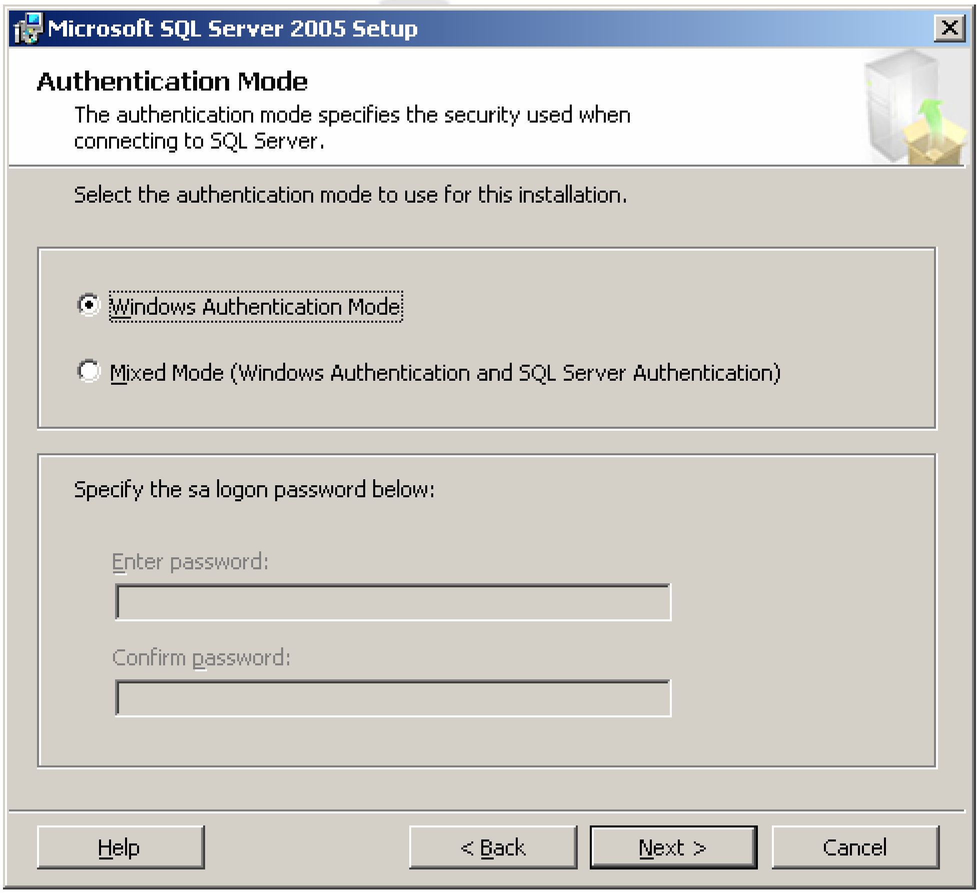
☑ SQL Server

☑ SQL Browser

☑ SQL Server Agent

☑ Analysis Services

Use the **Authentication Mode** page of the Microsoft SQL Server Installation Wizard to choose the security mode used to authenticate client and server connections to this installation. If you select **Mixed Mode**, you must enter and confirm the SQL Server system administrator (**sa**) password. After a device establishes a successful connection to SQL Server, the security mechanism is the same for both Windows Authentication and Mixed Mode.



Microsoft SQL Server 2005 Setup

Authentication Mode

The authentication mode specifies the security used when connecting to SQL Server.

Select the authentication mode to use for this installation.

● Windows Authentication Mode

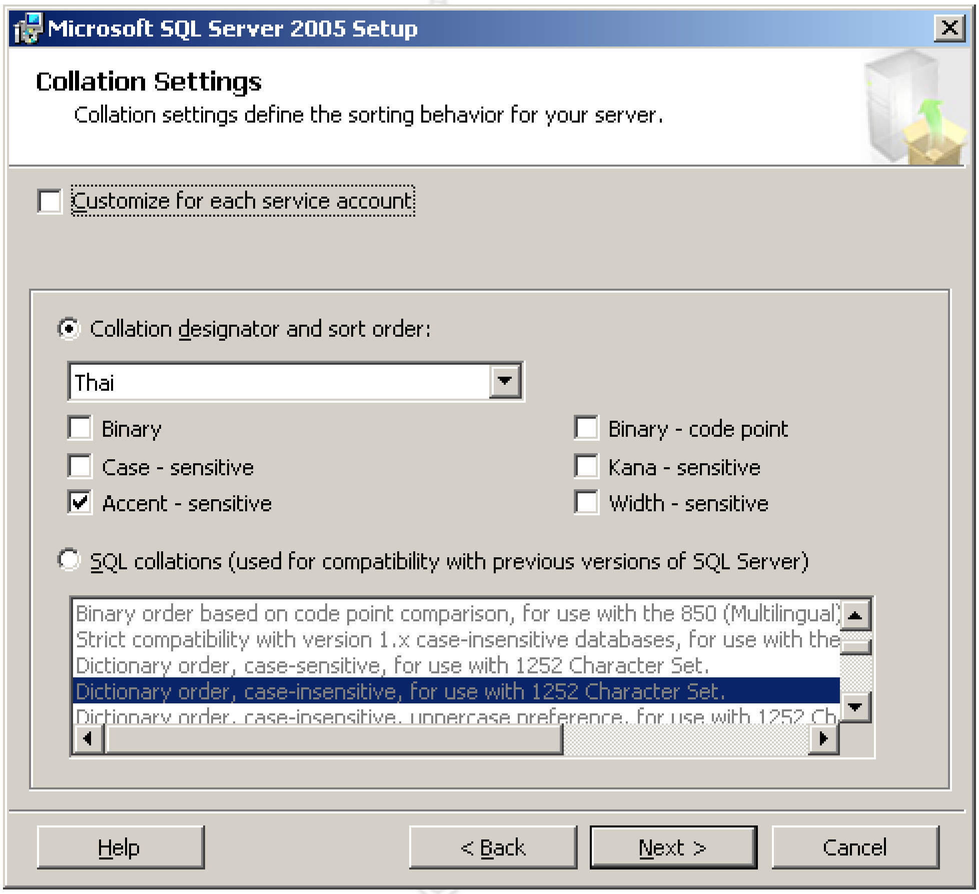
○ Mixed Mode (Windows Authentication and SQL Server Authentication)

Specify the sa logon password below:

Enter password:

Confirm password:

Use the **Collation Settings** page of the Microsoft SQL Server Installation Wizard to modify default collation settings used by the Database Engine and Analysis Services for language and sorting purposes. Choose the **Collation Designator** option to match collation settings of different installations of SQL Server or of another computer. Use the **SQL Collations** option to match settings that are compatible with the sort orders in earlier versions of SQL


Microsoft SQL Server 2005 Setup

Collation Settings

Collation settings define the sorting behavior for your server.

□ Customize for each service account

● Collation designator and sort order:

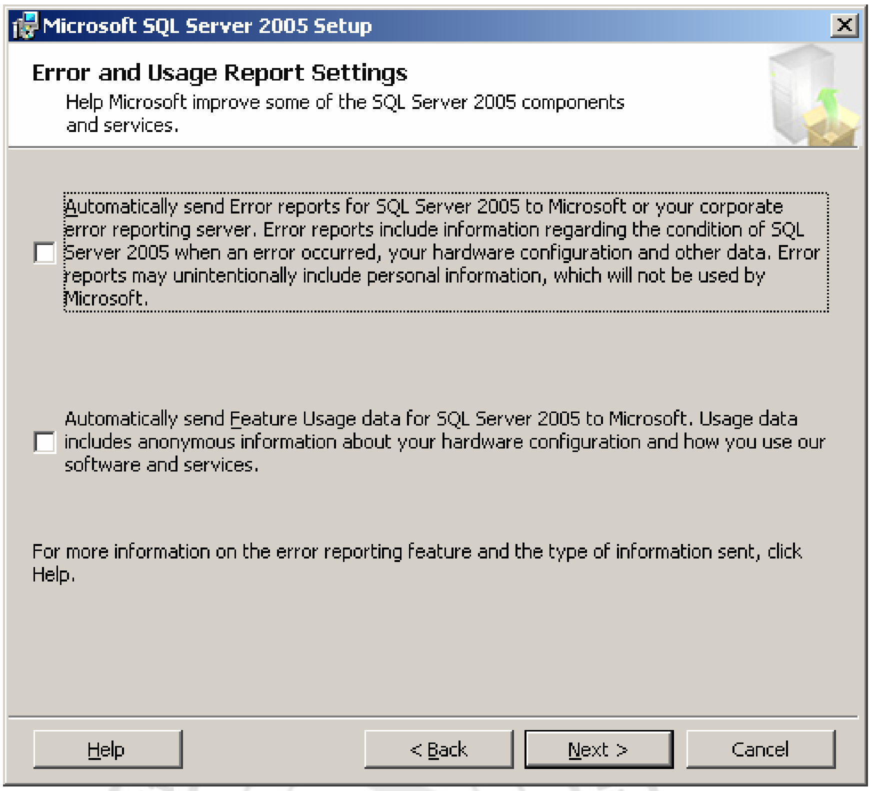
Thai

□ Binary
□ Case - sensitive
☑ Accent - sensitive
□ Binary - code point
□ Kana - sensitive
□ Width - sensitive

○ SQL collations (used for compatibility with previous versions of SQL Server)

Binary order based on code point comparison, for use with the 850 (Multilingual)
Strict compatibility with version 1.x case-insensitive databases, for use with the
Dictionary order, case-sensitive, for use with 1252 Character Set.
Dictionary order, case-insensitive, for use with 1252 Character Set.
Dictionary order, case-insensitive, uppercase preference, for use with 1252 Ch

Use the **Error and Usage Report Settings** page of the Microsoft SQL Server Installation Wizard to enable feature error and usage reporting functionality for SQL Server 2005.



