Chapter 7 Potential Applications for a Manufacturer

7.1 Introduction

The last two chapters present case studies illustrating the potential applications of data mining for a retailer (Chapter 5) and a service provider (Chapter 6), respectively. These illustrations revolve around customer analytics, which in turn, revolves around "getting to know your customers". They attempt to provide answers to business questions such as: (1) who are the customers? (2) what are their needs? (3) how do they behave? (4) where are they? and (5) why do they churn? In customer analytics, data mining can play a major role in customer relationship management (e.g., customer valuation, cross-/up-selling and churn modelling), market segmentation (e.g., customer profiling and response-based segmentation) and other areas. Data mining applications can contribute significantly to an organisation's competitive edge.

In contrast, this chapter focuses on a manufacturing environment and looks at potential data mining applications for a manufacturer. As highlighted by Gertosio and Dussauchoy (2004), industrial engineering (or more generally, manufacturing) presents unique opportunities for the application of data mining and for the development of new concepts and techniques in this field.

As in the case of business and commerce, there is also a data explosion in manufacturing. This comes about primarily because of the automation and computerisation of many industrial and manufacturing processes that aim to ensure a high quality of production at a low production cost. The increasingly intense competition in the global market has also led to the manufacturing industry actively seeking new technology solutions (e.g., intelligent and automated machines) that enable timely response to increasing customer demand for quick delivery of orders (Hou et al., 2003). A consequence of this automation and computerisation is the recording of large masses of real-time data during the manufacturing process to trace production steps to optimise the process.

These data can be a great asset to an organisation if "previously unknown, and potentially useful information" or "valid, novel, potentially useful, and ultimately understandable patterns" can be extracted from the data (see Frawley et al., 1992; and Fayyad et al., 1996). Data mining applications in manufacturing are well documented in the literature. For example, Gertosio and Dussauchoy (2004), in a recent article, have described a data mining application in a French truck manufacturer that exploits data sets of measures recorded during the test of diesel engines to significantly reduce (by about 25%) the processing time. The application has led to a real-time prediction model and a discovery support environment that allows non-statisticians to extract knowledge of other processes.

Other applications in manufacturing that have been reported in the literature include the use of data mining to: (1) detect and classify quality faults and provide parameter settings to correct the faults (Hou et al., 2003); (2) detect abnormalities and failures and schedule preventive maintenance (Smith, 2002); and (3) generate automatic, prioritised early warning signals about manufacturing problems (Adams, 2002).

Another important development in the manufacturing sector in the past decade is the emphasis on total quality management (TQM). It is now even more critical than before to make use of data generated during the manufacturing process to help improve the process and hence the quality of products. Data mining can play a significant role in TQM too. In addition, data mining applications in manufacturing may be developed for nonmanufacturing aspects such as the management of inventory or human resources. For a manufacturer, issues such as customer relationship management and churn modelling are relevant too.

7.2 Background Information for Data Mining Illustrations

To illustrate potential data mining applications for a manufacturer, consider a fictitious manufacturing company Hi-Quality Engineering. To facilitate discussion, the shortened name “HiQ” will be used from this point onwards to refer to this manufacturing company. Assume that HiQ has been in business for the last eight years and manufactures industrial conveyor belts which include pattern conveyor belts, abrasion-proof conveyor belts, oil-proof conveyor belts, high intensity conveyor belts and Aramid conveyor belts. Currently, one of the business objectives of HiQ is to look at ways to optimise its manufacturing operations by either reducing cost or improving productivity. It believes that it can make use of data mining and the data collected during the manufacturing process to achieve this.

For the data mining applications illustrated in this chapter, HiQ has several databases. The first database (filename = HiQ_assoc.sav) contains data as summarised in Table 7.1. This is used in the data mining application in the next section. Other databases will be discussed in later sections.

7.3 Application 1: Matching of Job and Worker

Six months ago, HiQ opened a factory in a neighbouring country. Reports from factory managers, complaints from workers and consistently low productivity figures have led HiQ management to wonder if there might be a poor fit between certain factory jobs and some workers. HiQ suspects that workers who do not quite fit their jobs perform badly and this affects the overall productivity of the factory, among other things.

Table 7.1 Variables in Database (HiQ_assoc.sav)

Variable	Definition	Label
ID	Identification code	1 to 2500
Fit	Fit between job and worker	0 = Bad 1 = Good
Jobtype	Type of job or work	A to E
Age	Age of worker	18 to 69
Educ	Education level of worker	1 = No formal education 2 = PSLE 3 = GCE O level 4 = GCE A level and above
Nation	Nationality of worker	1 = Local 2 = ASEAN 3 = Asian 4 = Others
Gender	Gender of worker	0 = Male 1 = Female
Train	Training programme of worker	1 = On the job 2 = Small group 3 = Large group 4 = External
YrsExp	Years of experience	1 = Less than 2 years 2 = 2 to less than 5 years 3 = 5 or more years
Ptype	Personality type assessment at interview	1 = Type I personality 2 = Type II personality

HiQ believes that it is likely that a particular factory job may require a particular worker profile for there to be a good match between job and worker. Currently, there are five different job types (labelled A to E) in this factory. To explore this possibility further, HiQ has decided to develop a data mining application to profile "bad-fit" and "good-fit" workers. The findings are expected to help HiQ employ the right kind of workers for the right factory jobs.

The data for mining are extracted from the database HiQ_assoc.sav. These comprise the variables in Table 7.1 and have been collected for a random sample of 500 workers for each of the five job types. Besides an identification code (ID), a performance-based measure of job-worker fit (Fit) and job type (Jobtype), the other variables (i.e., Age, Educ, Nation, Gender, Train, YrsExp and Ptype) relate to the demographic and background characteristics of the workers.

The SPSS Clementine Distribution and Matrix nodes are used to explore the fit between jobs and workers by job type. The results are summarised in Figure 7.1. As can be seen, job type E has the worst fit - in particular, there is bad fit among 53.00% of the workers. This is followed by job type D (37.00%). Given these findings, HiQ has decided to focus on job type E. Further analyses include a directed web diagram and association analyses for this segment of workers. The SPSS Clementine nodes Web, Apriori and GRI are used. The results are shown in Figure 7.2.

The directed web diagram (upper panel of Figure 7.2) shows that bad fit is strongly associated with workers with less than two years of experience, who possess Type II personality, who are Asian (vis-à-vis, say, local or ASEAN) and who are female. Bad fit is also moderately associated with a large-group type of training programme. On the other hand, good fit is moderately associated with workers who are male and who possess Type I personality.

