Chapter 4: Data Mining

Jason C. H. Chen, Ph.D.
Professor of MIS
School of Business Administration
Gonzaga University
Spokane, WA 99258
chen@jepson.gonzaga.edu

Learning Objectives

• Define data mining as an enabling technology for business intelligence
• Understand the objectives and benefits of business analytics and data mining
• Recognize the wide range of applications of data mining
• Learn the standardized data mining processes
  ➢ CRISP-DM
  ➢ SEMMA
  ➢ KDD

(Continued...)

Questions for the Opening Vignette

1. Why should retailers, especially omni-channel retailers, pay extra attention to advanced analytics and data mining?
2. What are the top challenges for multi-channel retailers? Can you think of other industry segments that face similar problems/challenges?
3. What are the sources of data that retailers such as Cabela’s use for their data mining projects?
4. What does it mean to have a “single view of the customer”? How can it be accomplished?
5. What type of analytics help did Cabela’s get from their efforts? Can you think of any other potential benefits of analytics for large-scale retailers like Cabela’s?
6. What was the reason for Cabela’s to bring together SAS and Teradata, the two leading vendors in the analytics marketplace?
7. What is in-database analytics, and why would you need it?

Learning Objectives

• Understand the steps involved in data preprocessing for data mining
• Learn different methods and algorithms of data mining
• Build awareness of the existing data mining software tools
  ➢ Commercial versus free/open source
• Understand the pitfalls and myths of data mining

Covers the following pages

• pp. 145 – 171
• p. 176 (Decision Trees only)
• pp. 186 – 195
• AC 4.6 (p.189-192)
• NO ECC

1. Why should retailers, especially omni-channel retailers, pay extra attention to advanced analytics and data mining?

Utilizing large and information-rich transactional and customer data (that they collect on a daily basis) to optimize their business processes is not a choice for large-scale retailers anymore, but a necessity to stay competitive.

2. What are the top challenges for multi-channel retailers? Can you think of other industry segments that face similar problems?

The retail industry is amongst the most challenging because of the change that they have to deal with constantly. Understanding customer needs, wants, likes, and dislikes is an ongoing challenge. As the volume and complexity of data increase, so does the time spent on preparing and analyzing it.

Prior to the integration of SAS and Teradata, data for modeling and scoring customers was stored in a data mart. This process required a large amount of time to construct, bringing together disparate data sources and keeping statisticians from working on analytics.
3. What are the sources of data that retailers such as Cabela’s use for their data mining projects?

- Cabela’s uses large and information-rich transactional and customer data (that they collect on a daily basis). In addition, through Web mining they track clickstream patterns of customers shopping online.

4. What does it mean to have “a single view of the customer”? How can it be accomplished?

- Having a single view of the customer means treating the customer as a single entity across whichever channels the customer utilizes. Shopping channels include brick-and-mortar, television, catalog, and e-commerce (through computers and mobile devices). Achieving this single view helps to better focus marketing efforts and drive increased sales.

5. What type of analytics help did Cabela’s get from their efforts? Can you think of any other potential benefits of analytics for large-scale retailers like Cabela’s?

- Cabela’s has long relied on SAS statistics and data mining tools to help analyze the data it gathers from sales transactions, market research, and demographic data associated with its large database of customers. Using SAS data mining tools, Cabela’s analysts create predictive models to optimize customer selection for all customer contacts. Cabela’s uses these prediction scores to maximize marketing spending across channels and within each customer’s personal contact strategy.

- These efforts have allowed Cabela’s to continue its growth in a profitable manner. In addition, dismantling the information silos, and integration of SAS and Teradata, enabled them to create “a holistic view of the customer.” Since this works so well for the sales/customer side of the business, it could also work in other areas as well. Supply chain is one example. Analytics could help produce a “holistic view of the vendor” as well.

6. What was the reason for Cabela’s to bring together SAS and Teradata, the two leading vendors in the analytics marketplace?

- Cabela’s was already using both for different elements of their business. Each of the two systems was producing actionable analysis of data.

- But by being separate, too much time was required to construct data marts, bringing together disparate data sources and keeping statisticians from working on analytics. Now, with the integration of the two systems, statisticians can leverage the power of SAS using the Teradata warehouse as one source of information.