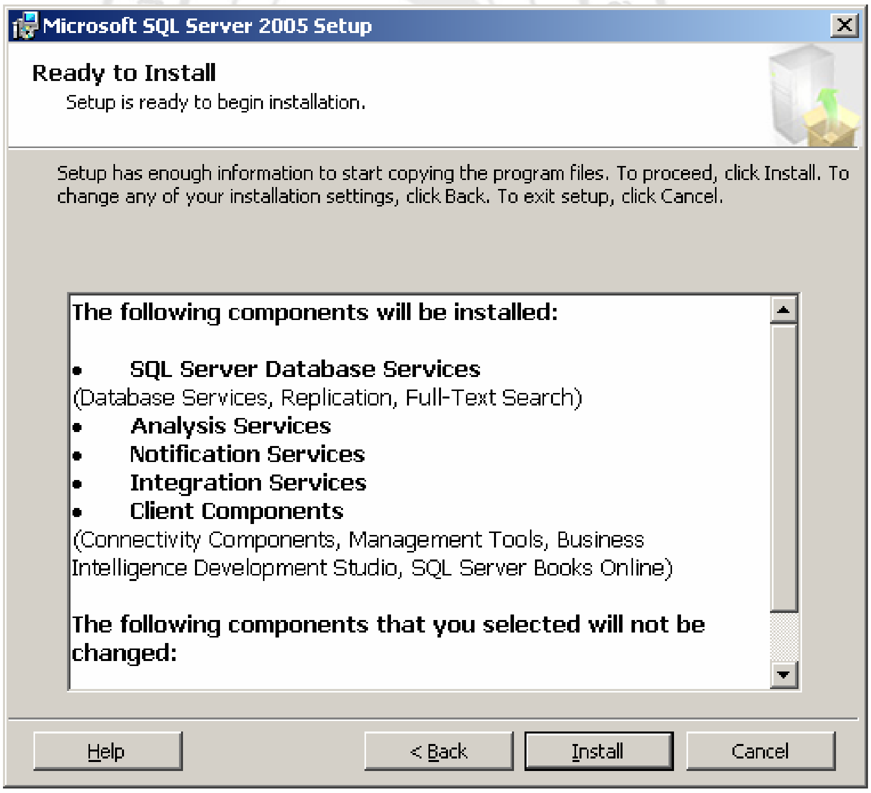
Error and Usage Report Settings

Help Microsoft improve some of the SQL Server 2005 components and services.

□ Automatically send Error reports for SQL Server 2005 to Microsoft or your corporate error reporting server. Error reports include information regarding the condition of SQL Server 2005 when an error occurred, your hardware configuration and other data. Error reports may unintentionally include personal information, which will not be used by Microsoft.

□ Automatically send Feature Usage data for SQL Server 2005 to Microsoft. Usage data includes anonymous information about your hardware configuration and how you use our software and services.

For more information on the error reporting feature and the type of information sent, click Help.



Ready to Install

Setup is ready to begin installation.

Setup has enough information to start copying the program files. To proceed, click Install. To change any of your installation settings, click Back. To exit setup, click Cancel.

The following components will be installed:

- SQL Server Database Services

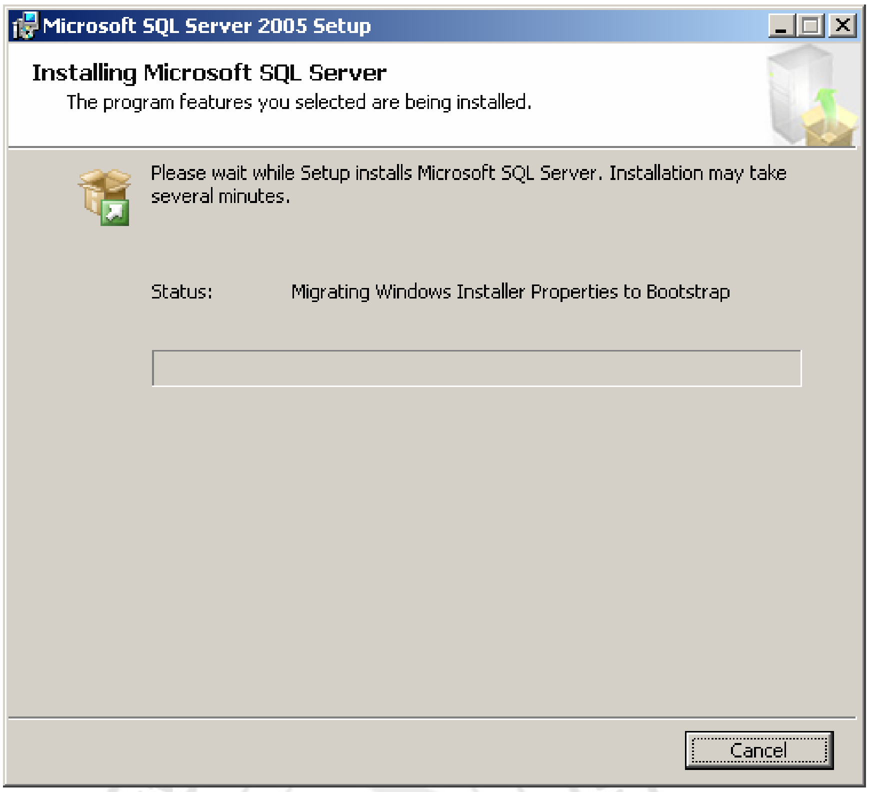
(Database Services, Replication, Full-Text Search)

- Analysis Services
- Notification Services
- Integration Services
- Client Components

(Connectivity Components, Management Tools, Business Intelligence Development Studio, SQL Server Books Online)

The following components that you selected will not be changed:

Use the **Setup Progress** page of the Microsoft SQL Server Installation Wizard to monitor the status of SQL Server Setup.



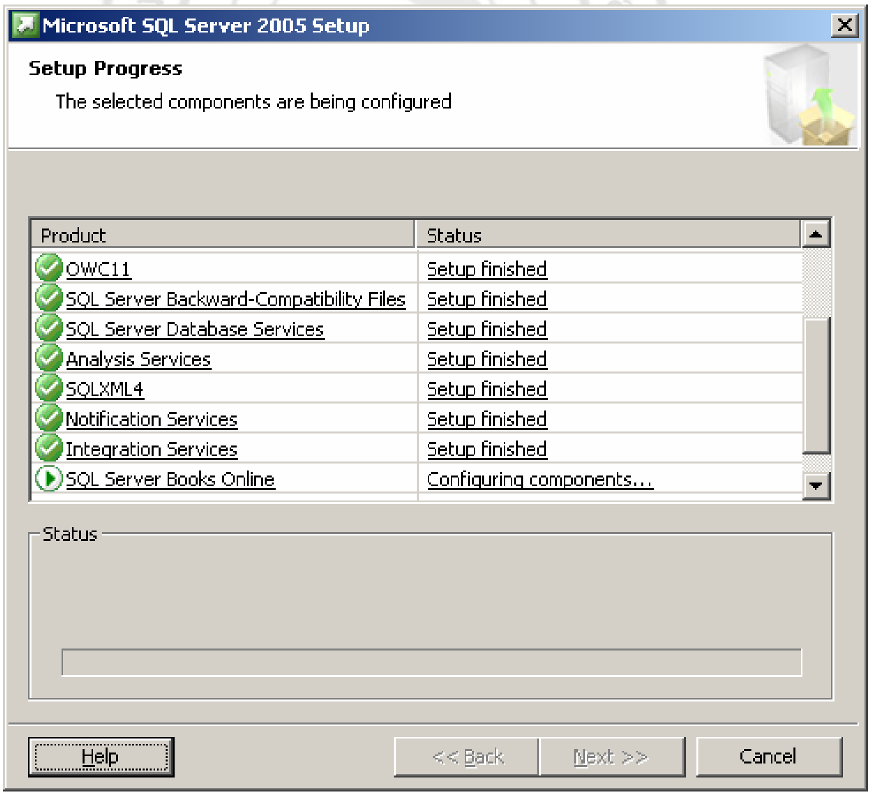
Microsoft SQL Server 2005 Setup

Installing Microsoft SQL Server

The program features you selected are being installed.

Please wait while Setup installs Microsoft SQL Server. Installation may take several minutes.

Status: Migrating Windows Installer Properties to Bootstrap



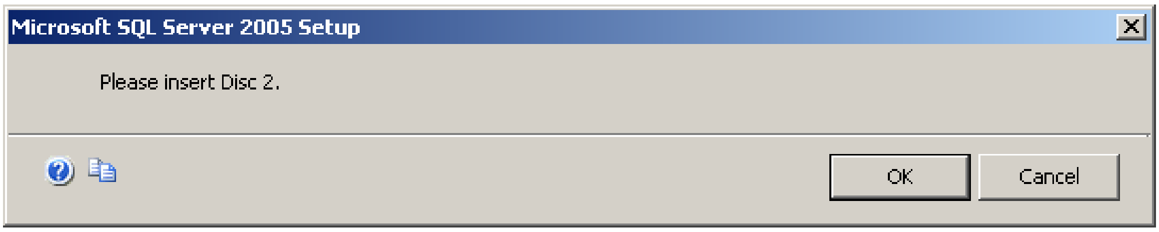
Microsoft SQL Server 2005 Setup

Setup Progress

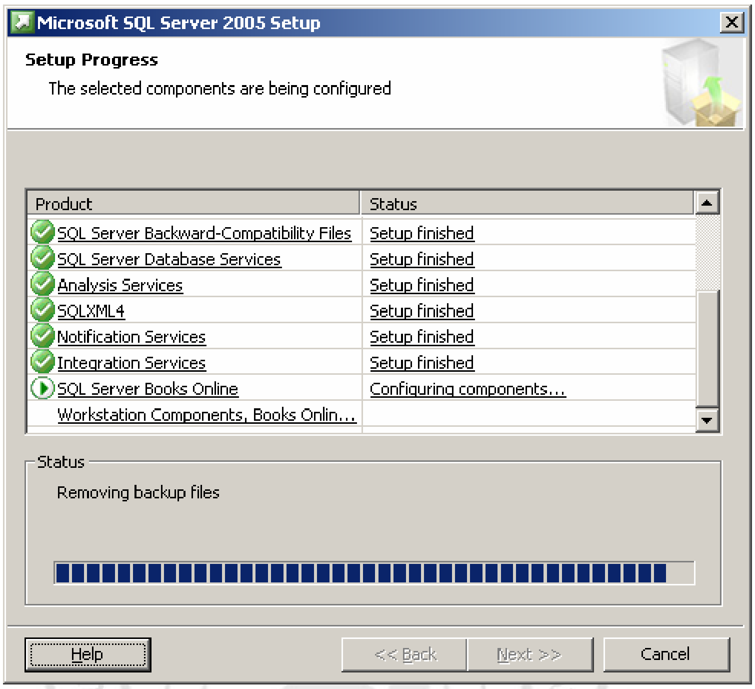
The selected components are being configured

Product	Status
✓ OWC11	Setup finished
✓ SQL Server Backward-Compatibility Files	Setup finished
✓ SQL Server Database Services	Setup finished
✓ Analysis Services	Setup finished
✓ SQLXML4	Setup finished
✓ Notification Services	Setup finished
✓ Integration Services	Setup finished
▶ SQL Server Books Online	Configuring components...