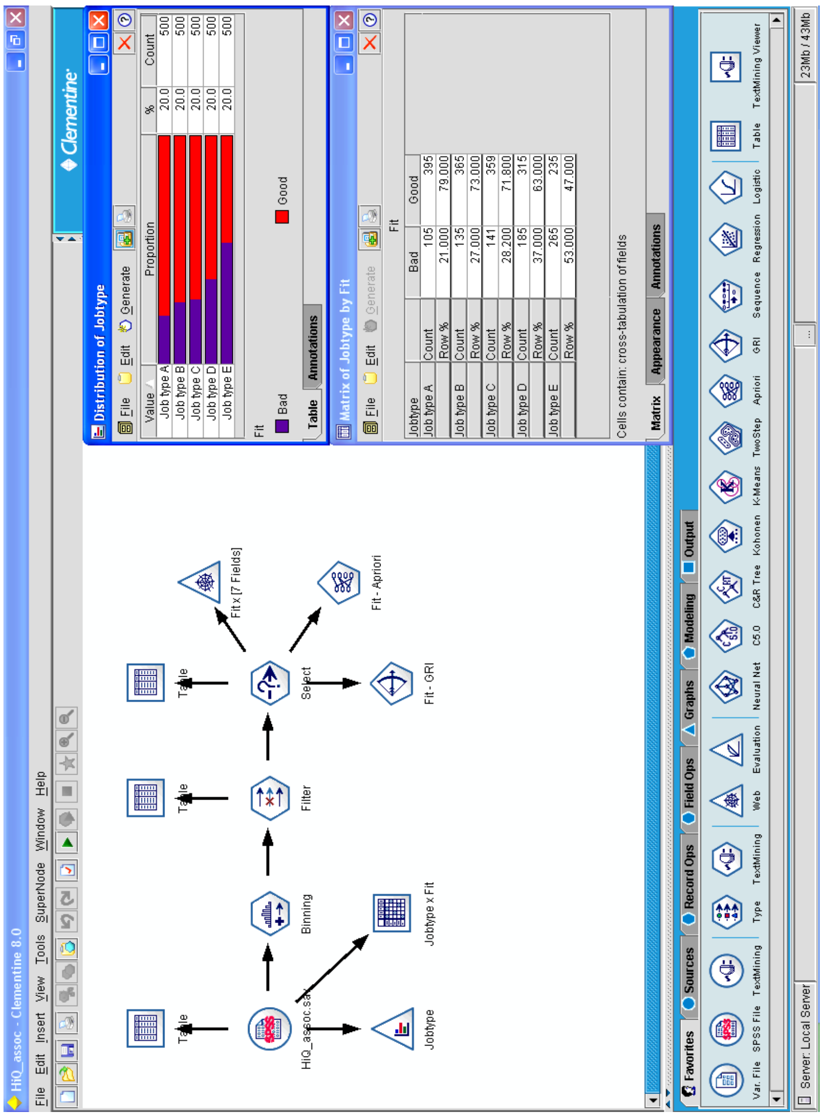

Figure 7.1 Exploration of Fit by Job Type

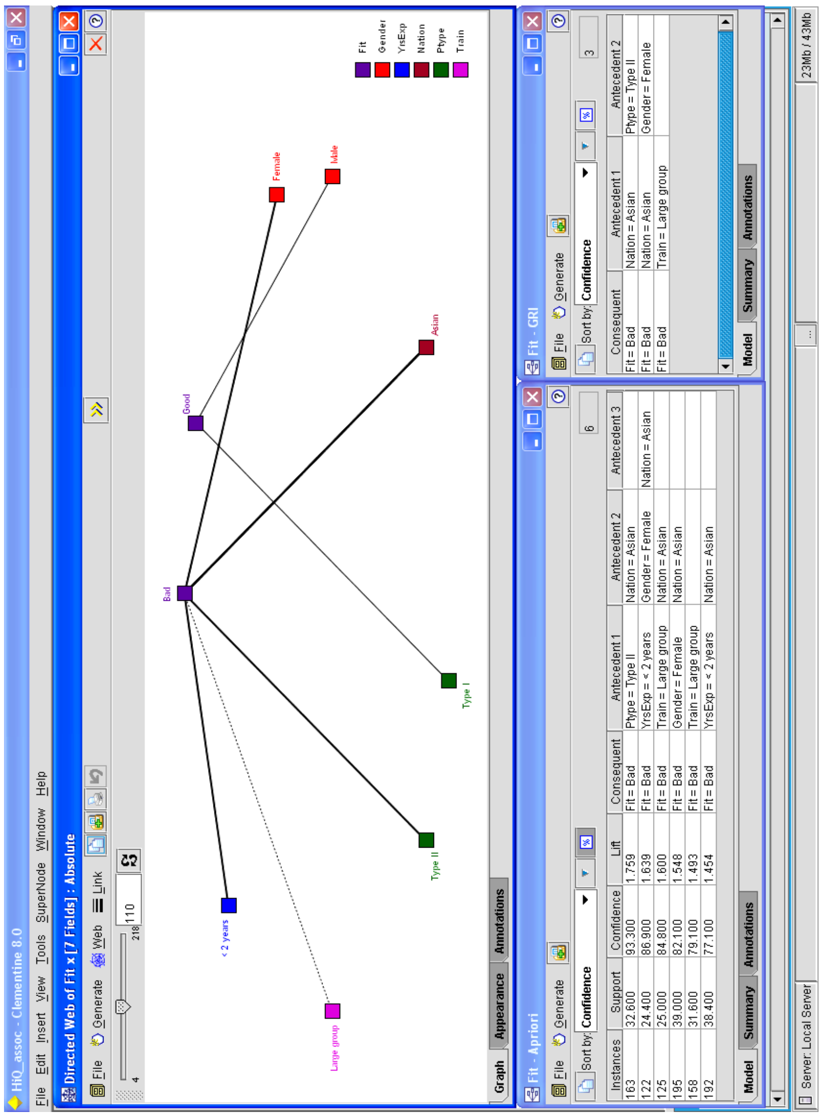

Figure 7.2 Directed Web Diagram and Association Analyses

HiQ plans to use these findings for its recruitment and deployment decisions for job type E. The aim is to recruit and deploy the right worker for the right job.

The web diagram shows only the relationships between good/bad fit and the demographic and background characteristics of the workers one at a time. More information can be extracted from association analysis, which has the ability to look at several antecedents simultaneously.