7. What is in-database analytics, and why would you need it?

- In-database analytics refers to the practice of applying analytics directly to a database or data warehouse rather than the traditional practice of first transforming into the analytics application’s data format. The time it takes to transform production data into a data warehouse format can be very long. In-database analytics eliminates this need.

Event Driven Alert - A Scenario

- Three transactions were posted in a credit account while the account holder was traveling in summer 2010:
  - Ranch 99 San Jose, California $102.33 Aug. 1, 2010
  - Exxon Austin, Texas  $99.12 August 3, 2010
  - Exxon Houston, Texas  $120.44 August 5, 2010

- What action will you (the credit company) take and why?

- How the scenario can be detected?

Q/Quotes

- “What will be the killer applications in the corporation?” – an “age-old” question

- “Data Mining”
  - replied by “Dr. Penzias” (Nobel laureate and former chief scientist of Bell Labs)
  - “Data mining will become much more important and companies will throw away nothing about their customers because it will be so valuable. If you’re not doing this, you’re out of business”

- “The latest strategic weapon for companies is analytical decision making (e.g., data mining) …”
Database vs. Datawarehouse

What is Data Mining?

• **Data mining** – the process of analyzing data to extract information (unknown patterns) not offered by the raw data alone
• To perform data mining users need data-mining tools
  - **Data-mining tool** – uses a variety of techniques to find patterns and relationships in large volumes of information and infers rules that predict future behavior and guide decision making
  - A wide range of data mining techniques are being used by organizations to gain a better understanding of their customers and their operations and to solve complex organizational problems.
• An example
  - **Grocery Store in UK (see next slide)**

CRM and Data Mining (BI) Example

• A Grocery store in U.K. with the following **“patterns”** found:
  - Every Thursday afternoon
  - Young Fathers (why?) shopping at store
  - Two of the followings are always included in their shopping list
    - **Diapers** and
    - **Beers**
• What other decisions should be made as a store manager (in terms of store layout)?
  - Short term vs. Long term
    - This is an example of **cross-selling**
    - Other types of promotion: **up-sell, bundled-sell**
  - (e.g., BI helps to find valuable information then decision makers make a timely/right decision for improving/creating competitive advantages.)

Q/A

• Can the “patterns” in the grocery store example be produced from its **Database**?
  - **Y/N**
  - Why?
  - It only can be produced from its **“Data Warehouse”** using a kind of “**data mining**” software.

Why Data Mining?

• More intense competition at the global scale
• Recognition of the value in data sources
• Availability of quality data on customers, vendors, transactions, Web, etc.
• Consolidation and integration of data repositories into data warehouses
• The exponential increase in data processing and storage capabilities; and decrease in cost
• Movement toward conversion of information resources into nonphysical form
Definitions of Data Mining – Summary

- **Technical** def.: a process that uses statistical, mathematical, and artificial intelligence techniques to extract and identify useful information and subsequent knowledge (or patterns) from large sets of data.
- **General** def.: the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases - Fayyad et al. (1996).
- Keywords in this definition: Process, nontrivial, valid, novel, potentially useful, understandable
- Other names: knowledge extraction, pattern analysis, knowledge discovery, information harvesting, pattern searching, data dredging.

Data Mining

**Characteristics/Objectives**

- Source of data for DM is often a consolidated data warehouse (not always!).
- DM environment is usually a client-server or a Web-based information systems architecture.
- Data is the most critical ingredient for DM which may include soft/unstructured data.
- The miner is often an end user.
- Striking it rich requires creative thinking.
- Data mining tools’ capabilities and ease of use are essential (Web, Parallel processing, etc.).

What Does DM Do? How Does it Work?

- DM builds models to extract/identify patterns from data
  - Pattern?
    - A mathematical (numeric and/or symbolic) relationship among data items
- Types of patterns
  - 1. **Prediction**
  - 2. **Association**
  - 3. **Cluster** (segmentation)
  - 4. **Sequential** (or time series) relationships

Data in Data Mining

- Data: a collection of facts usually obtained as the result of experiences, observations, or experiments.
- Data may consist of numbers, images, …
- Data: lowest level of abstraction (from which information and knowledge are derived).