Status



Please insert Disc 2.



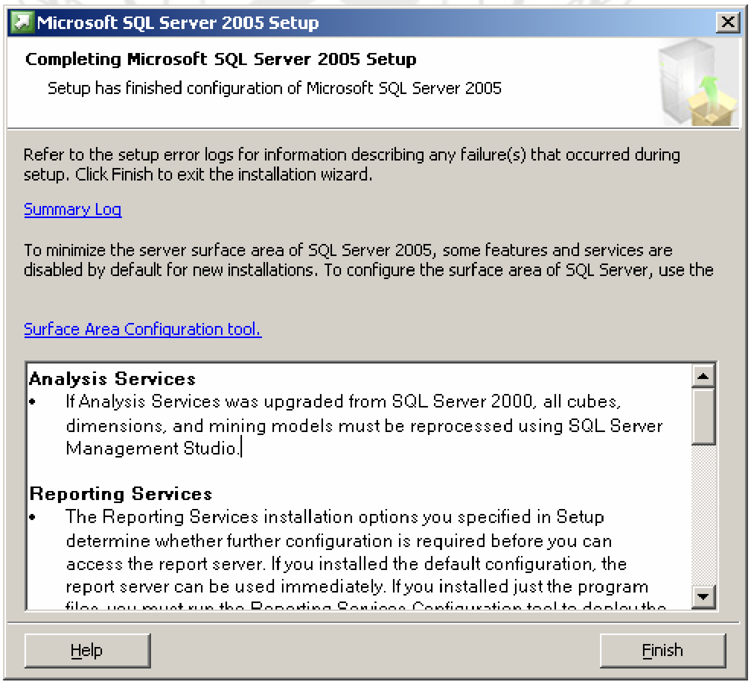
Setup Progress

The selected components are being configured

Product	Status
✓ SQL Server Backward-Compatibility Files	Setup finished
✓ SQL Server Database Services	Setup finished
✓ Analysis Services	Setup finished
✓ SQLXML4	Setup finished
✓ Notification Services	Setup finished
✓ Integration Services	Setup finished
SQL Server Books Online	Configuring components...
Workstation Components, Books Onlin...	

Status

Removing backup files



Completing Microsoft SQL Server 2005 Setup

Setup has finished configuration of Microsoft SQL Server 2005

Refer to the setup error logs for information describing any failure(s) that occurred during setup. Click Finish to exit the installation wizard.

Summary Log

To minimize the server surface area of SQL Server 2005, some features and services are disabled by default for new installations. To configure the surface area of SQL Server, use the

Surface Area Configuration tool.

Analysis Services

- If Analysis Services was upgraded from SQL Server 2000, all cubes, dimensions, and mining models must be reprocessed using SQL Server Management Studio.

Reporting Services

- The Reporting Services installation options you specified in Setup determine whether further configuration is required before you can access the report server. If you installed the default configuration, the report server can be used immediately. If you installed just the program files, you must run the Reporting Services Configuration tool to deploy the

Analysis Services

●If Analysis Services was upgraded from SQL Server 2000, all cubes, dimensions, and mining models must be reprocessed using SQL Server Management Studio.

Reporting Services

●The Reporting Services installation options you specified in Setup determine whether further configuration is required before you can access the report server. If you installed the default configuration, the report server can be used immediately. If you installed just the program files, you must run the Reporting Services Configuration tool to deploy the report server.

Integration Services

- SQL Server 2000 Data Transformation Services packages can run side by side with SQL Server 2005 Integration Services packages. Read SQL Server 2005 Books Online to learn more about migrating SQL Server 2000 packages to SQL Server 2005.

Notification Services

●You can deploy and manage Notification Services instances using SQL Server Management Studio or the nscontrol command-prompt utility, as well as programmatically.

Documentation and Samples

●To add Microsoft Visual Studio documentation to Business Intelligence Development Studio, install the MSDN Library from the SQL Server 2005 installation media.

●To install the .NET Framework SDK, see "Installing the .NET Framework SDK" in SQL Server Books Online.

- To install sample databases and code samples, see "Running Setup to Install AdventureWorks Sample Databases and Samples" in SQL Server 2005 Books Online.

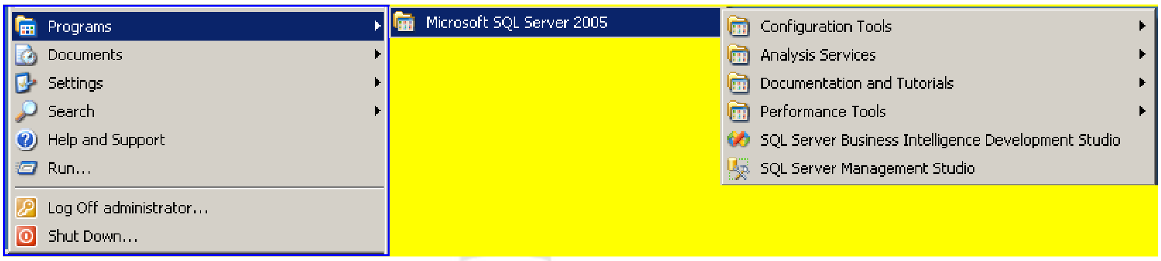
Step 3: Configure Your SQL Server 2005 Installation

After Setup completes the installation of Microsoft SQL Server 2005, you can further configure SQL Server using graphical and command prompt utilities. The following table describes support for tools used to manage an instance of SQL Server 2005.

Tool or utility	Description
SQL Server Management Studio	SQL Server Management Studio is used for editing and executing queries, and for launching standard wizard tasks. For more information, see Introducing SQL Server Management Studio.
SQL Server Profiler	SQL Server Profiler provides a graphical user interface for monitoring an instance of the SQL Server database engine or an instance of Analysis Services.
Database Engine Tuning Advisor	Database Engine Tuning Advisor helps create optimal sets of indexes, indexed views, and partitions.
Business Intelligence Development Studio	Business Intelligence Development Studio is an integrated development environment for Analysis Services and Integration Services solutions.
Command Prompt Utilities	Manage SQL Server objects from the command prompt.
SQL Server Configuration Manager	Manage server and client network configuration settings.
Import and Export Data	Integration Services provides a set of graphical tools and programmable objects for moving, copying, and transforming data.
SQL Server Setup	Install, upgrade to, or change components in an instance of SQL Server 2005.

To complete the installation of the samples, after Setup, perform one of the following steps:

- From the **Start** menu, click **All Programs**, click **Microsoft SQL Server 2005**, click **SQL Server Business Intelligence Development Studio**.



APPENDIX C

SYSTEM ADMINISTRATOR MANUAL

This appendix is presented data mining software. SQL Server 2005 Data Mining, produced by Microsoft. This section shows about source data description and building risk models with Microsoft Decision Trees algorithms. This section contain step-by-step for building insurance risk models as following steps.

Step 1: Creating risk insurance Data source

Step 2: Building Model

Step 3: Creating a Microsoft Decision tree

Step 4: Specific algorithm parameter

Step 5: Testing the Accuracy of the Mining Models

Step 6: Viewing the Lift Chart

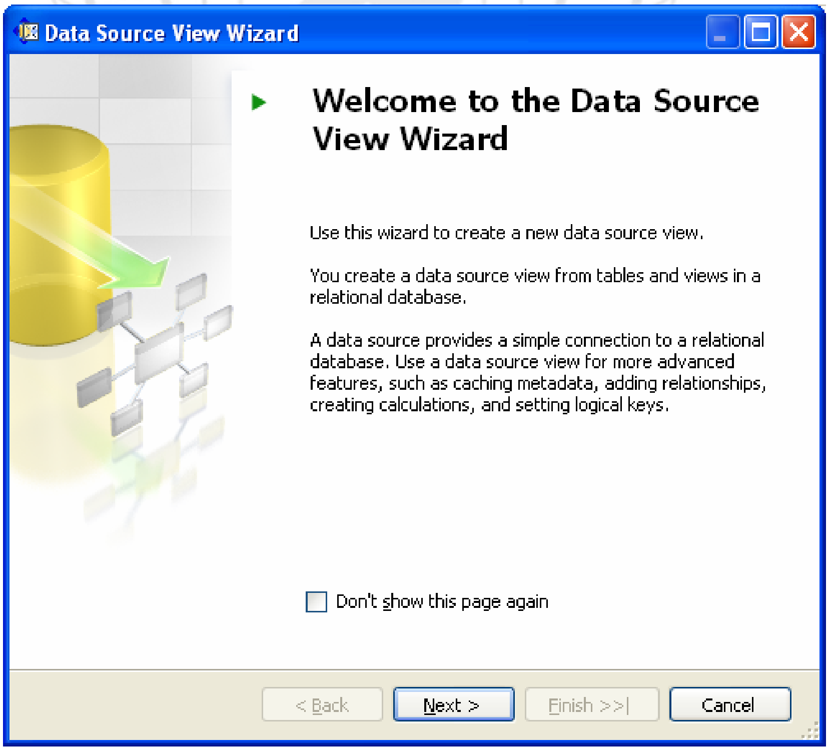
Step 7: The Classification Matrix

Step 8: Package for risk mining

Step 1. Creating risk insurance Data source

A data source view provides an abstraction of the data source. By using data source views, can select the tables that relate to your particular project, establish relationships between tables, and add calculated columns and named views without modifying the original data source