Apriori association analysis (see the lower left panel of Figure 7.2) yields the following rules for bad fit between jobs and workers (in descending order of confidence): (1) Asian workers with Type II personality; (2) female Asian workers with less than two years of experience; (3) Asian workers with large-group type of training; (4) Asian workers who are female; (5) workers who have undergone large-group type of training programmes; and (6) Asian workers with less than two years of experience. The GRI results (presented in the lower right panel) are a subset of the Apriori results. These association findings can help HiQ with its recruitment and deployment decisions.

The analyses above are focussed only on workers performing job type E. The same analyses can also be done for the other job types.

7.4 Application 2: Determination of Optimal Conditions for Storage

The objective of this data mining application is to uncover the relationships between the strain rate of rubber (one of the raw materials used in the manufacturing process) and storage conditions. It is hoped that based on the uncovered relationships, the optimal conditions for storage can be determined.

Rubber is one of the essential materials used in the manufacture of conveyor belts. In view of the sheer volume, HiQ does not have a warehouse to store the material. Instead, rubber is kept in several nearby storage facilities. These facilities have varying storage conditions, even within the same facility. Also, to explore the different effects of varying storage conditions on rubber, these storage facilities and HiQ have been experimenting with different settings of the storage conditions (e.g., humidity and temperature) in the past.

To produce high quality conveyor belts, the strain rate of the stored rubber is tested and monitored regularly to ensure that the rubber is in a good condition before entering the manufacturing process. There are five levels of strain rate: the first level represents excellent condition while the second to fourth levels represent an acceptable range of deterioration (in decreasing order of acceptability from the second to fourth levels). The fifth level of strain signifies that the rubber is defective and cannot be used in the manufacturing process. The data for analysis is captured by the database HiQ_cluster.sav (see Table 7.2).

As shown, the database comprises the storage conditions (i.e., the number of storage days, type of packaging material, light intensity, air pressure, ozone level, humidity, temperature and number of items in each packaging), the strain rate of rubber and a serial number to identify each storage record or observation.

To determine the optimal conditions for storage, HiQ has decided to embark on a two-stage analysis. In the first stage of the data mining application, segmentation is performed on the data (using clustering) to group the observations into several segments based on the storage conditions (as clustering criteria). HiQ has found this to be necessary as there are just too many possible combinations of storage conditions to manage. Hence, clustering is used here as a data reduction device to collapse sets of similar combinations of storage conditions into a manageable number of clusters. In the second stage, visualisation is used to examine the relationships between the strain rate and the storage conditions (i.e., cluster profile) within each cluster.

Table 7.2 Variables in Database (HiQ_cluster.sav)

Variable	Definition	Label
SerNum	Serial number	1 to 2050
SDays	Number of storage days	
PkMat	Type of packaging material	A = Type A packaging B = Type B packaging
LIntens	Light intensity	Low = Low light intensity Normal = Normal light intensity High = High light intensity
APress	Air pressure	Normal = Normal pressure High = High pressure
OzLev	Ozone level	Low = Low ozone level Normal = Normal ozone level High = High ozone level
Humid	Humidity (%)	
Temp	Temperature (^{o}C)	
NItems	Number of items in packaging	
Strain	Strain rate of rubber	1 to 5 (in decreasing order of acceptability)

Before clustering is attempted, the Statistics node in SPSS Clementine is used to compute the descriptive statistics of selected metric variables (i.e., SDays, Humid and Temp). The results are summarised in Figure 7.3.

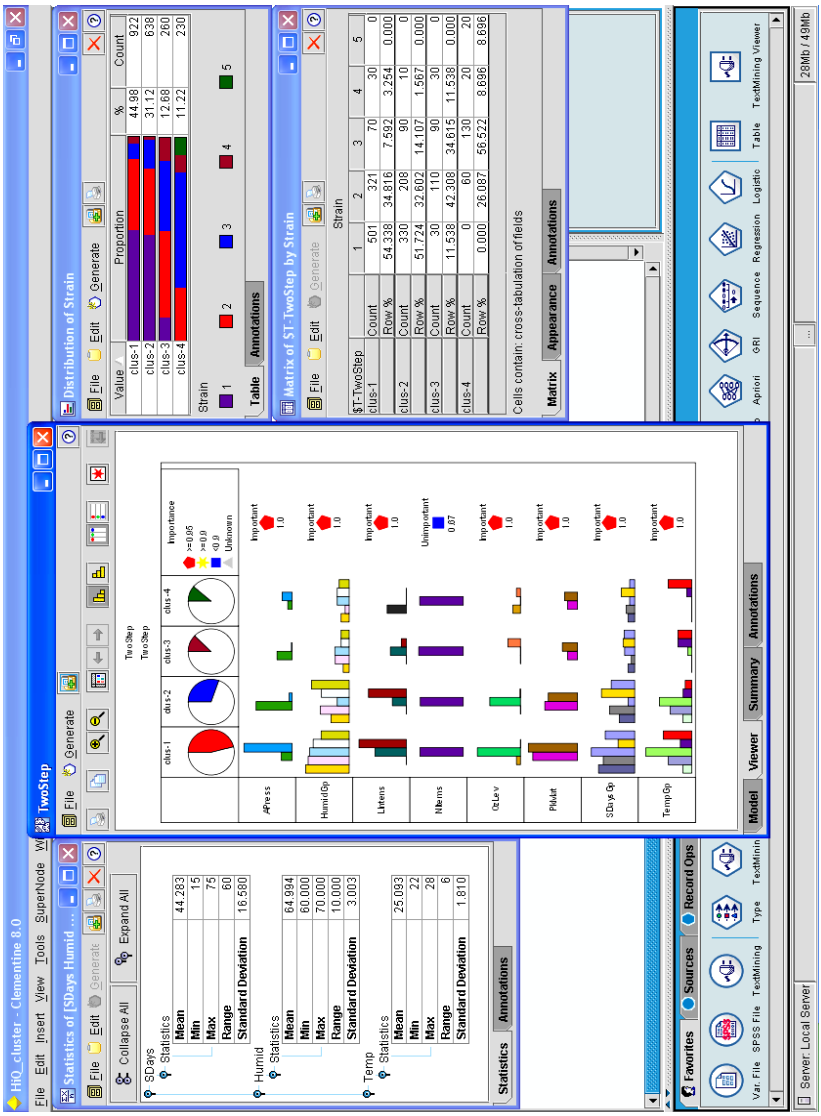

Figure 7.3 Preliminary and Clustering Results

As shown (see the left panel of Figure 7.3), these variables are measured on quite different scales. For example, the value for SDays ranges from 15 to 75, for Humid from 60 to 70, and for Temp from 22 to 28. As variables with larger values are likely to dominate the clustering solution, binning is used to mitigate this problem. In particular, three new variables are derived using SPSS Clementine Binning nodes to categorise the values of SDays, Humid and Temp into five quintiles of approximately equal counts of observations.

