19

Banking and Commercial Applications

Objectives:

- The wholesale side of a commercial bank has a significant need to invest in business intelligence and data mining technology because wholesale banking provides a large percentage of a commercial bank's revenue.
- A business intelligence solution, which summarizes this information in the form of query-based reports, augmented by the predictive power of data mining technology, can greatly enhance the corporate decision making process.
- Various groups of analysts at the bank employ data mining software ranging from proprietary solutions (neural networks in stock brokerage, decision trees in credit risk), PC-based single algorithm tools (knowledge Seeker), to SAS on MVS and Intelligent Miner on SP2.
- DM supports a unit view of patterns in data, especially in cases where similar project objectives and data deliver discrepancies result.
- The KDD process used in commercial applications performs data cleaning and data preparation, which has to be done before the actual data mining can take place.
- Decision support is like a symphony of data. It is best to look at decision support from a functional (rather than a tool-based) perspective and to look at available technologies in terms of the functions they provide.
- The model has to be integrated into existing systems, applied on a day-to-day basis, and is shown to meet a target in cost savings.
- To sell products and service performances via Internet in a really permanent and successful manner, this means very special requirements to today's e-commerce platforms.
- The diversification of product and performance portfolios and an improved customer relationship management (CRM) on the basis of asset situations, venturesomeness, payment habits, and consumer behavior will become the driving forces of the use of data mining in e-commerce.

S. Sumathi and S.N. Sivanandam: *Banking and Commercial Applications*, Studies in Computational Intelligence (SCI) **29**, 439–472 (2006)
www.springerlink.com

- Retail data mining can help identify customer buying behaviors, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies, and reduce the cost of business.
- The most innovative retailers of today are those who use business intelligence to gain sustained competitive advantage.

Abstract. The applications of data mining in banking, e-commerce, retail and commercial areas with case studies are illustrated in this section. It discusses the application of data mining in the wholesale banking industry, illustrates some of the associated challengers, and recommends the development of a domain-specific knowledge-encoding tool. Although we focus on wholesale banking, these observations have direct parallels elsewhere, including retail banking, pharmaceutical, and manufacturing industries. Section 19.2 includes a case study based on the distributed data mining. The case study discussed was taken from Global Information Technology, Bank of Montreal, Canada. Another case study Dimensional Systems, Cambridge, MA, is discussed in Section 19.4 based on the decision support systems.

There is a tremendous interest in data mining applications in the commercial world. Many companies, after their initial success with their data mining research and pilot projects, are starting to move these projects into real-world business applications. Today most of the data mining development has come from the research (academic) world. As a result, a lot of business issues have not been addressed and so far they have only received cursory attention. The section highlights some of these issues to the KDD research community so that we could bridge the gap between the research (academic) world and the business world.

Data mining is the "non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data." To sell products and service performances via Internet in a really permanent and successful manner means very special requirements to today's e-commerce platforms. The virtual market in the network of the networks develops faster and, in the future, will be more important than the market of the real world. The offers are more extensive, the purchasing decisions by the customers are made faster, and last but not least: The competitor's product is only a mouse-click click away from the respective Web pages! Section 19.6 outlines the types of data available for mining, describes operation and limitations of the mining algorithms, and discusses the marketing and ethical issues that arise.

Retail data mining can help identify customer buying behaviors, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies, and reduce the cost of business.

Section 19.8 explores the various applications of business intelligence in the retail industry. Business intelligence refers to a host of technologies like data warehousing, online analytical processing (OLAP), and data mining, which seek to turn data into actionable information.

19.1 Bringing Data Mining to the Forefront of Business Intelligence in Wholesale Banking

Although the basic algorithms of data mining technology have been available for many years, data mining has not yet realized its full potential as an integral part of some business intelligence solutions. As in most industries, the success of a data mining implementation as part of a viable business intelligence solution depends primarily on the accessibility of the data, the level of integration of the data mining software to existing databases, the ease of data manipulation, and the degree of "built-in" domain knowledge. Each of these issues merits careful thought and analysis, although the focus here is on the last issue. The wholesale side of a commercial bank has a significant need to invest in business intelligence and data mining technology because wholesale banking it: provides a large percentage of a commercial bank's revenue, thus making the potential return on investment attractive; needs to manage client relationships across a broad spectrum: from a relatively small corporate entity, to a large multinational corporation with potentially dozens of subsidiaries and closely aligned business partners; and is affected by a large number of factors, including macroeconomic factors such as international economic forces, industry-specific trends, the real estate market, and microeconomic factors such as a particular client's economic health and leadership.

A business intelligence solution that summarizes this information in the form of query-based reports, augmented by the predictive power of data mining technology, can greatly enhance the corporate decision-making process. In a typical data mining project, these data are brought together at an appropriate level of generality that describes product usage as well as client information. Although obtaining data from internal sources and external vendors is not difficult, creating an appropriate data set for mining is challenging for many reasons, a few of which are given here:

"Customer" is not easily defined. A large wholesale bank may work with a "parent" company as well as potentially dozens of subsidiaries, each with its own set of legal and financial constraints.

Product cohorts. These can lead to an array of analytical problem because a corporation's usage of one product set may indirectly their use of another product set because of legal or financial structure.

Patterns of product usage are sporadic, cyclical or inter-related. This perennial problem is found in many industries but is particularly exacerbated in wholesale banking because of product cohorts and complex corporate relationships.

Given that this list is not nearly exhaustive, simply the creation of an appropriate mining database is time consuming and challenging. Furthermore, consideration of these factors in the data mining analysis is crucial

for data mining to gain acceptance within wholesale baking marketing channels. Lines of corporate influence, product cohorts, and data quality issues are often known by domain experts but are not directly reflected in the data.

To illustrate, consider a product demand forecasting analysis, which is a typical application of data mining in this industry, where the task is to predict a corporation's demand for a banking product given other product usage information, relevant economic ratios, and other macroeconomic variables. Building a good predictor here requires a clear understanding of the relationship between product groups. A forecasting analysis might indicate the "usage of cash management products imply usage of investment banking products," but this may be information that is implied by bank policies, or is necessary because of other economic considerations. A tool that reflects these internal relationships when delivering results would greatly improve the feasibility of data mining in this industry. The details of such a tool fall outside of the scope of this section, but a few points can be made regarding it. A simple tool may include the following characteristics:

- A straightforward interface to link known concepts together, such as known lines of corporate influence, known product cohorts, etc;
- A data manipulation feature to effectively deal with hierarchical variables. An internal validation of the data mining results against a table of known encoded relationships;
- A graphical display of known relationship, with an overlay of the "predicted" or "discovered" relationships would greatly aid in understanding the model and in the delivery of executive information; and
- A feedback mechanism that gives the analyst the ability to differentially enhance the strength some relationships to enable "what if" analyses.

19.2 Distributed Data Mining Through a Centralized Solution – A Case Study

This is to present a corporate data mining solution, which supports the current and future large-scale analytical needs of most of the Bank of Montreal lines of business. This case study is taken from Global Information Technology, Bank of Montreal, Canada.

19.2.1 Background

Data mining has quickly matured out of isolated, small-scale, PC-based, single algorithm techniques to a robust analytical solution, which utilize a combination of various artificial intelligence algorithms, massively parallel technology, direct both-way access to relational databases and opened systems with published Application Programming Interfaces. In the banking industry, data mining techniques have been accepted by the statisticians' community and

utilized side by side with more traditional statistical modeling techniques. Various groups of analysts at the Bank employ data mining software ranging from proprietary solutions (neural networks in stock brokerage, decision trees in credit risk), PC-based single algorithm tools (Knowledge Seeker), to SAS on MVS and Intelligent Miner on SP2. Among these groups the analytical, technological, and statistical skill sets and expertise vary.

Growing interest in Data Mining technology at the Bank is driven by numerous factors:

- *Strategic/political:*
- Growing awareness of DM (at the executive as well as business analyst levels)
- Recognizing need for elastic knowledge-based decision making – gaining competitive advantage by responding quickly to changes
- Growing competition in the industry
- Internal competitiveness among divisions, departments, and their respective leaders
- Influence of various consulting companies
- Increased demand by the business (analysts, product managers, branch managers) to be granted direct access to the information
- *Practical:*
- Growing volumes of relatively clean data stored in data warehouse, data marts, and operational data stores.
- Limitations in standard analytical/statistical approaches in terms of number of variables considered for a model, their scarcity, cardinality, and the (high) number of categorical variables requiring special treatment.
- Increased demand by analysts to work with large data samples or even full data sets (fraud detection, credit risk, etc.)
- The need to analyze data from several or all lines of business (cross selling, credit risk, product cannibalization, customer behavior).