1. In Solution Explorer, right-click **Data Source Views**, and select **New Data Source View**. Click next.



Welcome to the Data Source View Wizard

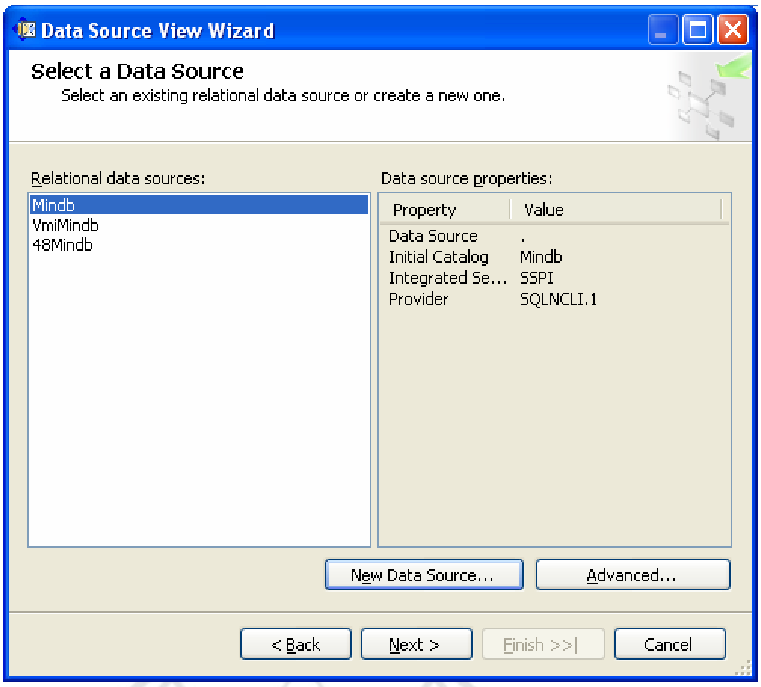
Use this wizard to create a new data source view.

You create a data source view from tables and views in a relational database.

A data source provides a simple connection to a relational database. Use a data source view for more advanced features, such as caching metadata, adding relationships, creating calculations, and setting logical keys.

☐ Don't show this page again

2. Select data source Mindb and Click Next.



Data Source View Wizard

Select a Data Source

Select an existing relational data source or create a new one.

Relational data sources:

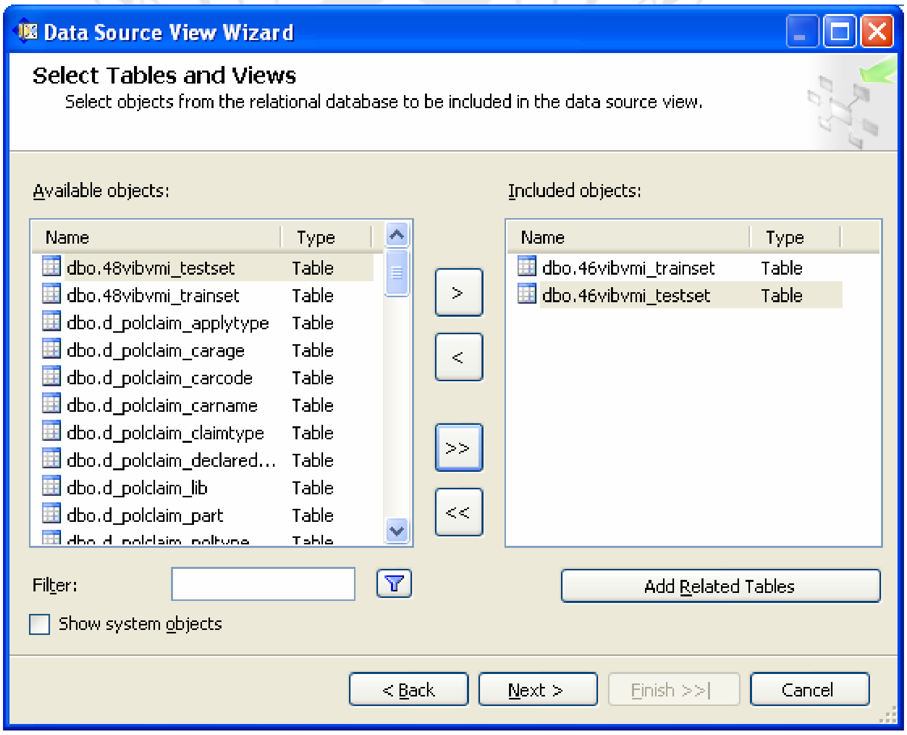
- Mindb
- VmiMindb
- 48Mindb

Data source properties:

Property	Value
Data Source	.
Initial Catalog	Mindb
Integrated Se...	SSPI
Provider	SQLNCLI.1

3. Select the following tables, and then click the right arrow to include them in the new data source view and click next

- dbo.46vibvmi_trainset
- dbo.46vibvmi_testset



Data Source View Wizard

Select Tables and Views

Select objects from the relational database to be included in the data source view.

Available objects:

Name	Type
dbo.48vibvmi_testset	Table
dbo.48vibvmi_trainset	Table
dbo.d_polclaim_applytype	Table
dbo.d_polclaim_carage	Table
dbo.d_polclaim_carcode	Table
dbo.d_polclaim_carname	Table
dbo.d_polclaim_claimtype	Table
dbo.d_polclaim_declared...	Table
dbo.d_polclaim_lib	Table
dbo.d_polclaim_part	Table
dbo.d_polclaim_poltype	Table

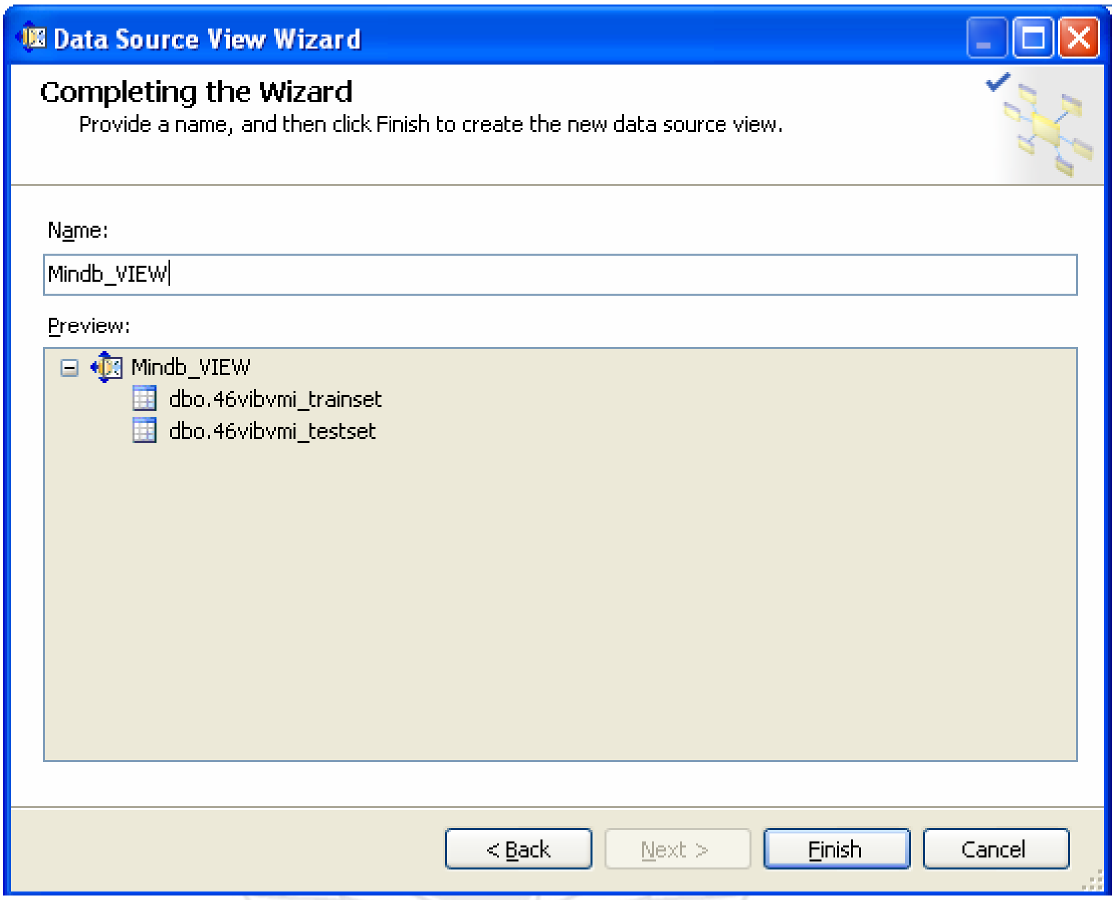
Included objects:

Name	Type
dbo.46vibvmi_trainset	Table
dbo.46vibvmi_testset	Table

Filter:

□ Show system objects

4 On the **Completing the Wizard** page, by default the data source view is named Mindb_VIEW Click **Finish**.



Completing the Wizard

Provide a name, and then click Finish to create the new data source view.