For clustering, the TwoStep node in SPSS Clementine is applied. The optimal solution has four clusters, with cluster sizes of 922, 638, 260 and 230 observations for Cluster 1 to Cluster 4, respectively. The results are summarised in Figure 7.3 (see middle panel). As clustering is done using storage conditions as the clustering criteria, observations in the same cluster have similar storage conditions and observations in different clusters have dissimilar storage conditions.

The clustering results in Figure 7.3 show the cluster profiles visually. To relate the strain rate to the cluster membership, the SPSS Clementine Distribution and Matrix nodes are used. The results are also shown in Figure 7.3 (in the right panel).

It can be seen that the storage conditions profiled by Cluster 1 (to be referred to as Cluster 1 storage conditions) are the most favourable in terms of their impact on the strain rate of rubber and the storage conditions profiled by Cluster 4 (to be referred to as Cluster 4 storage conditions) are the least favourable. Specifically, Cluster 1 storage conditions are associated with 54.34% of strain rate 1 and 34.81% of strain rate 2 (see lower right panel of Figure 7.3), giving a total of about 89.15% of favourable storage outcomes. The situation is similar for Cluster 2 storage conditions (with a total of 84.32% of favourable storage outcomes). However, Cluster 3 storage conditions are associated with 34.62% of strain rate 3 and 11.54% of strain rate 4 (both outcomes of which are less favourable). Finally and worst, Cluster 4 storage conditions are associated with unfavourable outcomes of 56.52% of strain rate 3, 8.70% of strain rate 4 and 8.70% of strain rate 5. Hence, it can be concluded that optimal storage conditions are those captured by Cluster 1 and Cluster 2.

To examine the cluster profiles in greater detail, more graphical nodes are used. The results are summarised in Figures 7.4. As can be seen, the number of storage days, packaging material, humidity and number of items per packaging do not seem to differentiate among the four cluster profiles significantly. Instead, the cluster profiles of storage conditions appear to be significantly differentiated as follows:

1) Cluster 4 has high levels of light intensity while Cluster 3 has a large proportion of low light intensity.
2) Cluster 2 has high levels of air pressure while Cluster 1 has a large proportion of normal air pressure.
3) Cluster 2 has low ozone levels while Cluster 3 has normal ozone levels. Further, Cluster 1 has mostly low ozone levels while Cluster 4 has a large proportion of high ozone levels.
4) Cluster 3 has high levels of temperature while Cluster 4 has very high levels of temperature.

Combining the incidence of the different strain rates and the cluster profiles, HiQ concludes that to minimise the strain on rubber, the storage facilities should have the following storage conditions: (1) mostly normal light intensity, (2) low ozone levels, and (3) relatively lower levels of temperature. The converse is also true in that the following storage conditions are associated with higher strain on the rubber: (1) low or high light intensity, (2) normal or high ozone levels, and (3) relatively higher levels of temperature. For light intensity and ozone levels, HiQ has already decided on the definition of the different levels at the data collection stage. For temperature, it is a binned variable. Figure 7.5 shows the binning boundaries and means for each temperature category.