The experience with the current status quo shows that satisfying the above needs locally and on ad hoc bases is neither practically manageable nor cost effective. Therefore, at the Bank they are in the process of implementing a robust *Centralized Data Mining Solution (CDMS)* supported by the creation of *Data Mining Center of Excellence*, an institution responsible for managing the new Hardware/Software (HW/SW) and its utilization by various DM groups, providing data transformation, managing DM metadata, and deploying precanned models to light users via the Intranet.

Benefits of CDMS: The CDMS provides high-speed links/gateways to the major data sources deployed in the organization (bank information warehouse, customer knowledge data mart, credit card, risk management, etc.) totaling more than 3TB of data. The likely power user groups are: database marketing, credit risk, credit card, transfer pricing, and others. These groups will be freed from tedious data transformation and HW/SW maintenance responsibilities.

It is the nature of data mining projects to require initially large space to store data. Contrary to what is said, in the data marts the storage requirements are limited to the initial data crunching and the life of the project. Therefore most of the storage could be freed in several weeks and made available to other projects. Equipped with massively parallel processing and sharing almost 0.5TB of DASD users will benefit from unprecedented processing power allowing them to run on large data sets.

Benefits in Costs: The data mining technology is expensive to acquire (SW and dedicated HW) and difficult to maintain (operating system, underlying database, data mining SW itself, gateways, front ends, etc.). And let us emphasize the support of various high-speed links and gateways, data marts, and other business critical sources of data. It is hard to imagine a LOB being able to justify a purchase of top-notch data mining technology, a solution that could cost an initial $600,000 of investment and $300,000 a year for support and maintenance. Such a system would likely be underutilized by the LOB most of the time.

Obviously, the CDMS is highly scaleable and in the long term is a cost-efficient solution – a win-win for all groups involved and for the organization as a whole. For its growth this point and corporate-managed data mining effort give additional power in negotiations with HW/SW vendors.

Data Mining Meta data: One important responsibility of the Data Mining Center of Excellence, which manages the CDMS, is the creation and management of corporate data mining metadata. This is a relational database outlining all data mining projects, models built, their frequency, contacts to creators, data sources involved, variables, and a models' version control. Details about particular model parameter settings, treatment of NULLs, outliers, etc. will be included. The metadata will be published on the Intranet and accessible by all Bank employees.

This solution will help in interpretation of DM results, especially in cases where similar project objectives and data deliver discrepancies in results. It supports a **unit view of patterns** in data.

19.3 Data Mining in Commercial Applications

19.3.1 Data Cleaning and Data Preparation

An important component of the KDD process is data cleaning and data preparation, which has to be done before the actual data mining can take place. The current state of art of data cleaning is far from being as sophisticated as, say data mining algorithms. Most of the data cleaning is done laboriously be humans. There are few general and automated tools that assist in this process.

The idea is to use some of the data mining techniques in data cleaning. We envision data cleaning to be performed in several stages, ranging from

application of simple techniques to increasingly more sophisticated methods. For example, one can use visualization techniques to look for missing data, imperfect data, etc. Or, one can use more sophisticated clustering tools to look for outliers, which may indicate potentially erroneous data. Statistical methods that look for missing data need to be studied in this context. Of course, the entire process will be an iterative one, with periodic evaluation by humans and the reapplication of appropriate cleaning tools.

Let us take an example. Consider a database maintained by a county office. This database contains information on every piece of real estate in the county. Such information might include the address, price owner, size, etc. Such databases may contain lots of errors, missing data, etc. Initially, we could use visualization techniques to locate each house on a map of the county. If a high-priced house appears in a low-income neighborhood, then that is possibly an error. Secondly, we could use multidimensional clustering techniques to look for "outliers" in clusters, which might point to erroneous data. If we have much prior knowledge about the data, we can assume a prior distribution that models the database and use Bayesian methods to compute posterior distributions. Then, we could selectively omit certain data items and re-compute posterior distributions again, thereby allowing us to detect outliers.

19.3.2 Involving Business Users in the KDD Process

Successful data mining applications in business world would require constant interaction and feedback between different users and the data mining process.

User interaction, in different aspects of the KDD process:

Previous knowledge:- The business user may have some underlying (but incomplete) knowledge about the data to be mined. Incorporating this knowledge into the data mining process will allow the data mining process to be more efficient.

On the other hand, the data mining process can also be used as a check against the knowledge that was supplied. For instance, mining may contradict some conventional wisdom. The system should allow for that. Thus we should provide methods to allow data mining algorithms to provide reasonably good results if there is not enough resources for a full-scale job.

Resource constraints:- The need to be able to mine information from a huge resource in real-time (or near real-time) is increasing. The demand of the real world does not allow the luxury of time.

Intermediate feedback:- Each step of the KDD process should provide some intermediate results. For example, results of the data cleaning process, decision of which algorithm to use, etc. The system should be able to provide appropriate feedback to appropriate personnel (users) so that they can make decisions to guide the data mining process.

19.3.3 Business Challenges for the KDD Process

The business issues are many, such as scalability, integration of current systems, data visualization, need for database system support, incremental processing, flexibility and so on. The important issues to be concentrated are:

Data Cleaning and Data Preparation

The research issues are many. How does one design algorithms that can work for heterogeneous datasets? Do the algorithms scale well? Can we perform automatic error "correction" in addition to just error detection?

Knowledge representation and incorporation

How to represent the underlying knowledge that is known? How to incorporate this knowledge into the appropriate KDD algorithm? How to discern whether the underlying knowledge is useful or counterproductive?

Algorithmic issues

To devise algorithms that would take resource constraints into account; to provide algorithms that can provide good intermediate results even if severe resource constraints are imposed. There has been some work on constraint algorithms and incremental algorithms. We need to devise method to incorporate them into the KDD process.

Data visualization and feedback

To provide effective methods for the data mining process and to provide feedback understandable by the user. Different kinds of users may require different kind of visualization techniques. Also, effective means of allowing user feedback to the data mining system need to be devised, ideally coupling these to the data visualization process.

19.4 Decision Support Systems – Case Study

Decision support is like a symphony of data. In the same way we need a variety of musical instruments functioning together for the purpose of playing a symphony, we need a variety of software tools working together for the purpose of doing decision support. It is best to look at decision support from a functional (rather than a tool-based) perspective and to look at available technologies in terms of the functions they provide. In other words, tool features get mapped to DSS (Decision Support System) functions. This case study is taken from Dimensional Systems, Cambridge, MA.

19.4.1 A Functional Perspective

Basic DSS functions

DSS is about synthesizing useful knowledge from large data sets. It is about integration, summarization, and abstraction as well as ratios, trends, and allocations. It is about comparing database generalizations with model-based assumptions and reconciling them when they are different. It is about good, data-facilitated creative thinking and the monitoring of those creative ideas that were implemented. It is about using all types of data wisely and understanding how derived data was calculated. It is about continuously learning and modifying goals and working assumptions based on data-driven models and experience. In short, decision support should function like a virtuous cycle of decision-making improvement. Let us laundry list these concepts to identify the minimum set of basic functions that comprise any DSS framework:

- Data collection,
- Data storage and access organization,
- Dimensional structuring,
- Data synthesis,
- Intuitive representations and access models,
- Predictive models,
- Model verification
- Knowledge sharing,
- Resource allocation strategies,
- Scenario analysis,
- Belief conflict resolution,
- Prescriptions and
- Decision implementation capabilities.

A Cognitive Metaphor for DSS: Beyond Closed Loop Systems

The best metaphor that one can think of for understanding how all these decision support functions fit together is a cognitive one. In contrast, earlier metaphors focused on the unidirectional flow of information from "raw" data to synthesized knowledge. Second-generation metaphors, currently in vogue, focus on bidirectional, closed loop systems wherein the results of DSS analysis are fed back into production systems. The hallmark of a third-generation cognitive metaphor is the interplay of two separate information loops. The first is akin to the closed loop system and one would characterize it as a data-driven loop. But in addition to that loop there exists an inner loop where data driven information meets model-driven goals and beliefs at the moment of decision. Although that inner loop is frequently provided by a living breathing person, it is a function that needs to take place and in automated systems needs to take place in the form of software within the overall decision support

system. AI workers have known for a long time that it takes a combination of data-driven and model-driven information to produce high-quality decisions.