Fig 4.3 A Simple Taxonomy for Data Mining Tasks

<table>
<thead>
<tr>
<th>Data Mining</th>
<th>Learning Method</th>
<th>Popular Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction</td>
<td>Supervised</td>
<td>Decision Trees, SVM</td>
</tr>
<tr>
<td>Classification</td>
<td>Supervised</td>
<td>Decision Trees, SVM</td>
</tr>
<tr>
<td>Regression</td>
<td>Supervised</td>
<td>Decision Trees, SVM</td>
</tr>
<tr>
<td>Association</td>
<td>Supervised</td>
<td>Decision Trees, SVM</td>
</tr>
<tr>
<td>Cluster</td>
<td>Supervised</td>
<td>Decision Trees, SVM</td>
</tr>
<tr>
<td>Sequential</td>
<td>Supervised</td>
<td>Decision Trees, SVM</td>
</tr>
</tbody>
</table>

Fig 4.4 Data Mining at the Intersection of Many Disciplines

- **DATA MINING**
- **Subject**
- **Pattern Recognition**
- **Machine Learning**
- **Mathematical Modeling**
- **Databases**
- **Management Science & Information Systems**
- **Pattern Recognition**
- **Machine Learning**
- **Mathematical Modeling**
**What are typical data-mining applications?**

- Data mining is an automated process of discovery and extraction of hidden and/or unexpected patterns of collected data in order to create models for decision making that predict future behavior based on analyses of past activity.
- There are two types of data-mining techniques:
  - **Unsupervised** data-mining characteristics:
    - No model or hypothesis exists before running the analysis (with "training data" set)
    - Analysts create a hypothesis after analysis is completed
    - Cluster analysis, a common technique in this category groups entities together that have similar characteristics
  - **Supervised** data-mining characteristics:
    - Analysts develop a model prior to their analysis
    - Classification and Decision tree
    - Apply statistical techniques to estimate parameters of a model
    - Regression analysis is a technique in this category that measures the impact of a set of variables on another variable
    - Neural networks predict values and make classifications.
    - Used for making predictions

---

**Decision Tree Example for MIS Classes (hypothetical data)**

- A decision tree is a hierarchical arrangement of criteria that predicts a classification or value. It’s an unsupervised data-mining technique that selects the most useful attributes for classifying entities on some criterion. It uses if…then rules in the decision process. Here are two examples.

---

**Market-Basket Analysis** is a supervised data-mining tool for determining sales patterns. It helps businesses create cross-selling opportunities (i.e., buying relevant products together). Two terms used with this type of analysis are:

- Support: the probability that two items will be purchased together (e.g., Fins and Mask will be purchased together)
- Confidence: a conditional probability estimate (e.g., proportion of the customers who bought a mask also bought fins)
- Lift: ratio of confidence to the base probability (e.g., ratio between customers of buying fins after buying mask, and those buying fins of walking into the store). The more the lift value is close to 2 this more independent between these items (e.g., unrelated).
Other Market Basket Application

- 75% of customers who buy Coke also by corn chips
  - Good for CRM analysis

Data Mining Applications (cont.)

- Computer hardware and software
- Science and engineering
- Government and defense
- Homeland security and law enforcement
- Travel industry
- Healthcare
- Medicine
- Entertainment industry
- Sports
- Etc.

Other Data Mining Tasks

- These are in addition to the primary DM tasks (prediction, association, clustering)
  - Time-series forecasting
    - Data can be used to develop models to extrapolate the future values of the same phenomenon.
    - Part of sequence or link analysis?
  - Visualization
    - Another data mining task that can be used with other DM techniques to gain a clearer understanding of underlying relationships.
  - Types of DM
    - Hypothesis-driven data mining
      - begin with a proposition by the user.
    - Discovery-driven data mining
      - Uncover hidden patterns, associations, relationships by different viewpoint.

