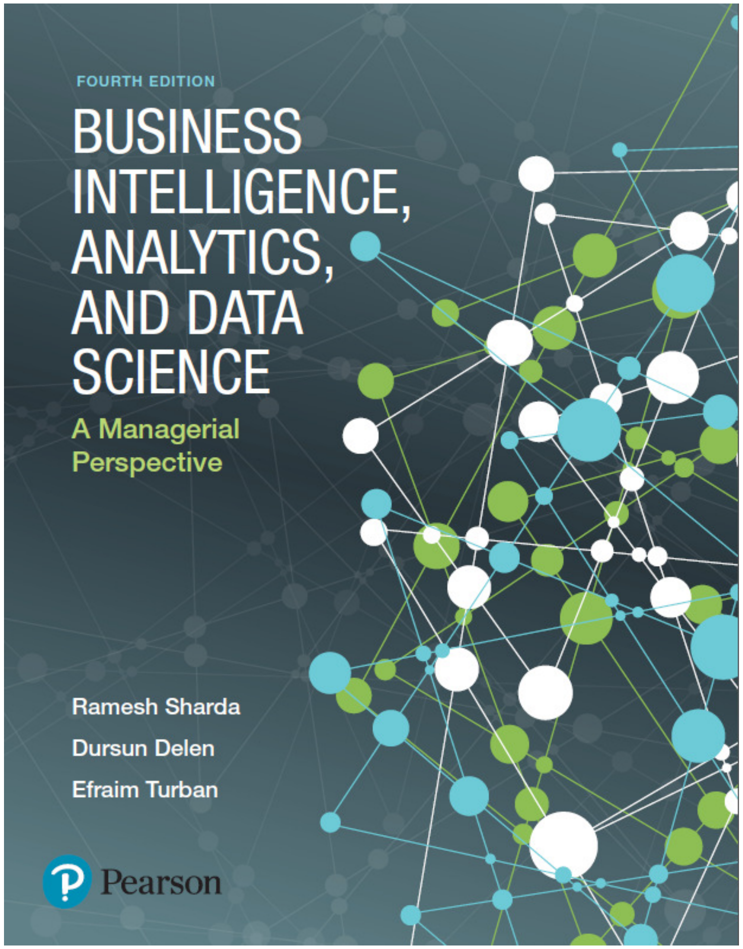
Business Intelligence, Analytics, and Data Science: A Managerial Perspective

Fourth Edition



Chapter 4 – Part C

Predictive Analytics I: Data Mining Process, Methods, and Algorithms

Pearson

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 versus predicting/usage speed
- Robustness
- Scalability
- Interpretability
  - Transparency, explainability

Accuracy of Classification Models

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

		True/Observed Class	
		Positive	Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
	Negative	False Negative Count (FN)	True Negative Count (TN)

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

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

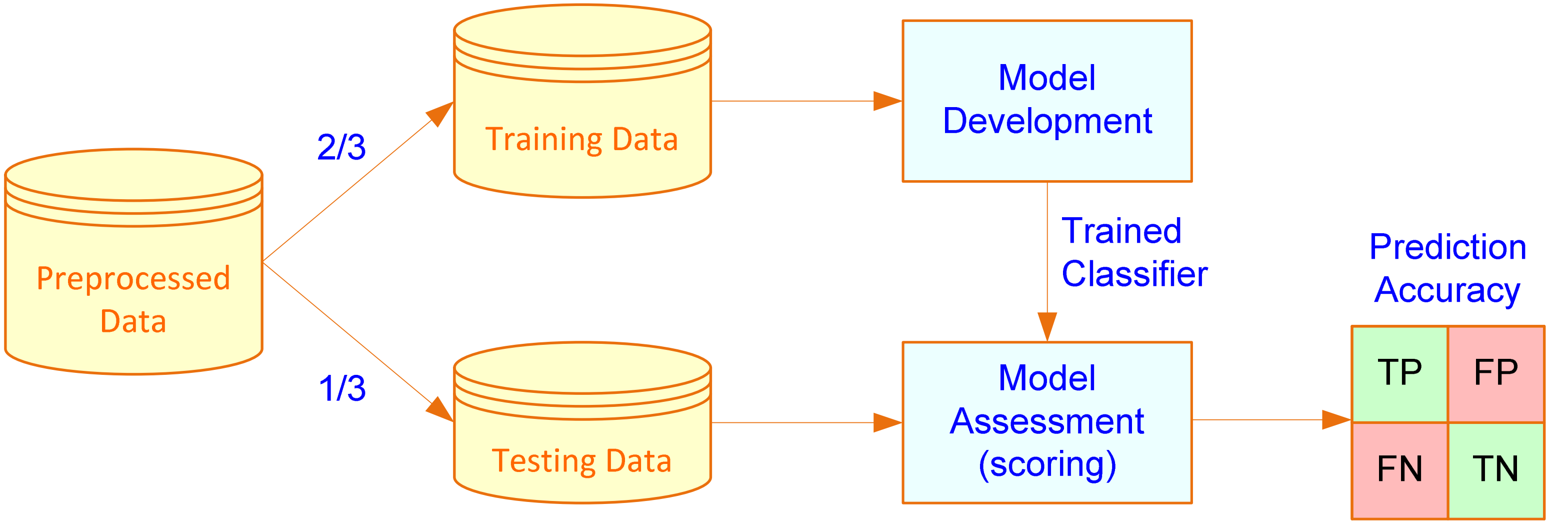
$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Pearson