Using a cognitive metaphor, the universe of DSS functions is composed of five distinct functional layers within which the two above-mentioned information loops interact: a sensory/motor layer, a primary memory layer, a data-based interpretive and understanding layer, a decision layer, and a model-driven layer of goals and beliefs.

Data-Driven Understanding

From 20,000 feet, data-driven understanding is the process of synthesizing knowledge from large disparate data sets. *Understanding* is a loaded term, so let us break it down into smaller chunks. The key components of understanding are describing, explaining, and predicting. The main obstacles to understanding are lack of tool integration, missing, meaningless, and uncertain data and lack of verification capabilities.

Descriptions form the basis. Examples include "The Cambridge store sold 500 pairs of shoes last week," "Our corporation did 35 million last year," "The Boston stores paid an average of 36 dollars per foot in 1997" and "Boston rent is twice as expensive as the rent in Portland, Maine." Descriptions are more than just measurements. Descriptive processes take whatever raw measurements there are, and through aggregations, ratios, and other certainty-preserving operations, creating a fleshed out multilevel multidimensional description.

In practice, there may be a variety of inferential techniques employed to arrive at a descriptive model of a business or organization. For example, inferential techniques may be used to guess what values may apply to what are otherwise missing cells. When there are a lot of blanks that need to be filled in for a descriptive model the process is akin to data archeology. Good data archeology requires the close integration of OLAP and data mining or statistics tools.

Explanatory modeling starts where descriptive modeling left off. Explanatory models are representations of relationships between descriptions. Such statements as "For every increase of 1% in the prime rate, housing sales decrease by 2%" represent explanations or relationships inferred from descriptions of both housing sales and interest rates. The functionality provided by statistics and data mining tools of all varieties belong in this arena. Regressions (the mother of all analyses), decision trees, neural nets, association rules, and clustering algorithms are examples of explanatory modeling.

Predictive modeling is just an extension of explanatory modeling. One cannot make a prediction without having at least one relationship that we are banking on. And while most all-mining activities aim at building predictive models, the key algorithms are in the discovery of the patterns. Predictions are just the extension of some pattern already discovered. That is why all

the data mining algorithm buzzwords, we hear are about pattern discovery techniques, not pattern extension.

OLAP tools do not provide for explanatory or predictive modeling. Data mining does not provide for dimensional structuring. Yet, it is best to perform data mining within an OLAP (multilevel, multidimensional) environment. For example, to design a new promotional campaign based on point-of-sale POS and demographic data we might use

- An OLAP tool to aggregate SKU-level data to product categories,
- The OLAP environment for exploratory analysis and new variable creation,
- Clustering to discover natural segments in the POS data,
- Visualization to interpret the clusters,
- OLAP to incorporate the clusters as higher level aggregation levels,
- More directed mining on the cluster-aggregated data,
- Visualization to interpret the mining results, and
- OLAP to further aggregate the mining-based predictions

all in order to support the data-based brainstorming for a new promotion campaign. In short, we needed to use a combination of OLAP, data mining, and visualization to accomplish a single BI task: promotion development. We can call this kind of integration **DSS fusion**.

The market as a whole is beginning to move in this direction. A number of OLAP companies are adding or claiming to add data mining capabilities, although not all of them are fully integrated with their OLAP products. For example, Holos is adding mining capabilities, Cognos has a simple mining application, MIS AG has a mining application, as does Pilot Software (D&B, Data Intelligence Group). We believe it will be easier for OLAP companies to add data mining capabilities than it will be for data mining companies to add OLAP capabilities.

Although it is good to see so many OLAP vendors offering mining capabilities, these capabilities still need to be better integrated. Mining functions should be as simple to invoke as ratios. It should be possible to perform data transformations from within an OLAP/mining environment. And mining should be fully integrated within the dimensional structure meaning that operations like drill down from an interface to the results of an association rule algorithm should work, and depending on how things were defined either return a set of associations already calculated for lower down in perhaps a product hierarchy, or trigger the calculation of such associations. Thus the same thinking that goes into an OLAP design of what should be precalculated and what can be calculated on demand can apply to data mining as well.

Missing, meaningless, and uncertain data are frequently present in data sets and pose a significant hurdle to understanding. Missing and meaningless data are logically distinct, both need to be distinguished from the value zero (frequently the default value assigned an empty cell), and need to be differentially processed. For example, in an averaging function, where empty cells

denoted *missing*, we would need to assign some kind of default value to the empty cells for the purposes of aggregation. In contrast, if the empty cells denoted *meaningless*, we could not assign a default value. Most OLAP and data mining tools lack good empty cell handling techniques.

Unlike missing and meaningless data, uncertain data is present as a data value in a data set. For example, a statement like "We predict that our new brand of ice cream will capture 5% of the market for ice cream products," needs to qualify the uncertainty associated with the estimate of 5%. Typical statistical measures of uncertainty include the variance and bias associated with an estimate. The overall picture of uncertainty can get a little more complicated as derived measures follow from business rules, which have their own sources of uncertainty. For example high-level sales forecasts based on aggregating lower level predicted sales data need to carry forward the uncertainties derived from the predictive models through the aggregation process. What is more, the predictive models themselves may rely on certain "rules of thumb" for their forecasting logic. As more assumptions become embedded in business data, OLAP tools especially will need to provide ways to process uncertainty.

Finally, in the same way as an astronaut's working environment is composed of fabricated living elements (temperature, pressure, oxygen, etc.), a DSS end user/analyst's working environment is composed of fabricated data elements (daily summaries, weekly aggregates, brand reports, and so on). Given the degree to which end users are dependent on derived data as their own inputs, it is crucial that DSS vendors provide better verification capabilities.

19.4.2 Decisions

Deciding is the process or function of combining goals and predictive models. To decide that prices need to be lowered on certain products is the result of a goal to maximize sales and a predictive model that relates sales to product price. To decide that a certain loan applicant should be denied credit is the result of a goal to minimize loan write-offs and a predictive model that relates certain loan applicant attributes with the likelihood of loan default.

If there were no goals, it would be impossible to decide what course of action to take as any action would be as acceptable as any other. Without the goal of maximizing sales, for example, there is no right decision concerning product pricing. And without a predictive model equating product prices with product sales there is no way to know which decision will be most likely to maximize sales.

Decision-making challenges may arise from

- the need to automate certain decision-making functions,
- the need to ensure consistent decisions,

- difficulties analyzing how a decision was made,
- complexities in the predictive models,
- difficulties interpreting stated goals,
- instability in the goals themselves,
- interpersonal dynamics,
- fluctuations in the predictive models, and
- conflicts between data-driven and model-driven beliefs.

Business rule automation tools focus on the first two challenges. Decision analysis tools focus on challenges 3–6. Group decision support tools focus on challenge 7. Challenges 8 and 9 lie a little further down the road.

Looking ahead, we will start to see self-modifying rule systems that continuously monitor the world to see if it behaved like predicted, and when it does not, then changing the predictive models it used to make rules. In the process, systems could try out different scenarios or predictive models and analyze how well the system would have fared under each scenario. We would also like to see rules bases connect to OLAP tools wherein the rules base was the source of cost allocation rules used in the OLAP system. Although OLAP tools provide a sophisticated calculation environment they would benefit from an organized method of defining and managing rules.

We should also be able to deduce rules given goals and predictive models, which brings to the next major category of decision-making software, decision analysis software. The need for decision analysis software kicks in where decisions are based on multiple predictive models with complex measures of uncertainty and where the goals themselves are variable. Typically this appeals to higher up in an organization. Decision analysis is closely related to operations research where there are several mutually exclusive goals and shared scarce resources and the trick is to maximize a global property like profit, stability, or happiness. In summary,

- The BI or DSS space is best looked at from a cognitively based functional approach rather than any tool-centric perspective,
- The center of decision support is at the intersection of data-driven understanding and model-driven goals and beliefs,
- Data collection and storage should be driven by decision-making needs,
- All media should be organized through an integrated semantic model of the enterprise,
- Data mining, OLAP, and data visualization can and ought to work together for the majority of decision support problems,
- The analytical knowledge required for interpreting complex derived data should be disseminated to end user/analysts,
- Descriptive and explanatory modeling should begin leveraging textual data,
- Model-based decision-influencing beliefs should be captured electronically and compared with data-driven versions of the same beliefs,

- Procedures for resolving conflict between data and/or model driven beliefs should be public, and
- Everything should be verifiable.