Data Mining Process

- A manifestation of best practices
- A systematic way to conduct DM projects
- Different groups have different versions
- Most common standard processes:
  - 1. CRISP-DM (Cross-Industry Standard Process for Data Mining)
  - 2. SEMMA (Sample, Explore, Modify, Model, and Assess)
  - 3. KDD (Knowledge Discovery in Databases)

Data Mining Applications

- Customer Relationship Management (CRM)
  - Grocery example in UK
- Banking & Other Financial
- Retailing and Logistics
- Manufacturing and Maintenance
- Brokerage and Securities Trading
- Insurance

What are typical data-mining applications?

<table>
<thead>
<tr>
<th>DM Capabilities</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Association (Unsupervised): Association between items</td>
<td>Market Basket Analysis: 75% of customers who buy Coke also buy corn chips (good for CRM analysis)</td>
</tr>
<tr>
<td>Clustering (Unsupervised): Grouping items according to statistical similarities</td>
<td>Customer Segmentation: Meals charged on a business-issued gold card are typically purchased on weekdays and include a mean value of greater than $250, whereas meals purchased using a personal platinum card are predominantly on weekends, have a mean value of $175 and include a bottle of wine more than 65% of the time.</td>
</tr>
<tr>
<td>Sequence/Temporal Pattern (Supervised): Time-based affinity (Statistical Analysis)</td>
<td>Decision Tree Analysis (Supervised Segmentation): Customers with excellent credit history have a debt/equity ratio of less than 10%.</td>
</tr>
<tr>
<td>Classification (Supervised): Assigns new records to existing classes</td>
<td>Other than sequence or link analysis?</td>
</tr>
<tr>
<td>Visualization</td>
<td>Another data mining task that can be used with other DM techniques to gain a clearer understanding of underlying relationships.</td>
</tr>
</tbody>
</table>
Dr. Chen, Business Intelligence

1. Data Mining Process: CRISP-DM

- **Step 1: Business Understanding**
- **Step 2: Data Understanding**
- **Step 3: Data Preparation (!)**
- **Step 4: Model Building**
- **Step 5: Testing and Evaluation**
- **Step 6: Deployment**

• The process is highly repetitive and experimental (DM: art versus science?)

2. Data Mining Process: SEMMA

- **Sample** (Generate a representative sample of the data)
- **Explore** (Classification methods: description of the data)
- **Assess** (Evaluate the accuracy and confidence of the models)
- **Model** (Use a variety of statistical and machine learning methods)
- **Modify** (Identify variables, transform variable representations)

3. Knowledge Discovery in Databases (KDD)

- KDD is a process of using data mining methods to find useful information and patterns in the data, as opposed to data mining, which involves using algorithms to identify patterns in data derived through the KDDS PROCESS.
- It is a comprehensive process that encompasses data mining.
- KDD consists of the following steps (Dunham (2003)):
  1. Data selection,
  2. Data preprocessing,
  3. Data transformation,
  4. Data mining, and
  5. Interpretation/evaluation.
Data Mining Methods: Classification

- Most frequently used DM method
- Part of the machine-learning family
- Employ supervised learning
- Learn from past data, classify new data
- The output variable is categorical (nominal or ordinal) in nature

- Classification versus regression?
- Classification versus clustering?

Assessment Methods for Classification

- Predictive accuracy
  - Hit rate
- Speed
  - Model building; predicting
- Robustness
- Scalability
- Interpretability
  - Transparency, explainability

Accuracy of Classification Models

- In classification problems, the primary source for accuracy estimation is the confusion matrix

<table>
<thead>
<tr>
<th>True Class</th>
<th>Predicted Class</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False Positive (FN)</td>
<td>True Negative (TN)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Accuracy = \( \frac{TP + TN}{TP + TN + FP + FN} \)

True Positive Rate = \( \frac{TP}{TP + FP} \)

True Negative Rate = \( \frac{TN}{TN + FP} \)

Precision = \( \frac{TP}{TP + FP} \)

Recall = \( \frac{TP}{TP + FN} \)

Estimation Methodologies for Classification

- Simple split (or holdout or test sample estimation)
  - Split the data into 2 mutually exclusive sets training (~70%) and testing (30%)

- Other estimation methodologies
  - Leave-one-out, bootstrapping, jackknifing
  - Area under the ROC curve

Estimation Methodologies for Classification

- k-Fold Cross Validation (rotation estimation)
  - Split the data into k mutually exclusive subsets
  - Use each subset as testing while using the rest of the subsets as training
  - Repeat the experimentation for k times
  - Aggregate the test results for true estimation of prediction accuracy training