Estimation Methodologies for Classification: Single/Simple Split

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



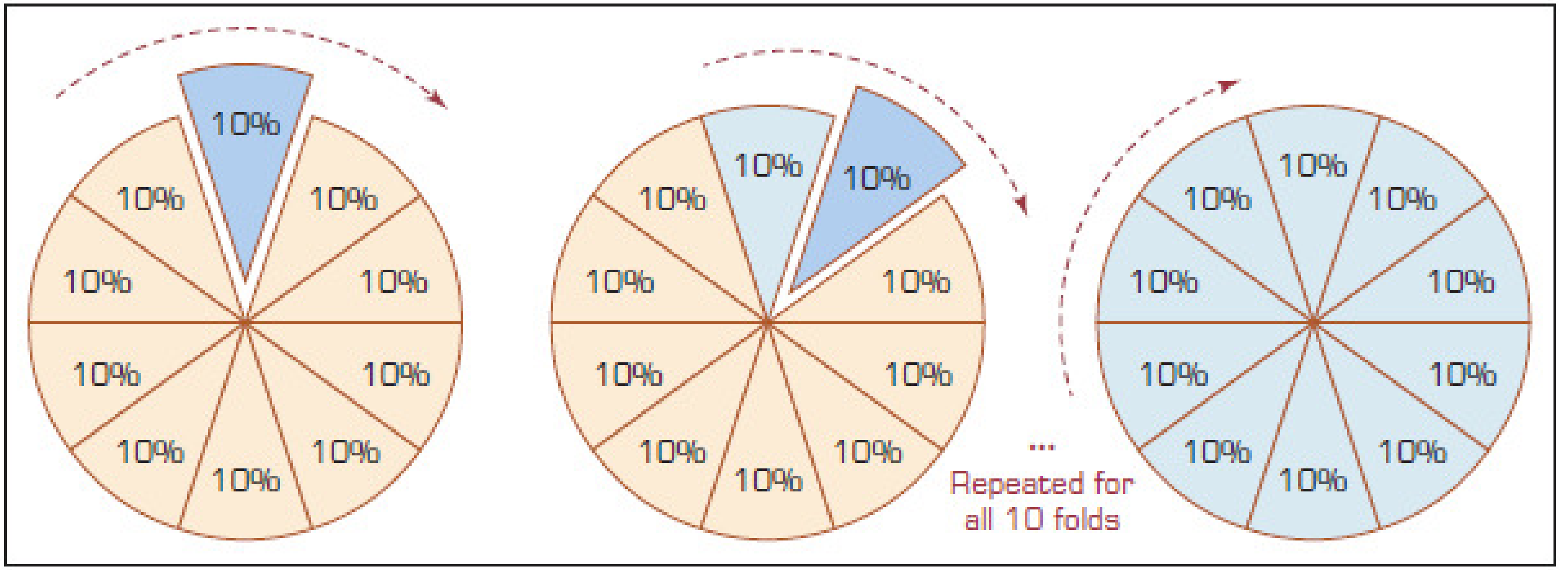
TP	FP
FN	TN

  - For Neural Networks, the data is split into three sub-sets (training [~60%], validation [~20%], testing [~20%])