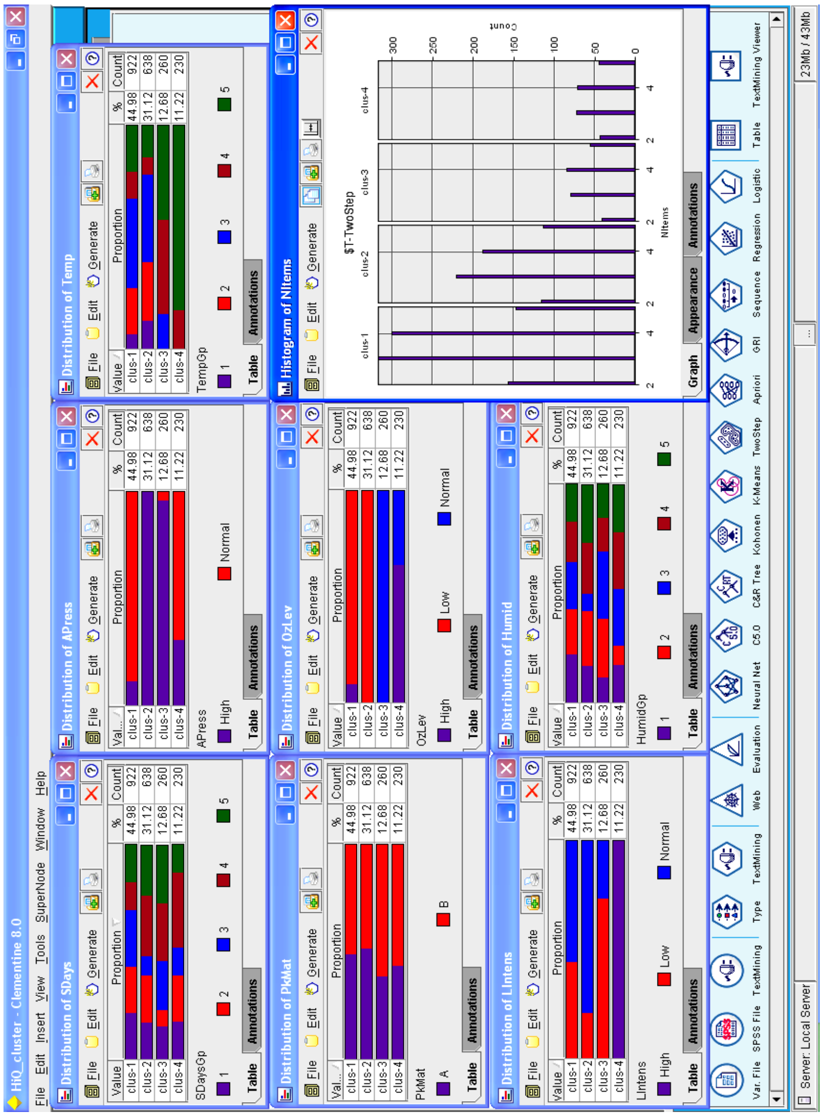

Figure 7.4 Cluster Profiles and Storage Conditions

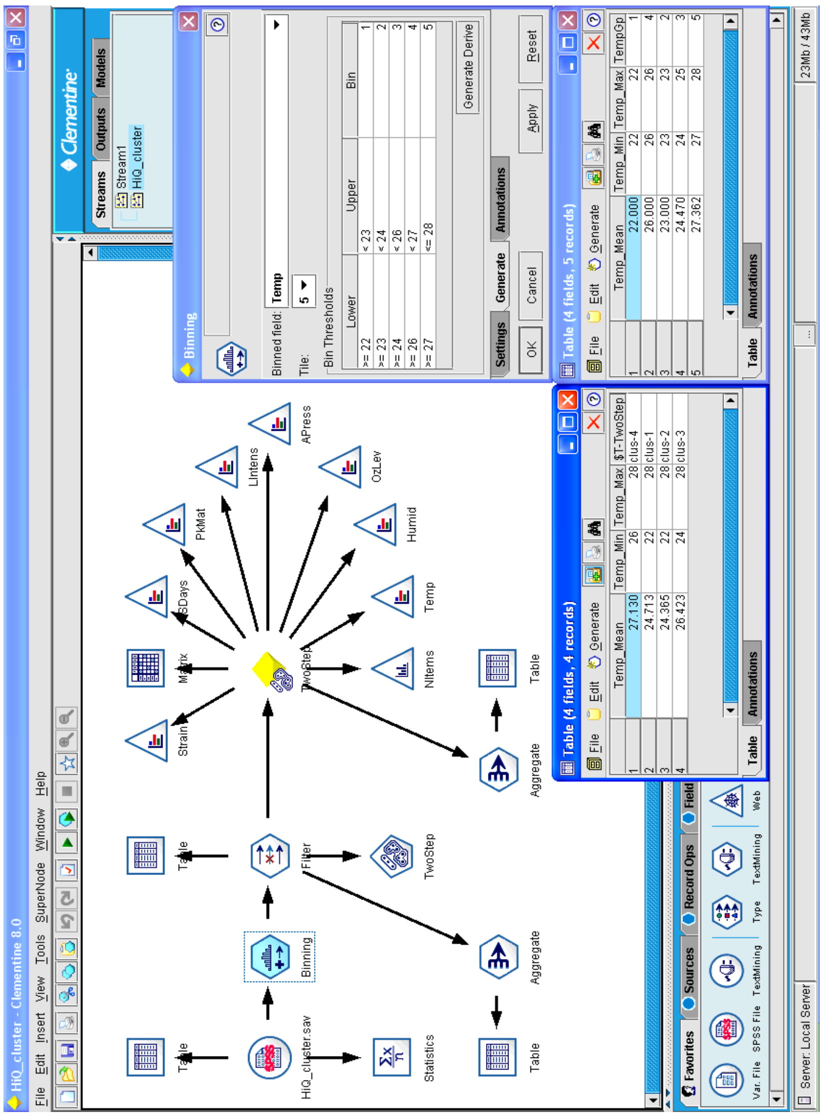

Figure 7.5 Further Analyses on Temperature

The results suggest that temperature levels below 24.5 degrees Celsius are desirable. Finally, although air pressure is significantly different among the four clusters, there is no systematic pattern between the levels of air pressure and the strain rate.

Based on the findings above, HiQ has decided to modify as well as monitor the storage conditions of the facilities to reduce the strain on the rubber stored. HiQ has also concluded that there is sufficient evidence to show that systematic relationships between the strain rate of rubber and storage conditions exist. Hence, it plans to move on to predictive modelling to investigate the relationships further at a later date when more data are collected. It also wants to assess the impact of the modified storage conditions on the strain on rubber.

7.5 Application 3: Fault Detection and Development of Operating Guidelines

In this final data mining application, HiQ seeks to understand the relationship between manufacturing parameters and output quality. The findings can facilitate the formulation of operating guidelines for machine operators on the specific settings of process parameters to correct detected quality problems. HiQ has decided to construct decision trees to generate rules that can be easily understood and used by the operators. The database used in this application (HiQ_pred.sav) is summarised in Table 7.3.

The database contains 6740 records, out of which 4500 (66.76%) are non-faulty and 2240 (33.24%) faulty. The breakdown of the faulty product types is as follows: 550 (8.16%) have bubbles in the conveyor belt (Type 1), 580 (8.61%) are wrinkled (Type 2), 550 (8.16%) are dead/cooled (Type 3) and 560 (8.31%) have uneven thickness (Type 4). Non-faulty belts are designated as Type 0.

Table 7.3 Variables in Database (HiQ_pred.sav)

Variable	Definition	Label
ID	Identification code	1 to 6740
Speed	Speed of revolution of finished product (m/min)	
Thick1	Thickness of product at Location 1 (mm)	
Thick2	Thickness of product at Location 2 (mm)	
Thick3	Thickness of product at Location 3 (mm)	
Temp1	Temperature at Roller 1 (^{o}C)	
Temp2	Temperature at Roller 2 (^{o}C)	
Temp3	Temperature at Roller 3 (^{o}C)	
Temp4	Temperature at Roller 4 (^{o}C)	
Fault	Fault type category	0 = Good 1 = Bubbles 2 = Wrinkled 3 = Dead/cooled 4 = Uneven

The following process parameters are also captured: speed of revolution (Speed), thickness at three locations (Thick1, Thick2 and Thick3) and temperature at four rollers (Temp1, Temp2, Temp3 and Temp4).

To provide an unbiased estimate of the performance of the models when applied to data outside the construction data set, HiQ_pred.sav is randomly partitioned into a 80% (5392 records) construction data set (filename = HiQ_predc.sav) and a 20% (1348 records) validation data set (filename = HiQ_predv.sav). In addition, the model which performs the best on the validation data set is selected as the best model for deployment.

HiQ believes that uneven thickness (i.e., Type 4) is the most serious fault. Hence, it wishes to specify unequal relative misclassification costs for the different fault types. By default, a correct fault detection is assigned a misclassification cost of 0 whereas an incorrect fault detection is assigned a misclassification cost of 1. For Type 4, however, an incorrect classification as no fault (i.e., Type 0) is assigned a misclassification cost of 3 and an incorrect classification as Type 1, Type 2 or Type 3 is assigned a misclassification cost of 2. These are summarised in Table 7.4 and incorporated into the predictive modelling process. The Clementine data mining stream for this application is given in Figure 7.6. It shows also the derivation of new variables as described in a later paragraph.

Table 7.4 Relative Misclassification Costs

Fault Type: Actual	Fault Type: Predicted				
	0	1	2	3	4
0	0	1	1	1	1
1	1	0	1	1	1
2	1	1	0	1	1
3	1	1	1	0	1
4	3	2	2	2	0

Before modelling is performed, preliminary data exploration using visualisation tools is performed on the construction data set (HiQ_predc.sav). Results from the SPSS Clementine Histogram and Plot nodes are shown in Figure 7.7.