Name:

Mindb_VIEW

Preview:

- Mindb_VIEW
  - dbo.46vibvmi_trainset
  - dbo.46vibvmi_testset

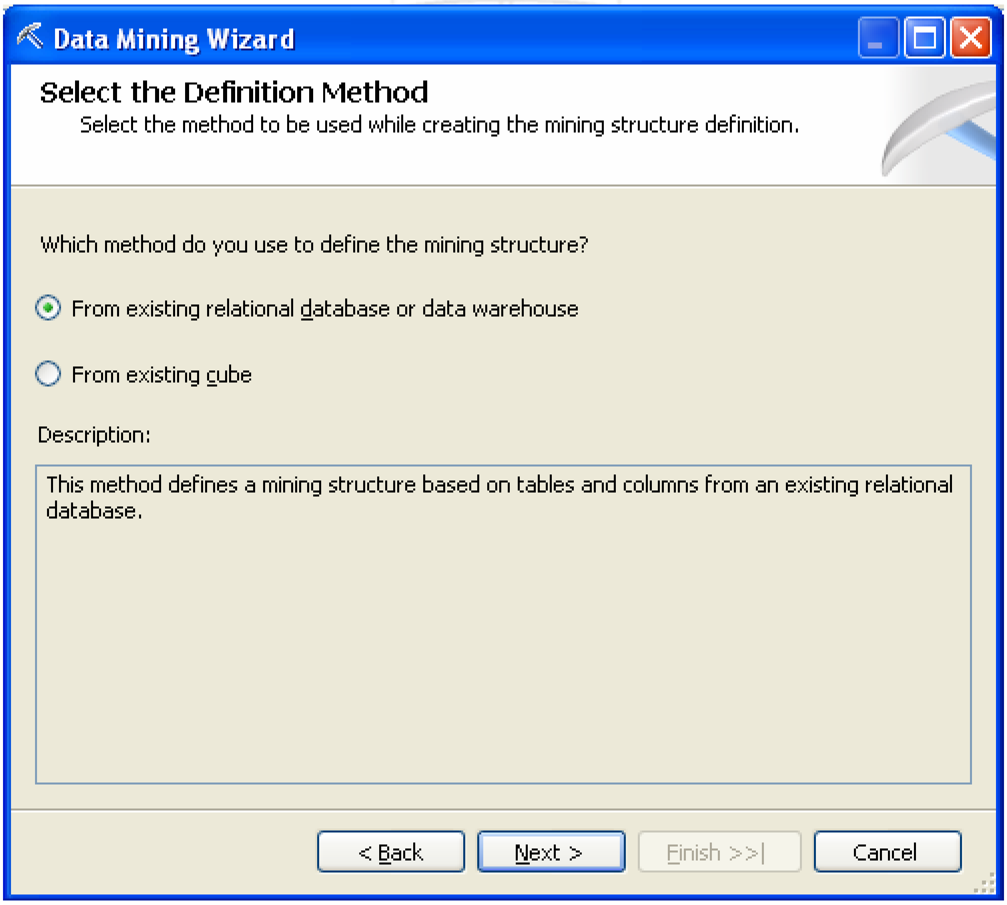
Step 2. Building Model

In this section you will create a insurance risk model. After you complete the tasks in this section, will have set of mining models that will suggest the most likely customers to be claimed with risk level.

1. In Solution Explorer, right-click Mining Structures and select New Mining Structure. The Data Mining Wizard opens.

2. On the Welcome to the Data Mining Wizard page, click Next.

3. On the Select the Definition Method page, verify that From existing relational database or data warehouse is selected, and then click Next.



Select the Definition Method

Select the method to be used while creating the mining structure definition.

Which method do you use to define the mining structure?

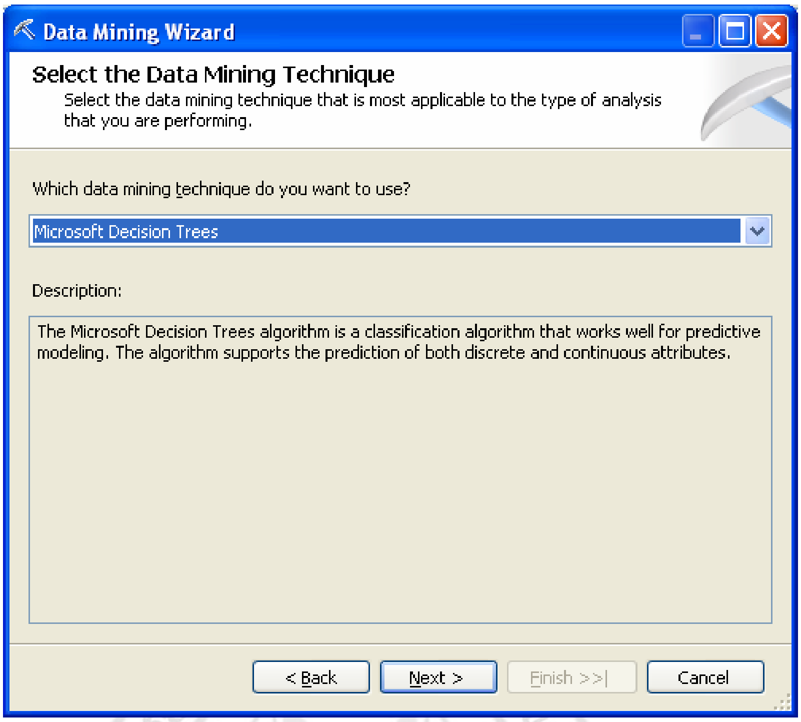
● From existing relational database or data warehouse

○ From existing cube

Description:

This method defines a mining structure based on tables and columns from an existing relational database.

4. On the Select the Data Mining Technique page, under Which data mining technique do you want to use?, select Microsoft Decision Trees. Click Next.



Select the Data Mining Technique

Select the data mining technique that is most applicable to the type of analysis that you are performing.

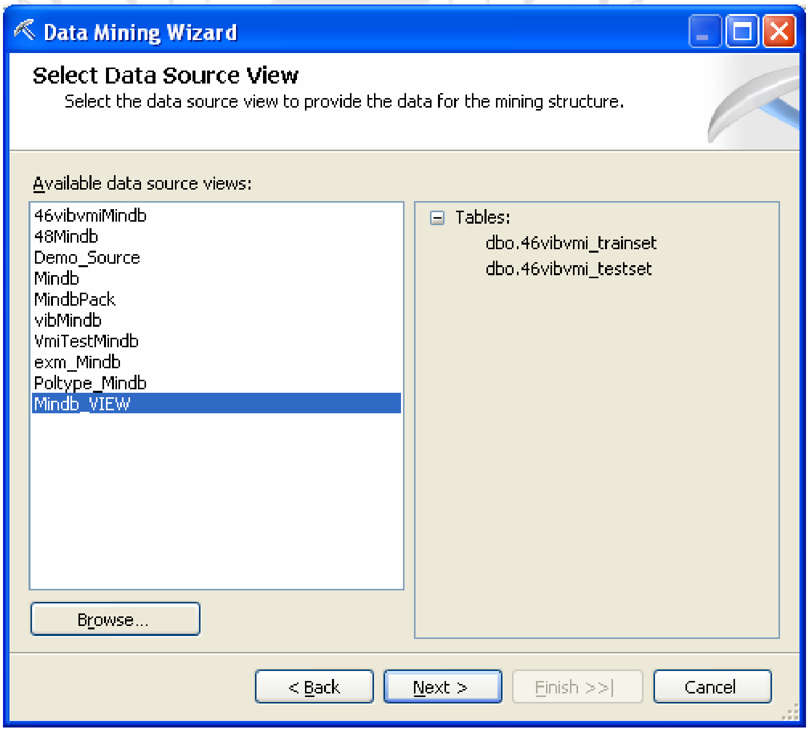
Which data mining technique do you want to use?

Microsoft Decision Trees

Description:

The Microsoft Decision Trees algorithm is a classification algorithm that works well for predictive modeling. The algorithm supports the prediction of both discrete and continuous attributes.

5. On the Select Data Source View page, select Mindb_ViewClick view the tables in the data source view, and then click Next.



Select Data Source View

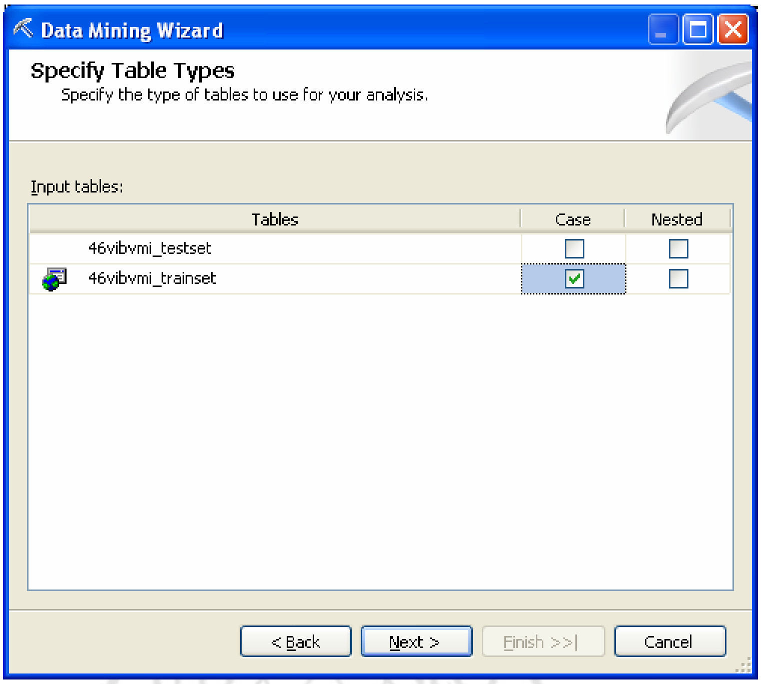
Select the data source view to provide the data for the mining structure.

Available data source views:

46vibvmiMindb
48Mindb
Demo_Source
Mindb
MindbPack
vibMindb
VmiTestMindb
exm_Mindb
Poltype_Mindb
Mindb_VIEW

⊟ Tables:
dbo.46vibvmi_trainset
dbo.46vibvmi_testset

6. On the Specify Table Types page, select the check box in the Case column next to the 46vibvmi_trainset table, and then click Next.