19.5 Keys to the Commercial Success of Data Mining – Case Studies

Some techniques to the commercial success of data mining are presented based on (i) The experience at BT Laboratories, Data Mining Group, Ipswich and (ii) A service provider's view taken from a global multidisciplinary professional services firm – Arthur Anderson, Zurich, Switzerland.

19.5.1 Case Study 1: Commercial Success Criteria

Success criteria are much more commercially oriented than they were in the early days. It is no longer enough to produce a convincing demonstration system or model. It has to be integrated into existing systems, applied on a day-to-day basis, and be shown to meet a target in cost savings. Defining and estimating the costs and benefits of a proposed project can be difficult, but we have found that the more effort that is spent on this early on, the better.

Reasons for Failure

We have carried out many data mining projects over the years. In our experience, the key to a successful data mining project is not obtaining some data and finding a useful pattern in it. Only a small number of our projects have been unsuccessful in the sense that we could not find anything useful in the data. More usual reasons for failure are:

- inadequately defined objectives,
- an inadequately thought-out exploitation route for results;
- external factors such as changes in the business environment that make the objectives of the project no longer relevant, and re-organizations.

To avoid these problems, we spent a lot of effort in the definition and planning phase of data mining project and have produced our own data mining project guidelines, in conjunction with Syntegra, BT's systems integration arm, for the company to follow.

A commercial data mining application takes more skills and people than just the data miners; the successful projects are those in which data mining is just seen as a component of an overall project or system. This tends to focus the objectives and deliverables of the data mining aspect, and avoids it being overhyped, if that is possible!

Data Extraction and Preprocessing

We still experience delays and difficulties in obtaining the necessary data for projects, particularly if the data is not in a data warehouse. And a lot of effort is spent preprocessing the data prior to analysis. This is not, however, a severe problem if planned for.

Our Rule

We have found that the role of the group has changed in recent years. Up until about a year ago, most of our nonresearch work was data analysis, i.e., doing data mining for another part of the company. We still do a lot of this, but often we find ourselves acting as consultants, advising other parts of the company about data mining. For example:

- Will data mining alone solve our problem?
- What tools do we need?
- What skills do we need?
- Which tool or supplier should we go with?

This changing role is our response to the increase in demand for data mining and also the increasing choice of suppliers and tools.

Areas for Improvement in DM Software

The suggestions for improvements to existing DM software concur with those cited in the literature, for example: integration with relational databases, scalability, etc. Obviously, the onus is not just on the data mining tool developers to achieve integration, but also major database vendors. Databases should support operations often used in data mining efficiently, for example, random sampling.

Several tools suppliers have developed data mining tools, which, it is claimed, can be used by nonspecialists, often via an easy-to-use interface or data visualization. While this is certainly an improvement, we believe such tools will not find a large market of nonspecialists. This is because nonspecialists do not want to have data mining made easy for them – they really do not want to do data mining at all in a general context – what they want is to solve their particular problem, be it, targeting their marketing, highlighting potential fraudsters, etc. The likely outcome will be the development of "vertical" applications for common business problems in particular industry sectors, in which the data mining element is largely hidden from the user. Examples include systems designed to highlight customers likely to leave a mobile telecommunication company for the competition ("churners"), telephony fraud detection, and targeting personalized adverts at Internet users.

Problem Formation

Problem formulation, along with risk assessment and project planning, is very important. We spend a lot of time upfront with the customers of our work working out as well as we can exactly what the business benefit is hoped to be, and how it will be realized.

DM Lifecycle

We have produced our own guidelines for managing data mining projects, called M3, for use by BT. It uses a DM lifecycle similar to those in the literature in Fayyad 1996, but places emphasis on the early stages of a lifecycle: particularly problem formation, data investigation, risk assessment, feasibility check, cost-benefit analysis, project planning, role identification, and defining a clear exploitation route. This can act as a checklist for the analyst where all the relevant aspects have been covered and provide some tips on pitfalls to watch out for. For example, the role identification part of M3 defines several roles that may have to be involved in, and kept committed to, a successful data mining project: the analyst himself/herself, the domain expert, the database designer/administrator, the customer, the end user, the legal expert, the system developer, and the data subjects themselves.

19.5.2 Case Study 2: A Service Provider's View

Arthur Andersen is a global multidisciplinary professional services firm that helps its clients improve their business performance through assurance and business advisory services; business consulting; economic and financial consulting; and tax, legal, and business advisory services. Arthur Andersen Business Consulting in Switzerland offers services and provides implementations in the areas among others of cost management, revenue enhancement, activity-based management, performance management, knowledge management, transaction systems, collaborative systems, data warehousing, and OLAP and data mining technologies.

Since we are both a user of data mining software and provider of data mining solutions, our view includes both aspects: what are the benefits and application areas of data mining technology as well as directions for improvement for data mining software.

- *Interfaces.* As for any information processing system, it is important to get data in and out of the system. The advantage that data mining applications have is that they are usually built on top of a data warehouse. Hence, they do not have to provide access to a variety of operational systems, but just to a database directly. This is to be preferred due to the massive amounts of data that need to be transferred. However, the information chain does not stop with the data mining application. It is merely an

intermediate element. The information that comes out of such a system has to go either into an operational system, e.g., a mail targeting or a customer service application, or needs to be passed on to decision takers as part of a knowledge management framework. The amount of data could be small – in the case that only the model and some related information is needed – but more often it will larger – if customer records are to be augmented with credibility scores or customer segment information. In the latter case, this data could be written back to the data warehouse and from there sent to the operational system. By doing this, one could avoid building special interfaces for the various operational systems. Most data mining applications are also not intended for knowledge sharing in contrast to collaborative systems such as electronic mail and Web pages or management information systems. Hence, models that have been created in a data mining process such as decision trees or scorecards need to be further processed electronically to make this information accessible and usable.

- *Model Scoring.* Everyone knows about the things that can happen when one tries to learn a model from large datasets and that can decrease the quality of the model or even make it useless, like, for instance, over learning. Especially, when we try to anticipate the customer's future behavior, e.g., in churn management or credit scoring, we need to know about the quality of the model. Some data mining tools that originate in statistical programs allow model scoring, whereas tools that are more application driven and are geared toward the end user just offer the model at the end. Another point is that the need for a data mining applications originated in a business problem, which was then converted into a data-mining task. Hence, the score needs to be translated back into terms of the original business problem. For example, we need to know about the costs of the errors we will make due to the quality of the model. Data mining works on historical data; but usually new data is constantly flowing into the databases. Rebuilding the model and rescoring the model are two different tasks with different resource consumption. Model scoring could help us to decide when to rebuild the model.
- *Dimensions with Interrelated Values.* Some dimensions are not just a collection of values. First of all, there is the time dimension. The fact that the values in this dimension represent points in time puts them into a certain order and augments them with an adjacency relation. This gives them a completely different meaning compared to unrelated values like customer IDs, company departments, or products. But time is not the only dimension where we have additional information about its values. The geographical location is another dimension where values are linked by an adjacency relationship. We know that certain states are neighbors. This does not give them an order like it is the case in the time dimension but this relationship could help in the explanation of certain phenomena. We also know that countries are composed of smaller administrative districts

and in turn are located in one of the five continents. This knowledge forms a hierarchy on top of the values. Likewise, we put customers into customer segments and products into product groups. Sometimes we could treat this information as an additional attribute, however, this might not always be feasible, especially, if the hierarchies do not have clearly identifiable levels. But this hierarchy puts certain values into a closer relationship than others and it would only make sense to make use of this information.