Pearson

Estimation Methodologies for Classification: *k*-Fold Cross Validation (rotation estimation)

- Data is split into *k* mutual subsets and *k* number training/testing experiments are conducted



- FIGURE 4.10
  A Graphical Depiction of k-Fold Cross-Validation

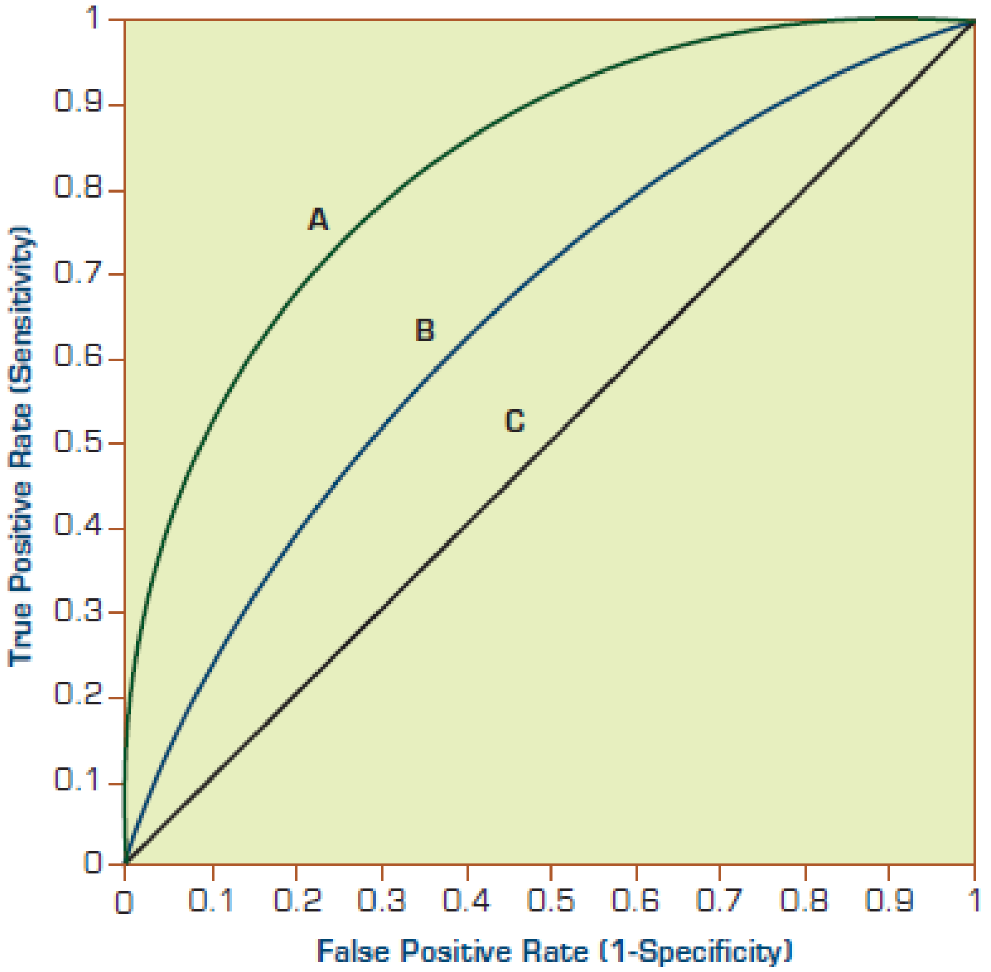
Additional Estimation Methodologies for Classification

- Leave-one-out
  - Similar to *k*-fold where *k* = number of samples
- Bootstrapping
  - Random sampling with replacement
- Jackknifing
  - Similar to leave-one-out
- Area Under the ROC Curve (AUC)
  - ROC: receiver operating characteristics (a term borrowed from radar image processing)

Area Under the ROC Curve (AUC) (1 of 2)