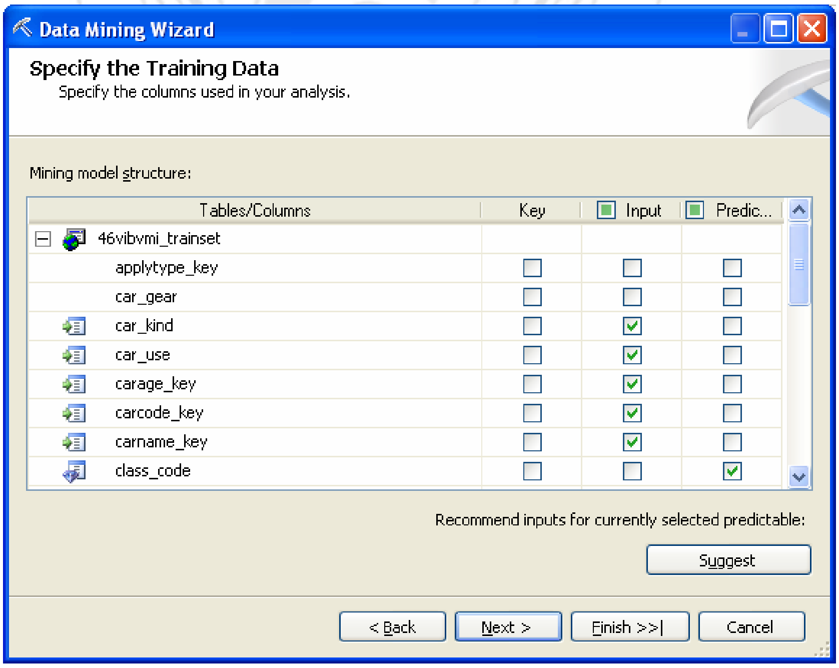
Specify Table Types

Specify the type of tables to use for your analysis.

Input tables:

Tables	Case	Nested
46vibvmi_testset	□	□
46vibvmi_trainset	☑	□

7. On the Specify the Training Data page, verify that the check box in the Key column is selected next to the Pol_no column. Select Input and Predictable next to the Class_code column. Select the Input check boxes next to and click next



Specify the Training Data

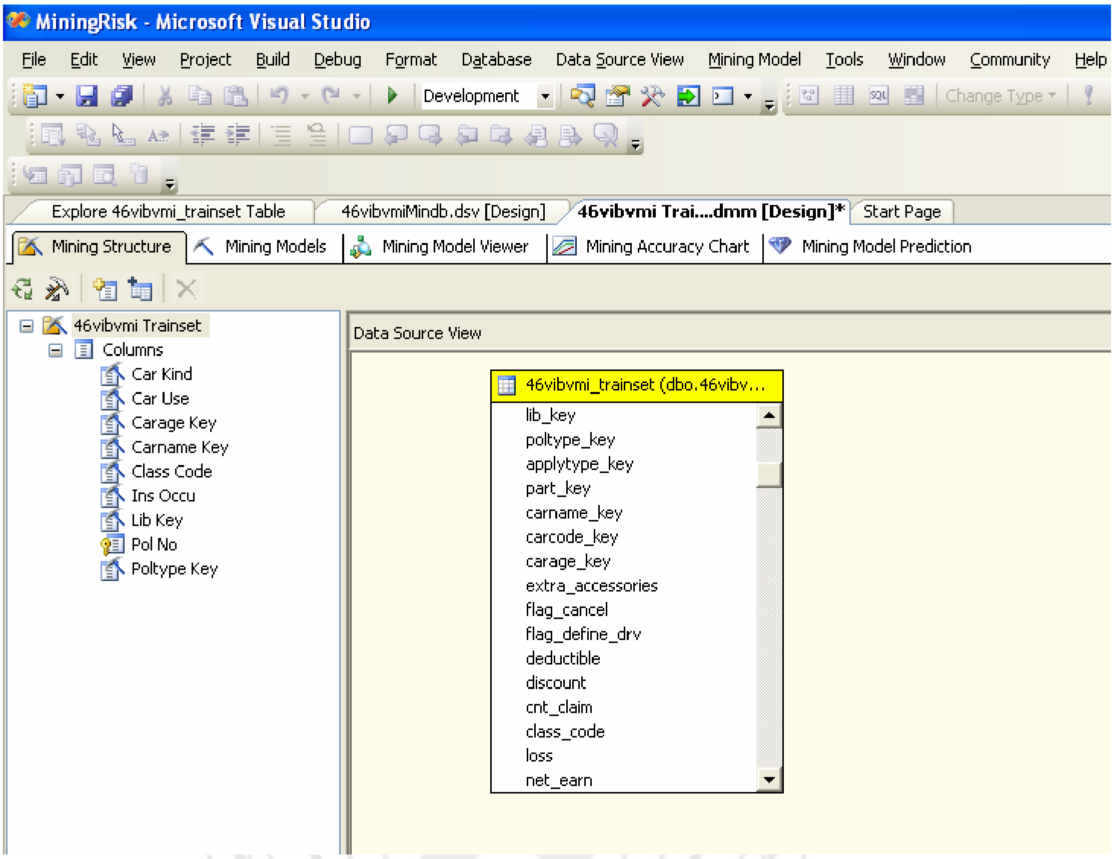
Specify the columns used in your analysis.

Mining model structure:

Tables/Columns	Key	■ Input	■ Predic...
⊟ 46vibvmi_trainset			
applytype_key	□	□	□
car_gear	□	□	□
car_kind	□	☑	□
car_use	□	☑	□
carage_key	□	☑	□
carcode_key	□	☑	□
carname_key	□	☑	□
class_code	□	□	☑

Recommend inputs for currently selected predictable:

8. On the Completing the Wizard page, in Mining structure name, type entropy name and click Finish. In this step we have mining model with decision tree algorithm.



Step 3. Creating a Microsoft Decision tree

The initial mining structure that you created in the previous task contains a single mining model that is based on the Microsoft Decision Trees algorithm. In this task, you will define additional models by using the **Mining Models** tab of Data Mining Designer. In this task will define a model and a set algorithm parameter.

1. On **Mining Models** tab in Data Mining Designer in Business Intelligence Development Studio.

2. Right-click the **Structure** column and select **New Mining Model**. The **New Mining Model** dialog box opens.

3. In **Model name**, type **K2**

4. In **Algorithm name**, select **Microsoft Decision Tree** and click **OK**.

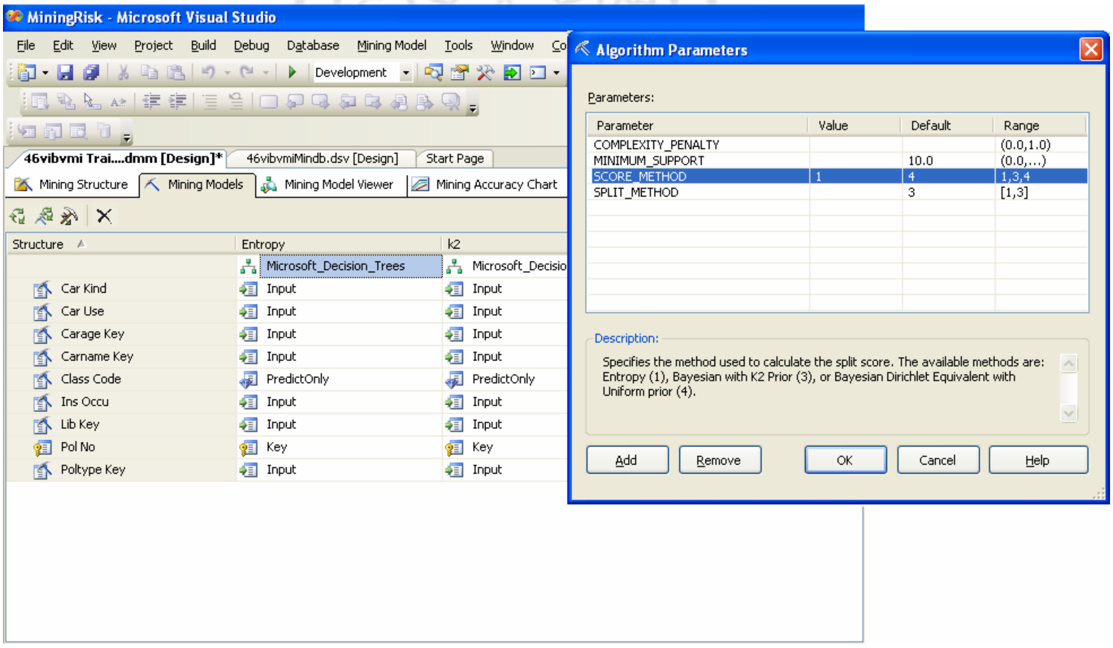
Step 4. Specific algorithm parameter

Microsoft decision tree has method used to calculate the split score. It used to control the tree growth, tree shape the methods are Entropy and Bayesian with K2.

1. On **Mining Models** tab in Data Mining Designer in Business Intelligence Development Studio.

2. Right-click the **Structure** column and select **Set Algorithm Parameters**. The **Algorithm Parameters** dialog box opens.