- *Operational Use vs. Ad hoc Analysis.* We see data mining software as the extension of current decision support systems, such as reporting tools and OLAP technology. We gain more power in two directions: (1) the range of questions we can ask the system to answer and (2) the amount of data that can be processed. Reporting software, especially one that produces paper, is mainly used for regularly, e.g., monthly, quarterly, or yearly, standard reports. There are usually only a few different types of reports and they contain mostly too many details for one user and too few for another. If such a report gives rise to a question that it cannot answer then there is need for an ad hoc analysis, which almost always means calling up someone in IT and have them design and print a special report. OLAP technology tries to cover both: regularly standard reports and ad hoc queries. By providing drill-up and -down as well as drill-through functionality it can serve the users with the right level of detail. Starting from such an electronic report, the user can furthermore adapt it by slicing and dicing to view the data in different ways and perform an ad hoc analysis. We believe that data mining will be used in much like the same ways. There will be standard applications (runs of data mining algorithms and presentations of the results) that will be scheduled regularly. This might be a specific task, which has been designed in a collaborative effort by a business user and a statistician. The configurations and sets of parameters need to be stored. We would like to have a user interface where most if not all of the configuration complexity can be hidden. Another request might be to compare the results of a sequence of such regularly performed applications and to look at changes over time, like for instance review the changes of the criteria for grouping customers into segments. On the other hand, there remains the need for ad hoc analyses. There might be questions arising from standard applications or we need to prepare and monitor the results of nonregular actions. Data mining software will have to provide the means to satisfy both needs.
- *Business Templates and Automated Model Selection.* After having been occupied with implementing all the necessary technology for visualizing multidimensional data and building user interfaces, OLAP software providers are now taking the next step and delivering templates and shells in order to reduce the amount of work required to build a solution to a specific business problem from scratch using a very generic tool. This idea tries to leverage the underlying commonalties of certain classes of business problems and to shift work being done several times by several end users or by

service providers to work being done once by the software producer. Also, now that much of the technology is implemented in all of the different OLAP tools, providing business templates could be a market differentiator for a software company. With suppliers of data mining software trying to add on more and more technology, they still have the focus on the technology as opposed to the focus on the business problem to solve. However, we need specific solutions to specific problems. In addition to the problems of software and user interfaces – something that data mining applications share with OLAP tools – one also needs to select the right algorithm and parameters or the right model from a set of runs of potentially different algorithms. Focusing on a particular business application helps in reducing the interface problem to few sources and target operational systems. By drawing on the experiences of business users of the software and combining them with the expertise that was needed to implement the algorithms for building business templates, software vendors could help reduce costs and efforts of implementing data mining. It could not only differentiate them from other tools that are built on the same technology but also decrease the likelihood of a user blaming the vendor for failures while using its software.

- *Business Applications.* One of the requirements for OLAP software is the ability to build business applications with a number of pages of different standardized views of maybe different data, while hiding most of the complexity of the underlying technology. These pages include settings of which dimensions of the data to show at what level of detail, which have been determined by, for instance, business analysts. This facilitates the use of the software for nontechnical end users and especially for operational, regularly, standardized usage. In our opinion, data mining technology will go in a similar direction. Configurations such as the selection of the algorithm, certain parameters, and the form of representing the results could be developed through the collaboration of business users and statisticians, leaving only a few knobs if at all for the standard user. The page could be extended by other user interface elements such as push buttons or sliders to provide access to preset configurations or standardized views of the results. Being able to build business applications with a user interface and functionality adapted to the end user is, from our point of view, one of the prerequisites to make data mining technology more accessible to a broader range of business users. One wish would also be that data mining software would lay a common ground for communication between business users and statisticians. For instance, being able to demonstrate the effects of algorithm or parameter selections directly on the user's data or, on the other hand, to collectively interpret the results might help the understanding of each other's problems. Also by providing means to translate the business problem into a statistical one and the statistical answer back into a business one, the software could facilitate both communities while building relationship between elements of the business and the statistical world.

- *New Application Areas.* Most of data mining technology is applied to customer data. One of its characteristics is that it is usually stored in a relational database and that there are a huge number of records. However, since in the company we also help the customers in the areas of performance management and activity-based management, we were wondering whether this could also be a possible application area. What if there are a large number of accounts, cost centers, or activities to be analyzed? We think that even a hundred different positions might be too much to be dealt with manually. One could, for instance, look at drastic changes over time, which probably no one would do for one hundred departments. Furthermore, we think that most of the ideas that have been developed regarding customers can also be applied to employees: employee satisfaction instead of customer satisfaction, employee retention instead of customer retention, employee profitability instead of customer profitability. Companies developed bonus programs, as one example, to implement these ideas. With possibly a large number of employees, this would open up a new application area.

Finally, we would propose performing data mining on multidimensional data. Here, the problem is less as regards the number of items per dimension but more with regard to the number of dimensions. With six of seven dimensions and only ten items per dimensions we already have one to ten million data elements. Let us say, we see a drop in revenue. Finding out which product in which configuration in which packaging sold in which location through which channel was mostly responsible for this observation might not be easy and could be time consuming. Furthermore, by only looking at the data on the top level, the business user might not even notice this event, which he should probably act upon at all because it is cancelled out by other data it is aggregated with. Performing multidimensional data mining could again be combined with business knowledge about the different domain that multidimensional OLAP technology has been applied to.

19.6 Data Mining Supports E-Commerce

Many users of the Internet are aware that each time they connect to an on-line shopping server, they leave behind a "footprint" in the site's server logs. The information contained in this footprint is innocuous, but it can be "mined." Data mining is the "non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data." This section outlines the types of data available for mining, describes operation and limitations of the mining algorithms, and discusses the marketing and ethical issues that arise.

A relatively high product and offer equality of the different offers, a decreasing customer loyalty, a high margin pressure, and the drastically

increasing globalization are further factors for the fact that the market chances become more and more narrow.

Web stores that want to guarantee their operators a substantial advantage over their competitors must be able to efficiently support an individualization of the mass marketing so far usual. This individualization and personification in the marketing will become the decisive instrument in the fight for the customers at the beginning of the new millennium.

The more exactly an offer, a price, a bonus, or a tariff is tailored to the customer and the more precise the usage habits, the consumer requirements, the reaction patterns, and the behavioral features of a customer can be predicted, the more positive his purchasing reaction in the store will be. The "tailored" courting and the systematic elimination of "consumer inhibition thresholds" on the basis of previously created customer profiles will open a completely new dimension in the Web marketing. The diversification of product and performance portfolios and an improved customer relationship management (CRM) on the basis of asset situations, venturesomeness, payment habits, and consumer behavior will become the driving forces of this new development.

As no other branch of the business else, the e-commerce is predestined to create the "glassy customer." As with the field of credit and customer card purchases, transaction histories can be examined. The analysis results then flow back to the marketing and form the basis of a selective micro marketing with an enormous degree of efficiency. Through "enriching" the collected data stocks with socio-graphical and microgeographical information (e.g., social background of residential area), the shop operator will allow the shop operator in the future to obtain even more sales-relevant knowledge on his respective customers.

Up selling and cross selling potentials (for example: a person who purchases A will also purchase B in 70% of all cases) can be discovered by means of basket analyses and can be utilized in sales-increasing manner. The hit quote on target-group-optimized offering actions (phone call, mail, or e-mail) can be improved substantially and, thus, unnecessary and cost-intensive "scattering losses" (e.g., stressed customers leave) can be avoided. With that, the "marketing shotgun" belongs to the past. The customer's binding to the product and to the offering enterprise is improved while the sales/distribution costs are reduced. Leaving-endangered customers or customers who have already left can be won back, and new customers can be counted more efficiently. The improvement of sales/traffic ratio must be the objectives of tomorrow's e-commerce solutions. With its data mining solutions, the Prudential System Software is already able to contribute with an essential step in this direction!

19.6.1 Data Mining Application Possibilities in Web Stores

The application possibilities of the data mining in Web stores become particularly clear from those persons' point of view who operate Web stores each day. Just at those points where many decisions are presently still made manually,

a decision-supporting data mining system developed by Prudential Systems can contribute to an enormous value increase of Web stores. On principle, two types of administrators operate Web stores: by system administrators and by business administrators.

The task of the system administrator is to make the basic settings immediately after a store has been installed and to change them later, to create the design of the store, to release the store front and, possibly, to block it, to technically maintain the business administration as well as to do database manipulations.

Business administrators are engaged with the actual business. They process orders; supervise the inventory; order with suppliers; administer customer account data; change product offers and descriptions; offer discounts for selected customers; define customer categories; and cause products to be summarized in lists, subcategories, categories, and main catalogs. Moreover, they react to selected, but not ordered baskets with special offers or a new product structure.

During his work, a business administrator has to make many decisions. However, the data that would be necessary for these decisions are present in a form that is not suitable for him. The data mining can fill this gap by the fact that exactly those decisions are automatically supported, which the business administrator presently still has to make manually or cannot make at all.

In the store, there is essentially distinguished between customers and products. Products can be summarized in lists, subcategories, categories, and main catalogs. The summary influences the presentation of the store and, thus, is relevant for the touch-and-buy behavior of the customers with respect to these products. However it is made by the business administrator without the possibility to actually consider the behavior of the customers because he does not have the needed information. Here analyses of the touch-and-buy behavior can give important hints on the classification.