- Area under the ROC curve

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \\
\text{True Positive Rate} = \frac{TP}{TP + FP} \\
\text{True Negative Rate} = \frac{TN}{TN + FP} \\
\text{Precision} = \frac{TP}{TP + FP} \\
\text{Recall} = \frac{TP}{TP + FN}
\]
Classification Techniques

- Decision tree analysis
- Statistical analysis
- Neural networks
- Support vector machines
- Case-based reasoning
- Bayesian classifiers
- Genetic algorithms
- Rough sets

Decision Trees

- Employs the divide and conquer method
- Recursively divides a training set until each division consists of examples from one class

A general algorithm for decision tree building

1. Create a root node and assign all of the training data to it.
2. Select the best splitting attribute.
3. Add a branch to the root node for each value of the split. Split the data into mutually exclusive subsets along the lines of the specific split.
4. Repeat the steps 2 and 3 for each and every leaf node until the stopping criteria is reached.

Decision Trees

- DT algorithms mainly differ on
  - Splitting criteria
    - Which variable to split first?
    - What values to use to split?
    - How many splits to form for each node?
  - Stopping criteria
    - When to stop building the tree
  - Pruning (generalization method)
    - Pre-pruning versus post-pruning
- Most popular DT algorithms include
  - ID3, C4.5, C5; CART; CHAID; M5

Decision Trees

- Alternative splitting criteria
  - Gini index determines the purity of a specific class as a result of a decision to branch along a particular attribute/value
    - Used in CART
  - Information gain uses entropy to measure the extent of uncertainty or randomness of a particular attribute/value split
    - Used in ID3, C4.5, C5
  - Chi-square statistics (used in CHAID)

Cluster Analysis for Data Mining

- Used for automatic identification of natural groupings of things
- Part of the machine-learning family
- Employs unsupervised learning
- Learns the clusters of things from past data, then assigns new instances
- There is not an output variable
- Also known as segmentation

Cluster Analysis for Data Mining

- Clustering results may be used to
  - Identify natural groupings of customers
  - Identify rules for assigning new cases to classes for targeting/diagnostic purposes
  - Provide characterization, definition, labeling of populations
  - Decrease the size and complexity of problems for other data mining methods
  - Identify outliers in a specific domain (e.g., rare-event detection)
Cluster Analysis for Data Mining

- Analysis methods
  - Statistical methods (including both hierarchical and nonhierarchical), such as k-means, k-modes, and so on
  - Neural networks (adaptive resonance theory [ART], self-organizing map [SOM])
  - Fuzzy logic (e.g., fuzzy c-means algorithm)
  - Genetic algorithms

Cluster Analysis for Data Mining

- How many clusters?
  - There is no “truly optimal” way to calculate it
  - Heuristics are often used
    - Look at the sparseness of clusters
    - Number of clusters = \((n/2)^{1/2}\) (n: no of data points)
    - Use Akaike information criterion (AIC)
    - Use Bayesian information criterion (BIC)
  - Most cluster analysis methods involve the use of a distance measure to calculate the closeness between pairs of items.
    - Euclidian versus Manhattan (rectilinear) distance

Cluster Analysis for Data Mining

- k-Means Clustering Algorithm
  - k: pre-determined number of clusters
  - Algorithm (Step 0: determine value of k)
    - Step 1: Randomly generate k random points as initial cluster centers.
    - Step 2: Assign each point to the nearest cluster center.
    - Step 3: Re-compute the new cluster centers.
    - Repetition step: Repeat steps 3 and 4 until some convergence criterion is met (usually that the assignment of points to clusters becomes stable).

Cluster Analysis for Data Mining

- Association Rule Mining
  - A very popular DM method in business
  - Finds interesting relationships (affinities) between variables (items or events)
  - Part of machine learning family
  - Employs unsupervised learning
  - There is no output variable
  - Also known as market basket analysis
  - Often used as an example to describe DM to ordinary people, such as the famous “relationship between diapers and beers!”