3. In **Score method**, type **1 value for entropy method and clicked OK**


MiningRisk - Microsoft Visual Studio
File Edit View Project Build Debug Database Mining Model Tools Window Co
Development
46vibvmi Trai....dmm [Design]*
46vibvmiMindb.dsv [Design]
Start Page
Mining Structure
Mining Models
Mining Model Viewer
Mining Accuracy Chart
Algorithm Parameters
Add
Remove
OK
Cancel
Help

Structure	Entropy	k2
	Microsoft_Decision_Trees	Microsoft_Decisio
Car Kind	Input	Input
Car Use	Input	Input
Carage Key	Input	Input
Carname Key	Input	Input
Class Code	PredictOnly	PredictOnly
Ins Occu	Input	Input
Lib Key	Input	Input
Pol No	Key	Key
Poltype Key	Input	Input

Parameters:

Parameter	Value	Default	Range
COMPLEXITY_PENALTY			(0.0,1.0)
MINIMUM_SUPPORT		10.0	(0.0,...)
SCORE_METHOD	1	4	1,3,4
SPLIT_METHOD		3	[1,3]

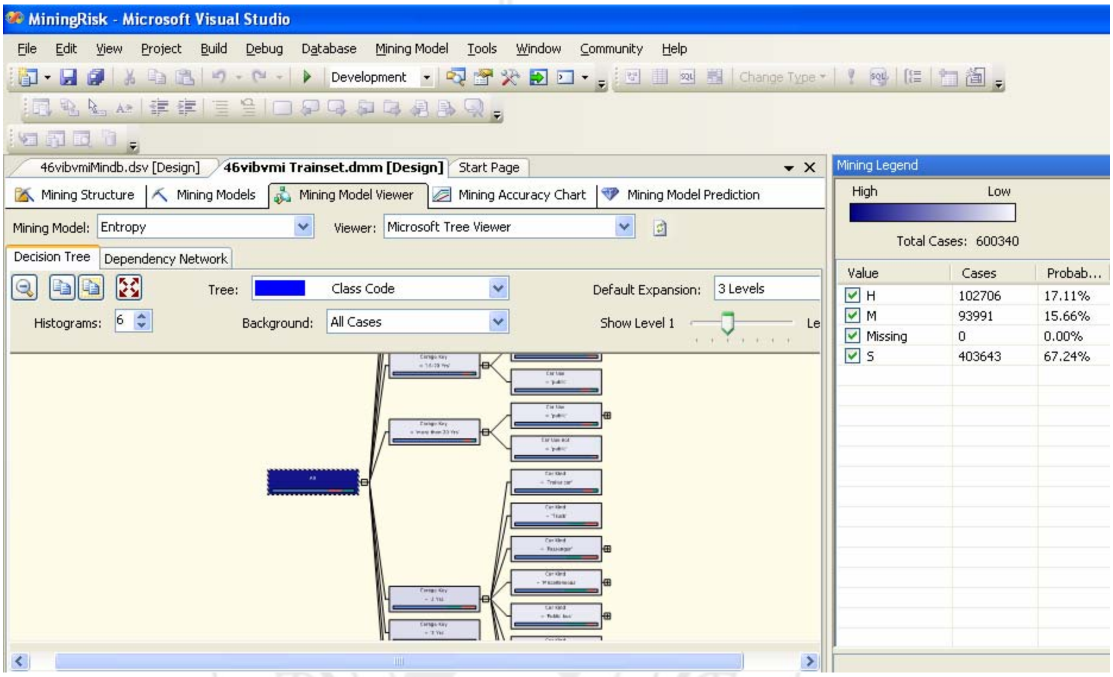
Description:

Specifies the method used to calculate the split score. The available methods are: Entropy (1), Bayesian with K2 Prior (3), or Bayesian Dirichlet Equivalent with Uniform prior (4).

After the models in your project are processed, you can view them by using the **Mining Model Viewer** tab in Data Mining Designer. You can use the **Mining Model** list at the top of the tab to examine the individual models in the mining structure.

Each node in the decision tree displays the following information:

- The condition that is required to reach that node from the node that comes before it.
- A histogram that describes the distribution of states for the predictable column in order of popularity. You can control how many states appear in the histogram by using the **Histograms** control.


MiningRisk - Microsoft Visual Studio
File Edit View Project Build Debug Database Mining Model Tools Window Community Help
Development
Change Type
46vibvmiMindb.dsv [Design]
46vibvmi Trainset.dmm [Design]
Start Page
Mining Structure
Mining Models
Mining Model Viewer
Mining Accuracy Chart
Mining Model Prediction
Decision Tree
Dependency Network
Mining Legend
High
Low

Mining Model: Entropy

Viewer: Microsoft Tree Viewer

Tree: Class Code

Default Expansion: 3 Levels

Histograms: 6

Background: All Cases

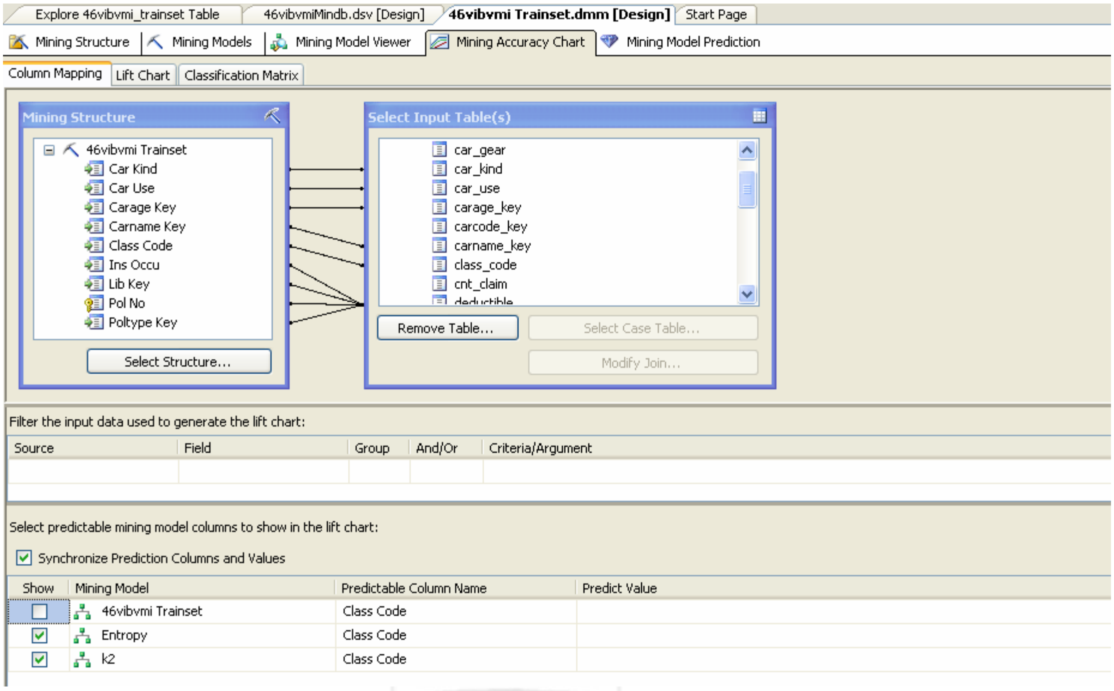
Show Level 1

Total Cases: 600340

Value	Cases	Probab...
☑ H	102706	17.11%
☑ M	93991	15.66%
☑ Missing	0	0.00%
☑ S	403643	67.24%

Step 5. Testing the Accuracy of the Mining Models

After you have built, processed, and explored the mining models you can test the models to determine how well they perform predictions and to determine whether one of the models performs better than the others. On the Mining Accuracy Chart tab for Data Mining can compare the results of each model directly against the results of the other models.



Mining Structure

- 46vibvmi Trainset
  - Car Kind
  - Car Use
  - Carage Key
  - Carname Key
  - Class Code
  - Ins Occu
  - Lib Key
  - Pol No
  - Poltype Key

Select Input Table(s)

- car_gear
- car_kind
- car_use
- carage_key
- carcode_key
- carname_key
- class_code
- cnt_claim
- deductible

Filter the input data used to generate the lift chart:

Source	Field	Group	And/Or	Criteria/Argument

Select predictable mining model columns to show in the lift chart:

☑ Synchronize Prediction Columns and Values

Show	Mining Model	Predictable Column Name	Predict Value
☐	46vibvmi Trainset	Class Code	
☑	Entropy	Class Code	
☑	k2	Class Code	

Step 6. Viewing the Lift Chart

The lift chart is important because it helps distinguish between models in a structure that are almost the same, to help you determine which model provides the best predictions. Similarly, the lift chart shows which type of algorithm performs the best predictions for a particular situation. The results of the comparison are then sorted and plotted on a graph. An ideal model, a theoretical model that predicts the result correctly 100 percent of the time, is also plotted on the graph.

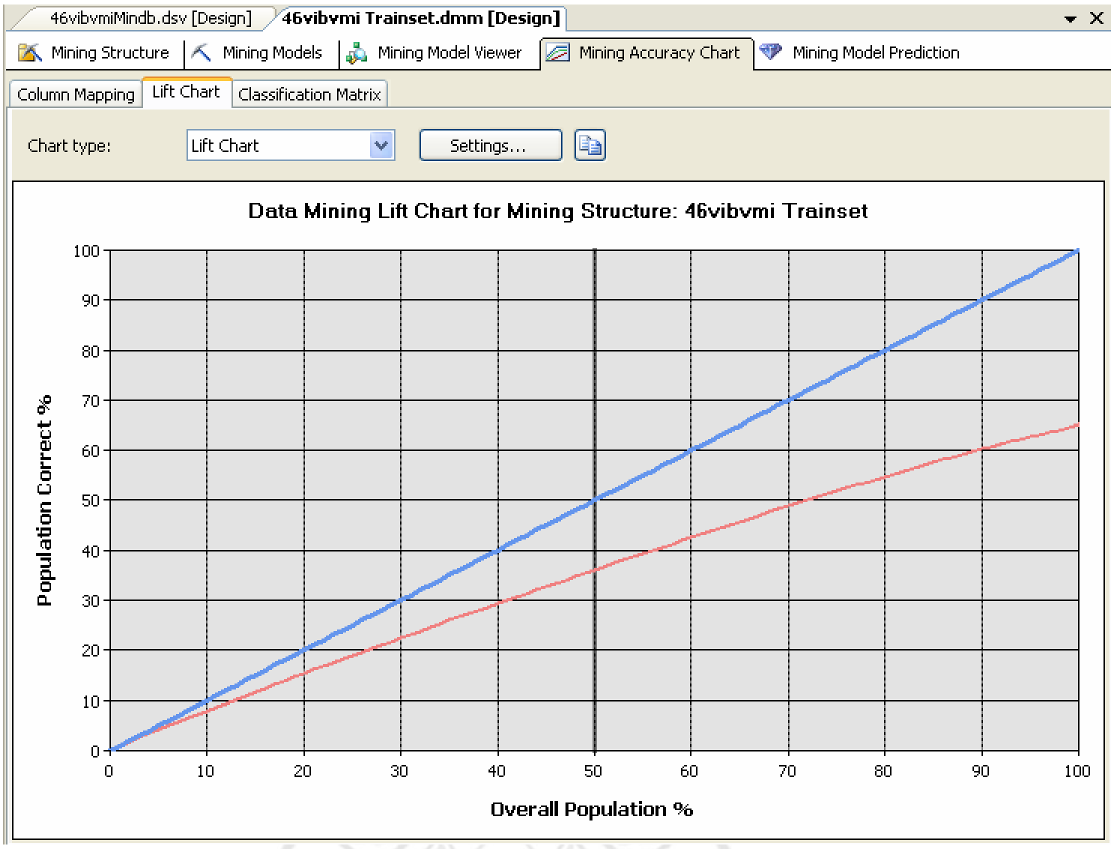
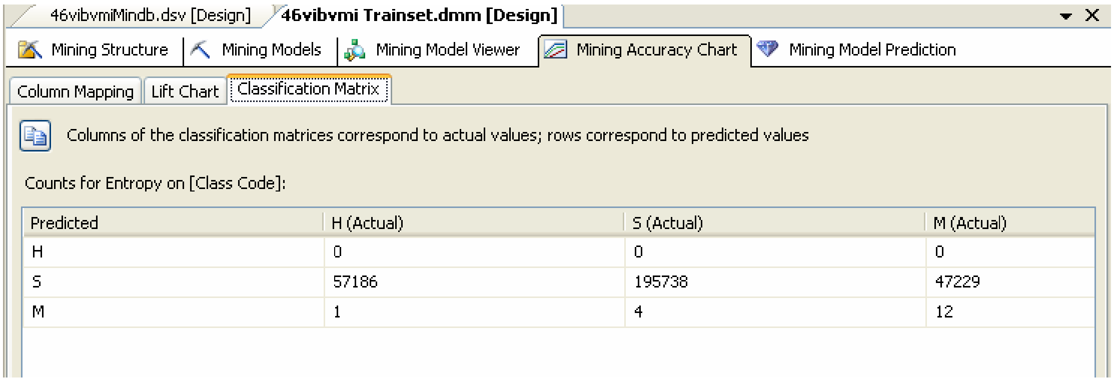