Moreover the business administrator classifies the customers to form customer categories, and customers can assign themselves to profiles. The assignments are always made manually without evaluating the behavior of the customers. So, for example, discounts are allocated manually, and the customers manually determine those catalogs and products that they want to view.

In a big store with many customers, it is surely impossible for the business administrator to perform the assignments completely or even optimally. Moreover, only a few customers will utilize the possibility to assign themselves to profiles or they select the wrong profile. Consequently it is required to cause this assignment process (i.e., the individualization of the store) to be carried out automatically. The data mining is the philosophy that can and will realize this. The PrudSys ECOMMINER developed by Prudential Systems is the tool that puts life into this philosophy!

Letting Business Users Loose: In the commercial world, the term *data mining* has become associated with large, multihundred-thousand dollar projects

taking several months and requiring the skills of PhD data miners with years of experience. A business manager wanting to use data mining to gain some competitive advantage might embark upon a large project and work with skilled data miners. But this can be cumbersome: large projects require many rounds of review, assessment of the set of vendors and consultants who offer products and services, project scooping, possibly buying new hardware, working with their IT department, etc. It is quite an obstacle. If this situation is one point on a spectrum, another point at the opposite end would be where applications already exist within the company that would allow a business person to mine their easily. The business manager might just point her Web browser at the site that contains the data and applications for analysis, look around for a while and discover, perhaps, that the jump in her group's expenses this quarter is primarily attributable to travel and hiring expenses, or that the majority of her lost customers were affluent single men who lived close to her competitor's store.

Giving power tools to novice business users can be a recipe for disaster – what if they are pointing the tools at data they do not fully understand? What if they build a predictive model and accidentally include input variables that were not all known before the action the model suggests needs to be taken? What if their data is not clean? And so forth. An experienced data miner is someone who has learned to use considerable caution and knows how to avoid pitfalls and traps.

But does this mean we cannot give "data mining" applications to business users? What if instead of just putting a nice GUI on top of a powerful data mining tool like a neural network and calling it easy-to-use, we built simpler analysis methods, whose results were easy to understand? What if the results were phrased as statements about the existing data, not predictions about future data that could be misconstrued. What if, instead of a raw tool that could be turned on any data they happened to have lying around, instead they had applications that were integrated with a data mart that was constantly updated with sales and other data sources, and analyses were restricted to use only data that was known to be clean? These "analysis methods" might look preposterously simple to expert data miners, yet they might be very useful to business users.

A robotics professor once asked his class to design a robot to wash dishes. The inventions that were turned in stretched the technology to its limits: microsensors and force feedback control to avoid breaking the dishes, vision systems to see the dirt, and joints with many degrees of freedom to allow reaching into the sink, wiping the dish, and setting it into the drying rack. The next day, the professor pointed out that dishwashing machines already exist, are relatively cheap to manufacture, and work quite well.

Large data mining projects certainly have their place – in some cases small improvements in a model can result in huge savings, making the investment easily worthwhile. In other cases, a large project may not be explored due to lack of resources. If no applications exist to allow an executive or manager to

look at their business and their data, there may be a huge missed opportunity cost from their lack of understanding of their business.

19.7 Data Mining for the Retail Industry

The retail industry is a major application area for data mining since it collects huge amounts of data on sales, customer shopping history, goods transportation, consumption and service records, and so on. The quantity of data collected continues to expand rapidly, especially due to the increasing ease, availability, and popularity of business conducted on the Web, or e-commerce. Today, many stores also have Web sites where customers can make purchases on-line. Some businesses, such as Amazon.com, exist solely on-line, without any bricks-and-mortar (i.e., physical) store locations. Retails data provide a rich source for data mining.

A few examples of data mining in the retail industry are outlined as follows.

Design and construction of data warehouses based on the benefits of data mining: Since retail data cover a wide spectrum (including sales, customers, employees, goods transportation, consumption and services), there can be many ways to design a data warehouse. The levels of detail to be included may also vary substantially. Since a major usage of a data warehouse is to support effective data analysis and data mining, the outcome of preliminary data mining exercises can be used to help guide the design and development of data warehouse structures. This involves deciding those dimensions and levels that are to be included and what preprocessing to perform in order to facilitate quality and efficient data mining.

Multidimensional analysis of sales, customers, products, time, and region: The retail industry requires timely information regarding customer needs, product sales, trends and fashions, as well as the quality, cost, profit, and service of commodities. It is therefore important to provide powerful multidimensional analysis and visualization tools, including the construction of sophisticated data cubes according to the needs of data analysis. The multifeature data cube is a useful data structure in retail data analysis since it facilitates analysis on aggregates with sophisticated conditions.

Analysis of the effectiveness of sales campaigns: The retail industry conducts sales campaigns using advertisements, coupons, and various kinds of discounts and bonuses to promote products and attract customers. Careful analysis of the effectiveness of sales campaigns can help improve company profits. Multidimensional analysis can be used for this purpose by comparing the amount of sales and the number of transactions containing the sales items during the sales period versus those containing the same items before or after the sales campaign. Moreover, association analysis may disclose which items are likely to be purchased together with the items on sale, especially in comparison with the sales before or after the campaign.

Customer retention-analysis of customer loyalty: With customer loyalty card information, one can register sequences of purchases of particular customers. Customer loyalty and purchase trends can be analyzed in a systematic way. Goods purchased at different periods by the same customers can be grouped into sequences. Sequential pattern mining can then be used to investigate changes in customer consumption or loyalty and suggest adjustments on the pricing and variety of goods in order to help retain customers and attract new customers.

Purchase recommendation and cross-reference of items: By mining associations from sales records, one may discover that a customer who buys a particular brand of perfume is likely to buy another set of items. Such information can be used to form purchase recommendations. Purchase recommendations can be advertised on the Web, in weekly flyers, or on sales receipts to help improve customers in selecting items, and increase sales. Similarly, information such as "hot items this week" or attractive deals can be displayed together with the associative information in order to promote sales.

19.8 Business Intelligence and Retailing

19.8.1 Applications of Data Warehousing and Data Mining in the Retail INDUSTRY

The information economy puts a premium on high-quality actionable information – exactly what business intelligence (BI) tools like data warehousing, data mining, and OLAP can provide to the retailers. A close look at the different retail organizational functions suggests that BI can play a crucial role in almost every function. It can give new and often surprising insights about customer behavior, thereby helping the retailers meet their ever-changing needs and desires.

On the supply side, BI can help retailers identify their best vendors and determine what separates them from not so good vendors. It can give retailers better understanding of inventory and its movement and also help improve storefront operations through better category management. Through a host of analyses and reports, BI can also improve retailers' internal organizational support functions like finance and human resource management.

As quite sensitively portrayed in the movie, large chain superstores have nearly forced small independent retailers to close down. At the same time, these large retailers have gained considerable power in the supply chain. They are increasingly dictating terms to the retailers and inventing new ways of attracting customers. But to hold the customer's imagination for long has remained an elusive dream. Changing tastes and preferences, increasing competition, demographic shifts, and the simple "let's try something new" attitude have all been blamed for customer disloyalty. No wonder retailers today are going that extra mile to reach and understand the customer. They are also

getting their act together by streamlining the supply chain, improving storefront operations, and actively exploring alternative channels like the Internet. Technology has played a key role in retailers' effort to compete in this volatile market. Sophisticated retailers have quickly evolved from basic automation to embrace new technologies like CRM, business intelligence, etc. This section explores the various applications of business intelligence in the retail industry. Business intelligence refers to a host of technologies like data warehousing, online analytical processing (OLAP), and data mining, which seek to turn data into actionable information.

19.8.2 Key Trends in the Retail Industry

Rise of superstores: Last two decades has seen the phenomenal rise of the "Chain of superstores" in both the US and Europe. Growing consolidation and globalization in the sector have seen the bargaining power of the retailer increase in the supply chain. We believe that in order to counter saturated domestic markets and increasing competition, leading superstores would continue to expand globally. WalMart acquired Britain's third largest supermarket chain ASDA, to establish itself in Europe. Similarly the grocery giant Safeway has significant presence in both the US and UK.

Customer Relationship Management as a key driver: Smart retailers have reoriented their business around the customer. In the mad rush to acquire new customers, they have realized it is equally important to retain the existing ones. Increased interaction and sophisticated analysis techniques have given retailers unprecedented access to the mind of the customer; and they are using this to develop one-to-one relation with the customer, design marketing and promotion campaigns, optimize store layout, and manage e-commerce operations. For example Safeway uses its ABC loyalty card to record each customer's individual transactions. This coupled with other relevant data has given Safeway tremendous knowledge about customer buying patterns – knowledge that has significantly helped in augmenting customer loyalty.