Association Rule Mining

- Input: the simple point-of-sale transaction data
- Output: Most frequent affinities among items
- Example: according to the transaction data…
  “Customer who bought a laptop computer and a virus protection software, also bought extended service plan 70 percent of the time”
- How do you use such a pattern/knowledge?
  - Put the items next to each other for ease of finding
  - Promote the items as a package (do not put one on sale if the other(s) are on sale)
  - Place items far apart from each other so that the customer has to walk the aisles to search for it, and by doing so potentially see and buy other items
Association Rule Mining

- A representative application of association rule mining includes:
  - **In business**: cross-marketing, cross-selling, store design, catalog design, e-commerce site design, optimization of online advertising, product pricing, and sales/promotion configuration
  - **In medicine**: relationships between symptoms and illnesses; diagnosis and patient characteristics and treatments (to be used in medical DSS); and genes and their functions (to be used in genomics projects)
- ...
Application Case 4.6

Data Mining Goes to Hollywood: Predicting Financial Success of Movies

Questions for Discussion
- Decision situation
- Problem
- Proposed solution
- Results
- Answer & discuss the case questions.

Elements/Concepts of ANN
- Processing element (PE)
- Information processing
- Network structure
  - Feedforward vs. recurrent vs. multi-layer...
- Learning parameters
  - Supervised/unsupervised, backpropagation, learning rate, momentum
- ANN Software – NN shells, integrated modules in comprehensive DM software, ...

Data Mining Software
- Commercial
  - IBM SPSS Modeler (formerly Clementine)
  - SAS - Enterprise Miner
  - IBM - Intelligent Miner
  - StatSoft – Statistica Data Miner
  - ... many more
- Free and/or Open Source
  - R
  - RapidMiner
  - Weka...

Big Data Software Tools and Platforms

Application Case 4.6
Data Mining Goes to Hollywood!

Questions for Discussion
- Decision situation
- Problem
- Proposed solution
- Results
- Answer & discuss the case questions.

Elements/Concepts of ANN
- Processing element (PE)
- Information processing
- Network structure
  - Feedforward vs. recurrent vs. multi-layer...
- Learning parameters
  - Supervised/unsupervised, backpropagation, learning rate, momentum
- ANN Software – NN shells, integrated modules in comprehensive DM software, ...

Data Mining Software
- Commercial
  - IBM SPSS Modeler (formerly Clementine)
  - SAS - Enterprise Miner
  - IBM - Intelligent Miner
  - StatSoft – Statistica Data Miner
  - ... many more
- Free and/or Open Source
  - R
  - RapidMiner
  - Weka...

Big Data Software Tools and Platforms

Application Case 4.6
Data Mining Goes to Hollywood!

Questions for Discussion
- Decision situation
- Problem
- Proposed solution
- Results
- Answer & discuss the case questions.
• 1. Why is it important for many Hollywood professionals to predict the financial success of movies?

• The movie industry is the “land of hunches and wild guesses” due to the difficulty associated with forecasting product demand, making the movie business in Hollywood a risky endeavor.

• If Hollywood could better predict financial success, this would mitigate some of the financial risk.

• 2. How can data mining be used for predicting financial success of movies before the start of their production process?

• The way Sharda and Delen did it was to use data from movies between 1998 and 2005 as training data, and movies of 2006 as test data. They applied individual and ensemble prediction models, and were able to identify significant variables impacting financial success.

• They also showed that by using sensitivity analysis, decision makers can predict with fairly high accuracy how much value a specific actor (or a specific release date, or the addition of more technical effects, etc.) brings to the financial success of a film, making the underlying system an invaluable decision aid.

• 3. How do you think Hollywood did, and perhaps still is performing, this task without the help of data mining tools and techniques?

• Most is done by gut feel and trial-and-error. This may keep the movie business as a financially risky endeavor, but also allows for creativity. Sometimes uncertainty is a good thing.