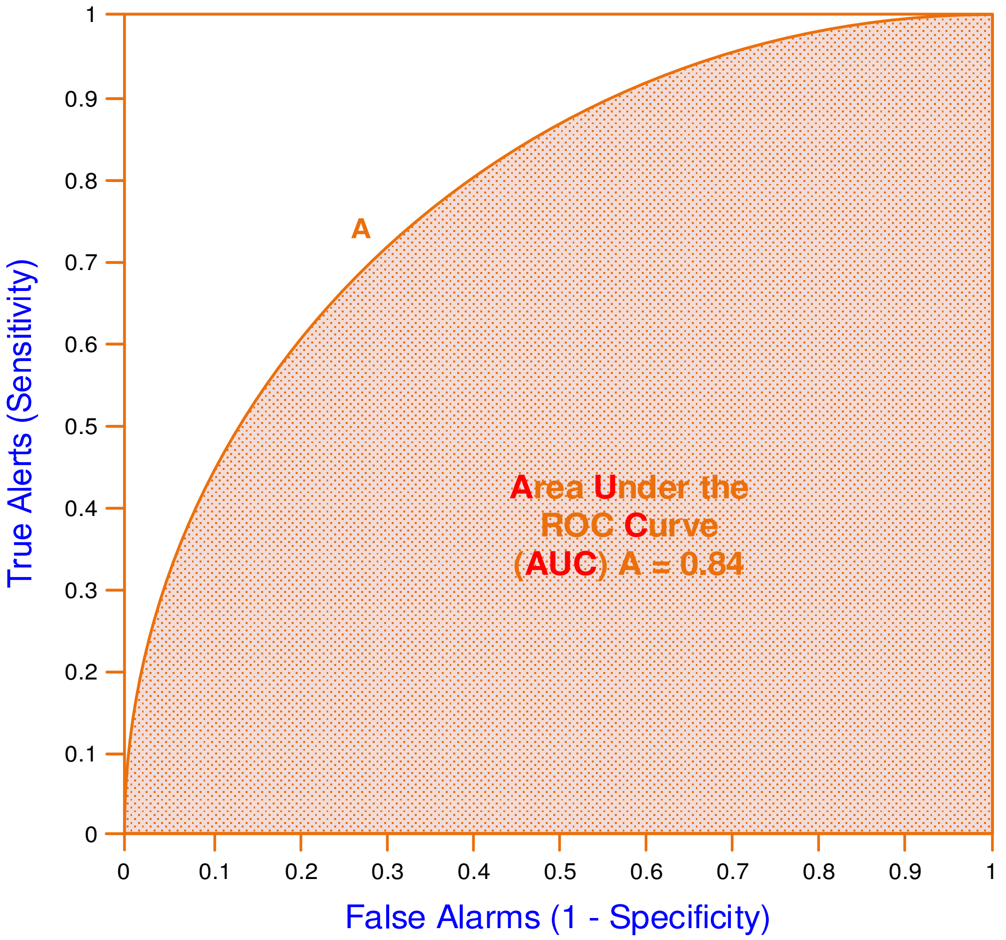
- Works with binary classification

- FIGURE 4.11 A Sample ROC Curve



Area Under the ROC Curve (AUC) (2 of 2)

- Produces values from 0 to 1.0
- Random chance is 0.5 and perfect classification is 1.0
- Produces a good assessment for skewed class distributions too!



Classification Techniques

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

Decision Trees

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

A general algorithm (steps) for building a decision tree

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 steps 2 and 3 for each and every leaf node until the stopping criteria is reached.