Supply Chain Management as a key driver: Increasingly retailers are handling their inbound logistics by setting up their own distribution networks. We believe that a vital criterion for success in future would be the ability to harness worldwide distribution and logistics network for purchasing. This global supply chain should ensure high levels of product availability that consumers want to buy.

Rise of Online Retailing: Some say that the Internet will completely change the face of retailing; others believe that the "touch and feel" factor would ultimately dominate and the Net will have only a marginal impact on the shopping behavior. Probably the truth lies somewhere in between. But one thing is sure – online retailing is here to stay. Many retailers realized that and have rushed to start their own e-commerce Web site. We believe that the key to success would be the effectiveness with which retailers integrate the Internet with their existing business model.

19.8.3 Business Intelligence Solutions for the Retail Industry

Business Intelligence (BI) refers to the ability to collect and analyze huge amount of data pertaining to the customers, vendors, markets, internal processes, and the business environment. A data warehouse is the corner stone of an enterprise-wide business intelligence solution; various analytical (OLAP) and data mining tools are used to turn data – stored in the data warehouse – into actionable information.

Delivering Customer Value

Figure 19.1 illustrates the different functions in a typical retail organization. Customer relationship management (CRM) forms the focal point from where the vital insights gained about the customers – using BI tools – are absorbed in the entire organization. BI also plays a critical role in all the other retail functions like supply chain management, store front operations, and channel management. This chapter is an introduction to the various BI applications in the different functions in the retail organization, including support functions like finance and human resources.

Customer Relationship Management

The CRM strategy should include:

(a) Operational CRM: Automating interaction with the customers and sales force, and
(b) Analytical CRM: Sophisticated analysis of the customer data generated by operational CRM and other sources like POS transactions, Web site transactions, and third-party data providers.

A typical retail organization has a huge customer base and often customer's needs are fairly differentiated. Without the means to analyze voluminous customer data, CRM strategy is bound to be a failure. Hence, we believe that analytical CRM forms the core of a retailer's customer relationship strategy.

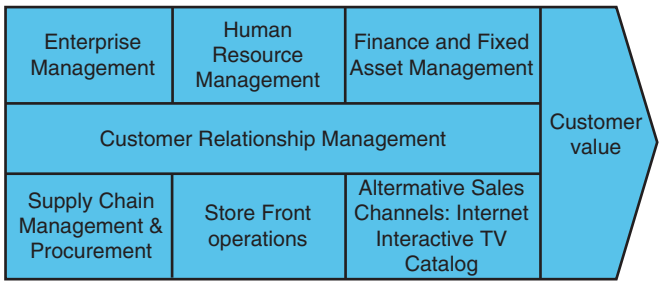

Fig. 19.1. Retail Organization

Marketing and sales functions are the primary beneficiaries of analytical CRM and the main touch points from where the insights gained about the customer is absorbed in the organization.

Analytical CRM uses the key business intelligence tools like data warehousing, data mining, and OLAP to present a unified view of the customer. Following are some of the uses of analytical CRM:

Customer Segmentation: Customer segmentation is a vital ingredient in a retail organization's marketing recipe. It can offer insights into how different segments respond to shifts in demographics, fashions, and trends. For example it can help classify customers in the following segments

- Customers who respond to new promotions
- Customers who respond to new product launches
- Customers who respond to discounts
- Customers who show propensity to purchase specific products

Campaign/Promotion Effectiveness Analysis: Once a campaign is launched its effectiveness can be studied across different media and in terms of costs and benefits; this greatly helps in understanding what goes into a successful marketing campaign. Campaign/promotion effectiveness analysis can answer questions like:

- Which media channels have been most successful in the past for various campaigns?
- Which geographic locations responded well to a particular campaign?
- What were the relative costs and benefits of this campaign?
- Which customer segments responded to the campaign?

Customer Lifetime Value: Not all customers are equally profitable. At the same time customers who are not very profitable today may have the potential of being profitable in future. Hence it is absolutely essential to identify customers with high lifetime value; the idea is to establish long-term relations with these customers. The basic methodology used to calculate customer lifetime value is – deduct the cost of servicing a customer from the expected future revenue generated by the customer, add to this the net value of new customers referred by this customer, and discount the result for the duration of the relationship. Though this sounds easy, there are a number of subjective variables like overall duration of the customer's relation with the retailer, gap between intermediate cash flows, and discount rate. We suggest data mining tools should be used to develop customized models for calculating customer lifetime value.

Customer Loyalty Analysis: It is more economical to retain an existing customer than to acquire a new one. To develop effective customer retention programs it is vital to analyze the reasons for customer attrition. Business intelligence helps in understanding customer attrition with respect to various factors influencing a customer and at times one can drill down to individual transactions, which might have resulted in the change of loyalty.

Cross Selling: Retailers use the vast amount of customer information available with them to cross sell other products at the time of purchase. This effort is largely based on the tastes of a particular customer, which can be analyzed using BI tools based on previous purchases. Retailers can also 'up sell' – sell more profitable products – to the customer at the time of contact.

Product Pricing: Pricing is one of the most crucial marketing decisions taken by retailers. Often an increase in price of a product can result in lower sales and customer adoption of replacement products. Using data warehousing and data mining, retailers can develop sophisticated price models for different products, which can establish price–sales relationships for the product and how changes in prices affect the sales of other products.

Target Marketing: Retailers can optimize the overall marketing and promotion effort by targeting campaigns to specific customers or groups of customers. Target marketing can be based on a very simple analysis of the buying habits of the customer or the customer group; but increasingly data mining tools are being used to define specific customer segments that are likely to respond to particular types of campaigns.

Supply Chain Management and Procurement

Supply chain management (SCM) promises unprecedented efficiencies in inventory control and procurement to the retailers. With cash registers equipped with bar-code scanners, retailers can now automatically manage the flow of products and transmit stock replenishment orders to the vendors. The data collected for this purpose can provide deep insights into the dynamics of the supply chain. However, most of the commercial SCM applications provide only transaction-based functionality for inventory management and procurement; they lack sophisticated analytical capabilities required to provide an integrated view of the supply chain. This is where data warehousing can provide critical information to help managers streamline their supply chain. Some of the applications of BI in supply chain management and procurement are:

Vendor Performance Analysis: Performance of each vendor can be analyzed on the basis of a number of factors like cost, delivery time, quality of products delivered, payment lead time, etc. In addition to this, the role of suppliers in specific product outages can be critically analyzed.

Inventory Control (Inventory levels, safety stock, lot size, and lead time analysis): Both current and historical reports on key inventory indicators like inventory levels, lot size, etc. can be generated from the data warehouse, thereby helping in both operational and strategic decisions relating to the inventory.

Product Movement and the Supply Chain : Some products move much faster off the shelf than others. On-time replenishment orders are very critical for these products. Analyzing the movement of specific products – using BI tools – can help in predicting when there will be need for reorder.

Demand Forecasting : It is one of the key applications of data mining. Complex demand forecasting models can be created using a number of factors like sales figures, basic economic indicators, environmental conditions, etc. If correctly implemented, a data warehouse can significantly help in improving the retailer's relations with suppliers and can complement the existing SCM application.

Storefront Operations

The information needs of the store manager are no longer restricted to the day-to-day operations. Today's consumer is much more sophisticated and she demands a compelling shopping experience. For this the store manager needs to have an in-depth understanding of her tastes and purchasing behavior. Data warehousing and data mining can help the manager gain this insight. Following are some of the uses of BI in storefront operations:

Market Basket Analysis: It is used to study natural affinities between products. One of the classic examples of market basket analysis is the beer–diaper affinity, which states that men who buy diapers are also likely to buy beer. This is an example of "two-product affinity." But in real life, market basket analysis can get extremely complex resulting in hitherto unknown affinities between a number of products. This analysis has various uses in the retail organization. One very common use is for *in-store product placement*. Another popular use is *product bundling*, i.e., grouping products to be sold in a single package deal. Other uses include designing the company's e-commerce Web site and product catalogs.

Category Management: It gives the retailer an insight into the right number of SKUs to stock in a particular category. The objective is to achieve maximum profitability from a category; too few SKUs would mean that the customer is not provided with adequate choice, and too many would mean that the SKUs are cannibalizing each other. It goes without saying that effective category management is vital for a retailer's survival in this market.