Data Mining Myths

• Data mining …
  ➢ provides instant solutions/predictions
  ➢ is not yet viable for business applications
  ➢ requires a separate, dedicated database
  ➢ can only be done by those with advanced degrees
  ➢ is only for large firms that have lots of customer data
  ➢ is another name for the good-old statistics
Data Mining Myths

Data mining is a powerful analytical tool that enables business executives to advance from describing the nature of the past to predicting the future. DM’s myths and realities are listed below:

<table>
<thead>
<tr>
<th>Myth</th>
<th>Reality</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM provides instant, crystal-ball-like solutions/predictions</td>
<td>DM is a multi-step process that requires deliberate, proactive design and use.</td>
</tr>
<tr>
<td>DM is not yet viable for business applications</td>
<td>The current state of the art is ready to go for almost any business.</td>
</tr>
<tr>
<td>DM requires a separate, dedicated database.</td>
<td>Because of advances in database technology, a dedicated database is not required, even though it may be desirable.</td>
</tr>
<tr>
<td>DM is only for large firms that have lots of customer data.</td>
<td>Newer Web-based tools enable managers at all educational levels to do data mining.</td>
</tr>
<tr>
<td>DM can only be done by those with advanced degrees.</td>
<td>If the data accurately reflect the business or its customers, a company can use data mining.</td>
</tr>
<tr>
<td>DM is another name for the good-old statistics.</td>
<td>DM is now available for easy use.</td>
</tr>
</tbody>
</table>

Common Data Mining Blunders

1. Selecting the wrong problem for data mining
2. Ignoring what your sponsor thinks data mining is and what it really can/cannot do
3. Not leaving sufficient time for data acquisition, selection, and preparation
4. Looking only at aggregated results and not at individual records/predictions
5. Being sloppy about keeping track of the data mining procedure and results
6. …more in the book

Extra Example on Correlation

**Business Scenario:**

Sarah is a regional sales manager for a nationwide supplier of fossil fuels for home heating. Recent volatility in market prices for heating oil specifically, coupled with wide variability in the size of each order for home heating oil, has Sarah concerned. She feels a need to understand the types of behaviors and other factors that may influence the demand for heating oil in the domestic market. What factors are related to heating oil usage, and how might she use a knowledge of such factors to better manage her inventory, and anticipate demand? Sarah believes that data mining can help her begin to formulate an understanding of these factors and interactions.

The following attributes are included in the data set:

- **Insulation:** This is a density rating, ranging from one to ten, indicating the thickness of each home’s insulation. A home with a density rating of one is poorly insulated, while a home with a density of ten has excellent insulation.
- **Temperature:** This is the average outdoor ambient temperature at each home for the most recent year, measured in degree Fahrenheit.
- **Heating_Oil:** This is the total number of units of heating oil purchased by the owner of each home in the most recent year.
- **Num_Occupants:** This is the total number of occupants living in each home.
- **Avg_Age:** This is the average age of those occupants.
- **Home_Size:** This is a rating, on a scale of one to eight, of the home’s overall size. The higher the number, the larger the home.

The following are included in the data set:

1. **File**  
   2. **New Process**  
   3. **Read CSV**  
   4. **Chapter04DataSet**

Step 2. Change from “Semicolon” to “Comma”
Dr. Chen,

Business Intelligence

The “Read CSV” with input data has been added to the process.

Next, add “correlation matrix” operator and connect all ports as shown. Click “Run”

The result is obtained (with “Meta Data View” shown.)

Result:

What do we obtain from this result?
We learned through our investigation, that the two most strongly correlated attributes in our data set are **Heating_Oil (or Insulation)** and **Avg_Age**, with a coefficient of **0.548** (or **0.643**). Thus, we know that in this data set, as the average age of the occupants in a home increases, so too does the heating oil usage in that home.

What we do not know is why that occurs. Data analysts often make the mistake of interpreting correlation as causation.

The assumption that correlation proves causation is dangerous and often false.

**Summary on Correlation**

- Correlation can be a quick and easy way to see how elements of a given problem may be interacting with one another. Whenever you find yourself asking how certain factors in a problem you’re trying to solve interact with one another, consider building a correlation matrix to find out.

  - For example,
    - 1) does customer satisfaction change based on time of year?
    - 2) Does the amount of rainfall change the price of a crop?
    - 3) Does household income influence which restaurants a person patronizes?

  - The answer to each of these questions is probably ‘yes’, but correlation can not only help us know if that’s true, it can also help us learn how strongly the interactions are when, and if, they occur.

• Questions, comments