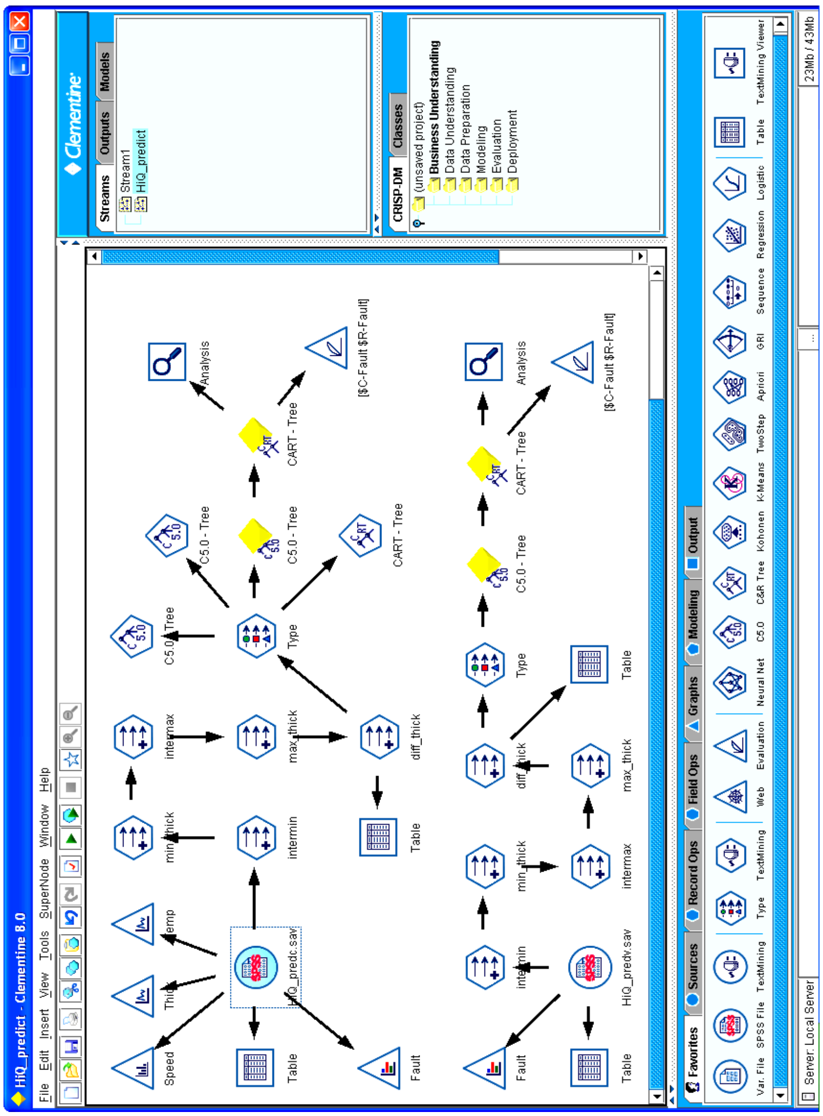

Figure 7.6 Data Mining Stream for Fault Detection

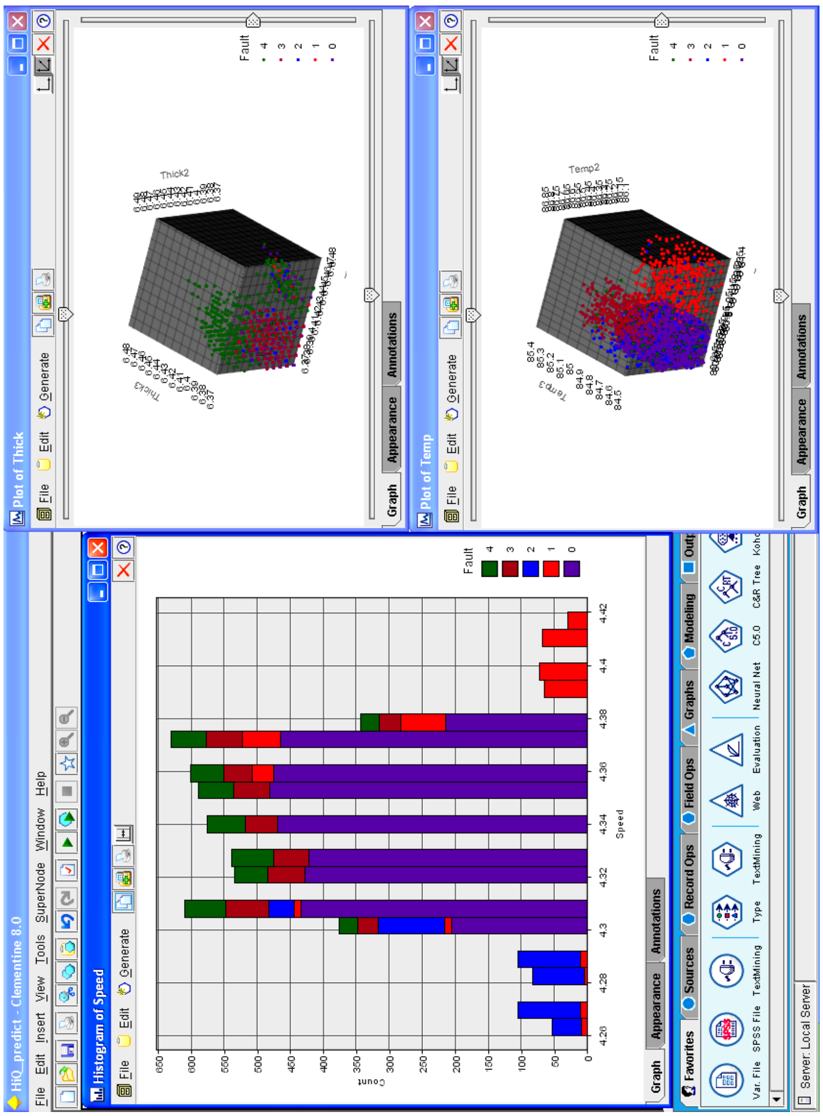

Figure 7.7 Exploratory Analysis Results

The histogram results show that: (1) all fault types are associated with revolution speeds of between 4.30 and 4.38; (2) higher revolution speeds above 4.38 produce only conveyor belts with bubbles (Type 1); and (3) lower revolution speeds below 4.32 produce wrinkled conveyor belts (Type 2).