Out-Of-Stock Analysis : This analysis probes into the various reasons resulting into an out-of-stock situation. Typically a number of variables are involved and it can get very complicated. An integral part of the analysis is calculating the lost revenue due to product stock out.

Alternative Sales Channels

The success of a retailer in future would depend on how effectively it manages multiple delivery channels like the Internet, interactive TV, catalogs, etc. A single customer is likely to interact with the retailer along multiple channels over a period of time. This calls for an integrated strategy to serve the customer well, which requires smooth flow of information across channels. To

ensure smooth flow of information customer data needs to be collected from different channels in one data warehouse. Customer relationship strategy can then be built around this customer-centric data warehouse. We have already seen how analytical CRM can provide analyses over the centralized data warehouse. In this section we explore how data warehousing and data mining can improve the effectiveness of a channel.

E-Business Analysis: The Internet has emerged as a powerful alternative channel for established retailers. Increasing competition from retailers operating purely over the Internet – commonly known as "e-tailers" – has forced the "bricks and mortar" retailers to quickly adopt this channel. Their success would largely depend on how they use the Net to complement their existing channels. Web logs and information forms filled over the Web are very rich sources of data that can provide insightful information about customer's browsing behavior, purchasing patterns, likes and dislikes, etc. Two main types of analysis done on the Web site data are:

Web Log Analysis: This involves analyzing the basic traffic information over the e-commerce Web site. This analysis is primarily required to optimize the operations over the Internet. It typically includes following analyses:

Site Navigation: An analysis of the typical route followed by the user while navigating the Web site. It also includes an analysis of the most popular pages in the Web site. This can significantly help in site optimization by making it more user friendly.

Referrer Analysis: An analysis of the sites, which are very prolific in diverting traffic to the company's Web site.

Error Analysis: An analysis of the errors encountered by the user while navigating the Web site. This can help in solving the errors and making the browsing experience more pleasurable.

Keyword Analysis: An analysis of the most popular keywords used by various users in Internet search engines to reach the retailer's e-commerce Web site.

Web Housing: This involves integration of Web log data with data from other sources like the POS transactions, third-party data vendors etc. Once the data is collected in a single customer-centric data warehouse, often referred to as Web house, all the applications already described under CRM can be implemented. Often a retailer wants to design specific campaigns for users who purchase from the e-commerce Web site. In this case, segmentation and profiling can be done specifically for the e-customers to understand their needs and browsing behavior. It can also be used to personalize the content of the e-commerce Web site for these users.

Channel Profitability: Data warehousing can help analyze channel profitability, and whether it makes sense for the retailer to continue building up expertise in that channel. The decision of continuing with a channel would also include a number of subjective factors like outlook of key enabling technologies for that channel. For example m-commerce, though not a very profitable channel today, has the potential to be a major alternative channel in the years to come.

Product–Channel Affinity: Some product categories sell particularly well on certain channels. Data warehousing can help identify hidden product-channel affinities and help the retailer design better promotion and marketing campaigns.

Enterprise Management

This typically involves the various activities performed by the top management; and the role of data warehousing and data mining is to provide the top management with reports and analyses to meet their decision-making requirements. One possible BI application in this area is:

Dashboard Reporting on KPIs: Key performance indicators like contribution margin, response rate, campaign costs, customer lifetime value can be presented in dashboard reports to the top management to facilitate decision-making process. Also alerts can be triggered if any KPI reaches a predefined threshold level. These reports can incorporate retail industry benchmarks, provided by third-party researchers, which can be used as threshold levels for various KPIs.

Human Resources

Data warehousing can significantly help in aligning the HR strategy to the overall business strategy. It can present an integrated view of the workforce and help in designing retention schemes, improve productivity, and curtail costs. Some BI applications in HR are:

Human Resource Reports/Analytics: Reports and analysis can be generated to support an integrated view of the workforce. Various analyses include staff movement and performance, workforce attrition by store, workforce performance by store, compensation and attrition, and other customized analyses and reports. The HR data can be integrated with benchmark figures for the industry and various reports can be generated to measure performance vis-a-vis industry benchmarks.

Manpower Allocation: This includes allocating manpower based on the demand projections. According to the seasonal variation in demand, temporary manpower can be hired to maintain service levels. The demand levels vary within one working day also, which can be used to allocate resources accordingly.

HR Portal: Employers need to maintain accurate employee data, which can be viewed by the employees for information relating to compensation, benefits, retirement facilities, etc. Payroll data can be integrated with data from other human resource management applications in the HR data warehouse. This data can then be circulated within the organization through the HR portal.

Training and Succession Planning: Accurate data about the skill sets of the workforce can be maintained in the data warehouse. This can be used to design training programs and for effective succession planning.

Finance and Fixed Asset Management

The role of financial reporting has undergone a paradigm shift during the last decade. It is no longer restricted to just financial statements required by the law; increasingly it is being used to help in strategic decision making. Also, many organizations have embraced a free information architecture, whereby financial information is openly available for internal use. Many analytics described till now use financial data. Many companies, across industries, have integrated financial data in their enterprise wide data warehouse or established separate financial data warehouse (FDW). Following are some of the uses of BI in finance:

Budgetary Analysis: Data warehousing facilitates analysis of budgeted versus actual expenditure for various cost heads like promotion campaigns, energy costs, salary, etc. OLAP tools can provide drill down facility whereby the reasons for cost overruns can be analyzed in more detail. It can also be used to allocate budgets for the coming financial period.

Fixed Asset Return Analysis: This is used to analyze financial viability of the fixed assets owned or leased by the company. It would typically involve measures like profitability per square foot of store space, total lease cost vs. profitability, etc.

Financial Ratio Analysis: Various financial ratios like debt–equity, liquidity ratios, etc. can be analyzed over a period of time. The ability to drill down and join inter-related reports and analyses, provided by all major OLAP tool vendors, can make ratio analysis much more intuitive.

Profitability Analysis: This includes profitability of individual stores, departments within the store, product categories, brands, and individual SKUs. A major component of profitability analysis is the costs incurred by stores/departments and the cost of acquiring, storing, and allocating shelf space to particular product categories, brands, or SKUs. It goes without saying that profitability analysis has an extremely universal appeal and would be required by other groups within the retail organization.

19.9 Summary

In many business settings the comprehensibility of the analyses as well as the degree of validation against known relationships help enhance the perception of the data mining activity, thus creating an environment where this new technology can take root. The success of data mining does not solely depend

on the quality of new algorithms, but also on the usability, comprehensibility, and degree of domain-specific knowledge integrated in the tool.

Data mining and analytics in general demand complex set of skills: business, statistics, databases, operations systems. CDMS, the DM Metadata, is a natural platform for sharing those skills and it helps to avoid duplications of effort.

This section discussed so far data mining in the area of customer relationship management for customer segmentation and customer retention analysis as well as for credit scoring. Industries in which this technology is traditionally applied are those that collect customer data for the purpose of billing. They have always been required to store this data and, thus, can make use of it without introducing new operational systems beforehand. We have helped banks, insurances, and Internet service providers. More and more industries are obtaining customer data while implementing information systems for marketing, sales, and customer service. Besides that, every company could benefit from data mining technology applied to the analysis of costs and revenues, activities and processes, and its performance in general.

Retailers are known for innovation. The most innovative retailers of today are those who are using business intelligence to gain sustained competitive advantage. These retailers have also realized that BI can be used strategically only when it is implemented with utmost care and complete support from the top management. We believe that unless all the user groups are consulted and the objectives clearly defined, BI solution cannot be a success. Also, like any other technology solution, BI cannot exist in vacuum. We strongly believe that it is just a means to an end. The wisdom, gathered by analyzing huge amount of data, should reach every corner of the retail organization. The end objective is to convert this wisdom into effective action. And for this the entire organization should be able to leverage the business intelligence network. Thus the chapter involved various case studies and discussed how data mining is suitable for banking and commercial applications.

19.10 Review Questions

1. With a case study, explain how distributed data mining provides a solution in banking.
2. What are the factors that led to data mining technology at the Bank?
3. Explain CDMS – centralized data mining solution.
4. Explain how data mining is used in various commercial applications.
5. State the basics of decision support system.
6. On what basis do the decision-making challenges arise?
7. How does the data mining supports e-commerce and retail industries?
8. What are the applications of data warehousing and mining in the retail industry?
9. What are the key trends in retail industry?
10. Write a short note on enterprise management.
11. Write a short note on finance and fixed asset management.