Chart type: Lift Chart

Step 7. The Classification Matrix

The Classification Matrix tab provides another way to examine how accurately the mining models in a structure create predictions. A classification matrix is built as a comparison of actual values that exist in the testing dataset against the values that the mining model predicts. The matrix is a valuable tool because it not only shows how frequently the model correctly predicted a value, but also shows which other values the model most frequently predicted incorrectly.



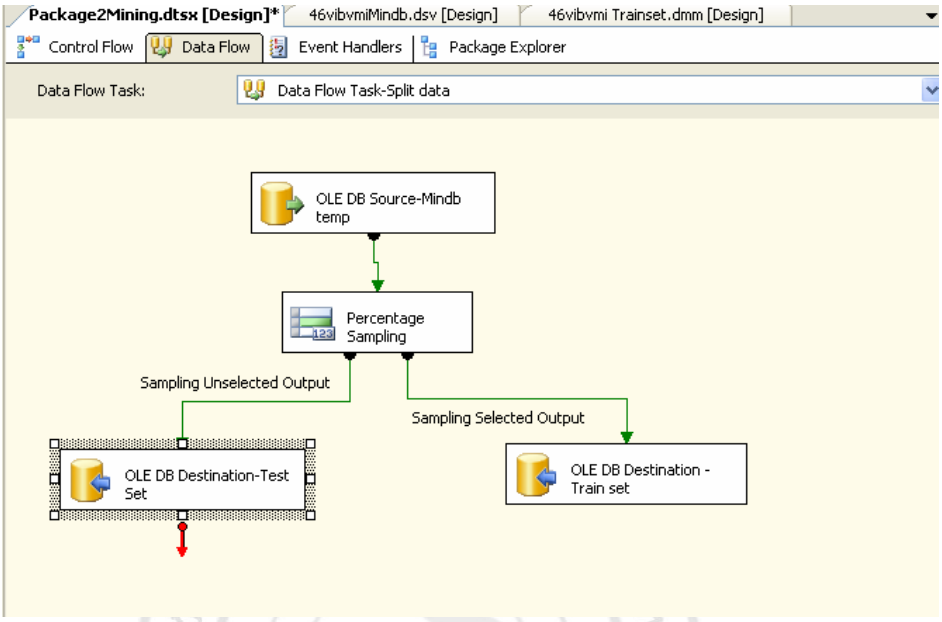
Columns of the classification matrices correspond to actual values; rows correspond to predicted values

Counts for Entropy on [Class Code]:

Predicted	H (Actual)	S (Actual)	M (Actual)
H	0	0	0
S	57186	195738	47229
M	1	4	12

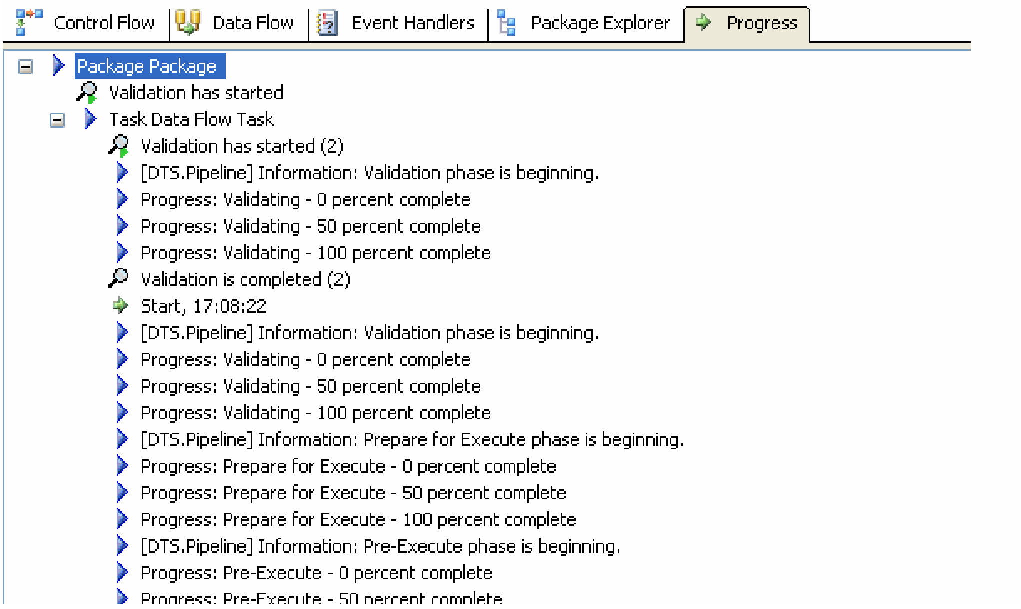
Step 8. Package for risk mining

The Integration Services tools provide the capability to design, create, deploy, and manage packages that address everyday business requirements. Package requirements are extraction transformation and load from data warehousing to destination table that dbo.46vibvmi_trainset and dbo.46vibvmi_testset.


Package2Mining.dtsx [Design]*
46vibvmiMindb.dsv [Design]
46vibvmi Trainset.dmm [Design]
Control Flow
Data Flow
Event Handlers
Package Explorer
OLE DB Source-Mindb temp
Percentage Sampling
Sampling Unselected Output
Sampling Selected Output
OLE DB Destination-Test Set
OLE DB Destination - Train set

Data Flow Task: Data Flow Task-Split data

Run package to load the data into destination. We can verify the status of each step by clicking on the progress tab.



```
Package Package
    Validation has started
  Task Data Flow Task
      Validation has started (2)
      [DTS.Pipeline] Information: Validation phase is beginning.
      Progress: Validating - 0 percent complete
      Progress: Validating - 50 percent complete
      Progress: Validating - 100 percent complete
      Validation is completed (2)
      Start, 17:08:22
      [DTS.Pipeline] Information: Validation phase is beginning.
      Progress: Validating - 0 percent complete
      Progress: Validating - 50 percent complete
      Progress: Validating - 100 percent complete
      [DTS.Pipeline] Information: Prepare for Execute phase is beginning.
      Progress: Prepare for Execute - 0 percent complete
      Progress: Prepare for Execute - 50 percent complete
      Progress: Prepare for Execute - 100 percent complete
      [DTS.Pipeline] Information: Pre-Execute phase is beginning.
      Progress: Pre-Execute - 0 percent complete
      Progress: Pre-Execute - 50 percent complete
```

APPENDIX D

DMX QUERY

This section shows, perform model for Report level risk motor insurance new customer by using DMX queries. This query returns the model's content. In the case of the decision tree model, each row in the query result represents a node in the decision tree. We will apply the model to predict risk level for new customer through the following query:

```
SELECT flattened
  (t.[pol_no]) as [Policy no],
  PredictHistogram([46vibvmi Trainset].[Class Code])
From
  [46vibvmi Trainset]
PREDICTION JOIN
  OPENQUERY([Mindb],
    'SELECT
      [pol_no],
      [lib_key],
      [poltype_key],
      [carname_key],
      [carcode_key],
      [carage_key],
      [discount],
      [class_code],
      [car_gear],
      [ins_occu],
      [car_use],
      [car_kind]
    FROM
      [dbo].[46vibvmi_testset]
    ') AS t
ON
  [46vibvmi Trainset].[Lib Key] = t.[lib_key] AND
  [46vibvmi Trainset].[Poltype Key] = t.[poltype_key] AND
  [46vibvmi Trainset].[Carname Key] = t.[carname_key] AND
  [46vibvmi Trainset].[Carcode Key] = t.[carcode_key] AND
  [46vibvmi Trainset].[Carage Key] = t.[carage_key] AND
  [46vibvmi Trainset].[Discount] = t.[discount] AND
  [46vibvmi Trainset].[Class Code] = t.[class_code] AND
  [46vibvmi Trainset].[Car Gear] = t.[car_gear] AND
  [46vibvmi Trainset].[Ins Occu] = t.[ins_occu] AND
  [46vibvmi Trainset].[Car Use] = t.[car_use] AND
[46vibvmi Trainset].[Car Kind] = t.[car_kind]
```

CURRICULUM VITAE

NAME	Miss Suparat Tuanakog
DATE OF BIRTH	19 September 1975
EDUCATION BACKGROUND	
Bachelor Degree	Bachelor of Computer Science Suan Sunandha Rajabhat University 1998
WORK EXPERIENCE	1995 – Present Senior Programmer in MIS Department. The Viriyah Insurance Company.