Through exploration, the rotating plots (see right panels of Figure 7.7) suggest that uneven belts (Type 4) are associated with high values in Thick1, Thick2 or Thick3, and also low values in Temp3. Further, belts with bubbles (Type 1), belts that are wrinkled (Type 2) and dead/cooled conveyor belts (Type 3) seem to have low values in Thick1. In addition, Type 2 is also associated with low values of Temp1 and Temp2.

After reviewing the preliminary results, the engineers in HiQ believe the maximum thickness measured in the three locations (i.e., maximum of [Thick1, Thick2, Thick3] – denoted as "max_thick"), minimum thickness measured in the three locations (i.e., minimum of [Thick1, Thick2, Thick3] – denoted as "min_thick") and the difference in thickness (i.e., "max_thick" – "min_thick" – denoted as "diff_thick") are likely to be more useful in differentiating the fault types than Thick1, Thick2 or Thick3 individually. Hence, three derived variables measuring maximum, minimum and difference in thickness are created. Also, the engineers believe that there may be interaction effects among the process parameters.

As this data mining application focuses not only on fault detection but also on the development of operating guidelines, it is decided that decision trees is the most appropriate tool to apply. Accordingly, the SPSS Clementine nodes C5.0 and C&R Tree (or CART – classification and regression tree) are applied to the data sets. The models are constructed on the construction data set (i.e., HiQ_predc.sav) and validated on the validation data set (i.e., HiQ_predv.sav). The validation results – both the accuracy rates and lift charts – are shown in Figure 7.8.

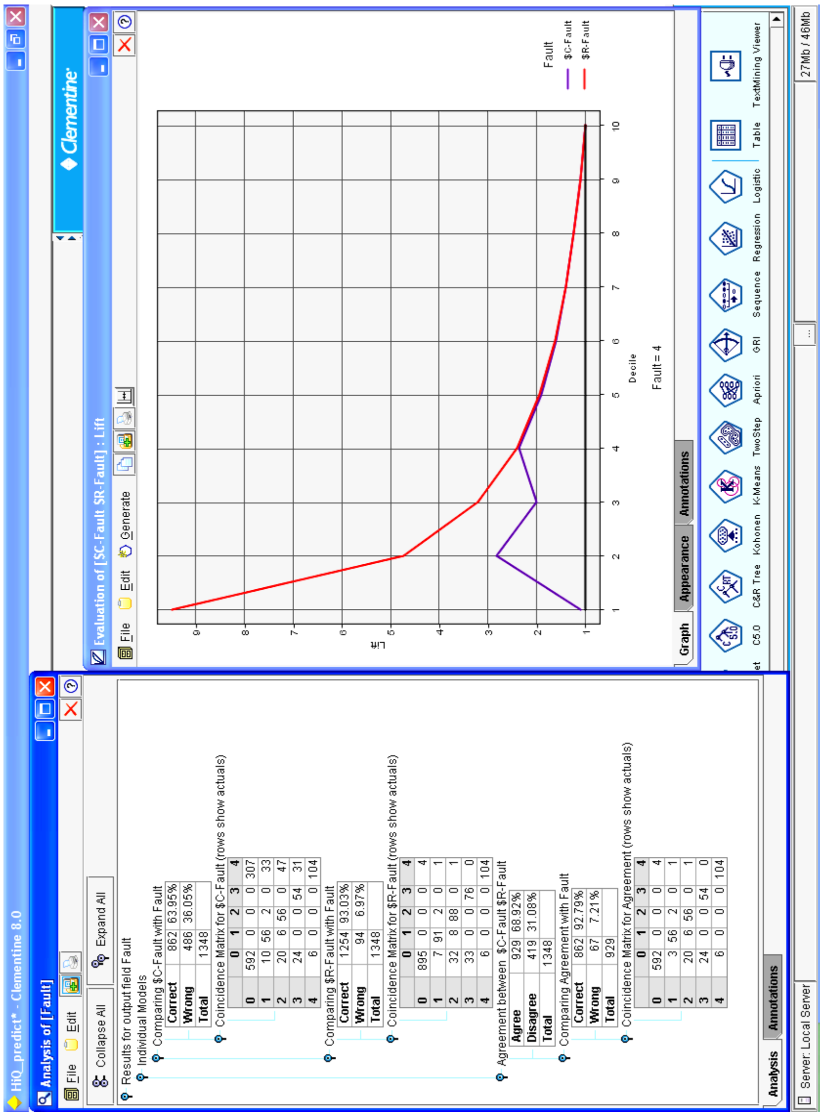

Figure 7.8 Validation Results

As mentioned earlier, different relative misclassification costs are incorporated in the predictive modelling process (i.e., the decision trees).

The left panel of Figure 7.8 shows that the CART decision tree performs significantly better than the C5.0 decision tree. In particular, their overall accuracy rates are 93.03% and 63.95%, respectively. Also, the CART lift chart dominates the C5.0 lift chart for the first four deciles. The lift charts are the same after that. Since the misclassification cost for Type 4 faults is the most costly, the lift charts are plotted to focus on the prediction of Type 4 faults.

Column 4 of the coincidence matrix (see left panel of Figure 7.8) shows that the overall hit rate for Type 4 faults is very much higher for the CART decision tree (104/[104 + 0 + 1 + 1 + 4] = 94.55%) than that for the C5.0 decision tree (104/[104 + 31 + 47 + 33 + 307] = 19.92%). In fact, the C5.0 decision tree predicts 307 instances of Type 0 faults (i.e., no faults) as Type 4 faults.

The final CART decision tree (based on the construction data set) is presented in Figure 7.9. The rules for fault detection and the formulation of operating guidelines for operators on the specific settings can be summarised as follows:

1) Type 0 (no fault conveyor belts): max_thick below 6.41 mm, Temp1 below 80.905 ^{o}C, speed above 4.295 m/min but below 4.385 m/min and Temp3 below 85.005 ^{o}C (Node 9);
2) Type 1 (bubbles in conveyor belts): max_thick below 6.41 mm and Temp1 above 80.905 ^{o}C (Node 4);
3) Type 1 (bubbles in conveyor belts): max_thick below 6.41 mm, Temp1 below 80.905 ^{o}C, Temp3 below 85.005 ^{o}C and speed above 4.385 m/min (Node 10);
4) Type 2 (wrinkled conveyor belts): max_thick below 6.41 mm, Temp1 below 80.905 ^{o}C and speed below 4.295 m/min (Node 5);

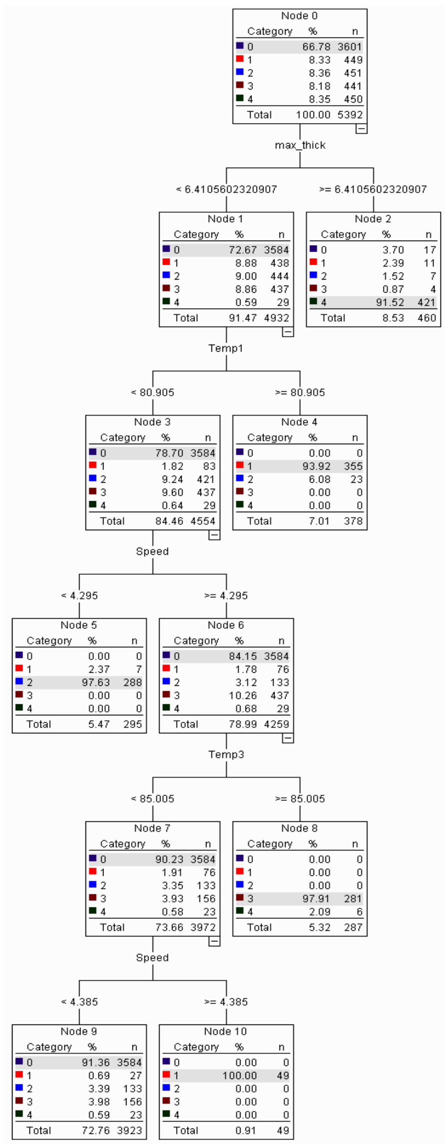

Figure 7.9 CART Decision Tree

5) Type 3 (dead/cooled conveyor belts): max_thick below 6.41 mm, Temp1 below 80.905 ^{o}C and speed above 4.295 m/min and Temp3 above 85.005 ^{o}C (Node 8); and

6) Type 4 (uneven conveyor belts): maximum thickness above 6.41 mm (Node 2).

To summarise, the manufacturing process parameters required to produce good conveyor belts are (see Type 0 listed above): a maximum thickness below 6.41mm, temperature at roller 1 below 80.905 ^{o}C, revolutionary speed between 4.295 and 4.385 and temperature at roller 3 below 85.005 ^{o}C.

The generated rules are useful in helping engineers to understand the operating parameters associated with a faulty conveyor belt and to develop guidelines to reset the operating parameters when the various types of faulty belts occur. More generally, the rules can serve as guidelines for operators to control the manufacturing process and produce good quality conveyor belts. In addition, engineers can also use the rules to train new machine operators.

7.7 Concluding Remarks

This chapter illustrates potential data mining applications in the manufacturing industry. It aims to link the earlier discussion on data mining to specific applications. The illustrative data mining applications are not exhaustive. With some imagination and creativity, a lot more applications can be developed for the manufacturing industry. There is no doubt that data mining can contribute towards enhancing the competitive edge of manufacturing firms.

Collectively, Chapters 5, 6 and 7 cover the retail industry, service industry and manufacturing industry. While these industries comprise profit-oriented and commercial organisations, data mining can be very useful for non-profit and non-commercial organisations too (e.g., statutory boards, government bodies and charitable organisations). For example, data mining can be used by crime investigators and security personnel (e.g., the police) to develop crime-related and security-related applications. Mena (2003) discusses investigative data mining in the areas of fraud, drug trafficking, money laundering, border controls, crime detection, early warning systems, intrusion and anomaly detection, facial recognition … etc. Further, data mining can also be used to detect tax evasion and tax fraud (Gavin, 2002).

Data mining can also be used in public (i.e., government funded) healthcare facilities and hospitals. The mass of data generated by healthcare transactions are too complex and voluminous to be processed and analysed by traditional methods. Here, data mining can improve decision making by discovering patterns and trends in the mass of complex data. The insights gained from data mining can influence cost, revenue and operating efficiency while maintaining a high level of care (Silver et al., 2001). As argued by Benko and Wilson (2003), healthcare organisations that perform data mining are better positioned to meet their long-term needs. Data mining applications can also benefit healthcare providers (i.e., hospital, clinics, doctors … etc.) and recipients (i.e., patients), say, by identifying effective treatments and best practices (e.g., Kolar, 2001).

For charitable organisations, data mining can be used to identify prospective donors and for academic researchers, data mining tools can supplement traditional statistics (see, for example, Koh et al., 2004). The list of potential data mining applications goes on.

Certainly, data mining is a very powerful methodology and technology that can be applied in many different commercial and non-commercial contexts.

Chapter References

Adams, L. (2002), "Mining factory data", *Quality*, Vol. 41 No. 5, pp. 28-31.

Benko, A. and Wilson, B. (2003), "Online decision support gives plans an edge", *Managed Healthcare Executive*, Vol. 13 No. 5, p. 20.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1992), "From data mining to knowledge discovery: an overview", in *Advances in Knowledge Discovery and Data Mining*, Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (eds.), AAAI Press/The MIT Press, California, pp. 1-34.

Frawley, W., Piatetsky-Shapiro, G. and Matheus, C. (1996), "Knowledge discovery in databases: an overview", *AI Magazine*, Vol. 14 No. 3, pp. 57-70.

Gavin, R. (2002), "States turn super tax sleuths – officials peruse millions of records in just minutes with data-mining systems", *Wall Street Journal*, May 8, p. B.6.

Gertosio, C. and Dussauchoy, A. (2004), "Knowledge discovery from industrial databases", *Journal of Intelligent Manufacturing*, Vol. 15 No. 1, pp. 29-37.

Hou, T. H., Liu, W. L. and Lin, L. (2003), "Intelligent remote monitoring and diagnosis of manufacturing processes using an integrated approach of neural networks and rough sets", *Journal of Intelligent Manufacturing*, Vol. 14 No. 2, pp. 239-253.

Koh, H. C., Sim, K. L. and Killough, L. N. (2004) "The interaction effects of lean production manufacturing practices, compensation, and information systems on production costs: a recursive partitioning model", *Advances in Managerial Accounting*, Volume 12, pp. 115-135.

Kolar, H. R. (2001), "Caring for healthcare", *Health Management Technology*, Vol. 22 No. 4, pp. 46-47.

Mena, J. (2003), *Investigative Data Mining for Security and Criminal Detection*. Butterworth Heinemann, Massachusetts.

Silver, M., Sakata, T., Su, H. C., Herman, C., Dolins, S. B. and O'Shea, M. J. (2001), "Case study: how to apply data mining techniques in a healthcare data warehouse", *Journal of Healthcare Information Management*, Vol. 15 No. 2, pp.155-164.

Smith, J. (2002), "Using data mining for plant maintenance", *Plant Engineering*, Vol. 56 No. 12, pp. 26-30.