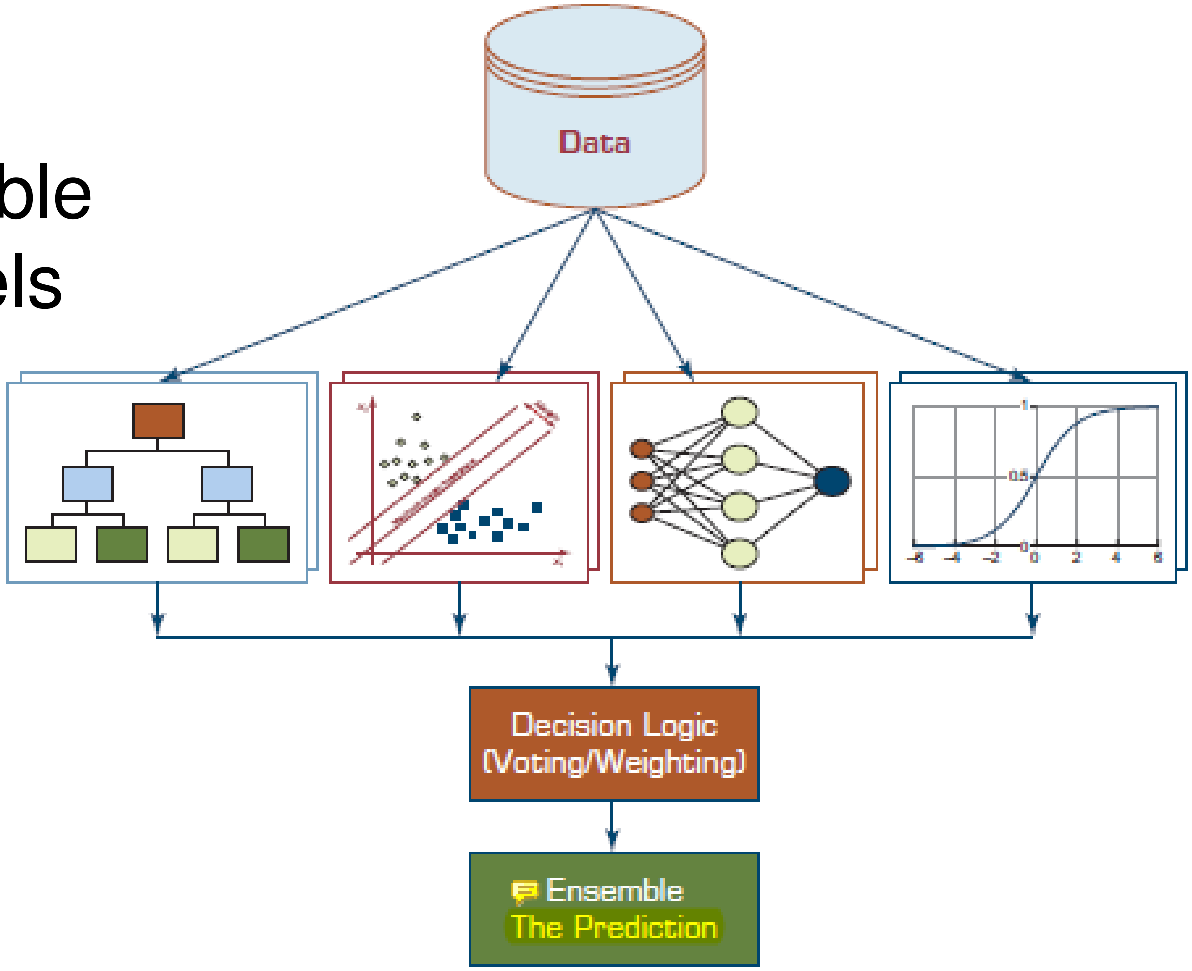
Decision Trees (1 of 2)

- DT algorithms mainly differ on
  1. Splitting criteria
     - Which variable, what value, etc.
  2. Stopping criteria
     - When to stop building the tree
  3. Pruning (generalization method)
     - Pre-pruning versus post-pruning
- Most popular DT algorithms include
  – ID3, C4.5, C5; CART; CHAID; M5

Ensemble Models for Predictive Analytics

- Produces more robust and reliable prediction models

- FIGURE 4.12 Graphical Illustration of a Heterogeneous Ensemble



Application Case 4.5

Influence Health Uses Advanced Predictive Analytics to Focus on the Factors That Really Influence People's Healthcare Decisions

Questions for Discussion

1. What did Influence Health do?
2. What were the challenges, the proposed solutions, and the obtained results?
3. How can data mining help companies in the healthcare industry (in ways other than the ones mentioned in this case)?

Cluster Analysis for Data Mining

- Used for automatic identification of natural groupings of things
- Part of the machine-learning family
- Employ unsupervised learning
- Learns the clusters of things from past data, then assigns new instances
- There is not an output/target variable
- In marketing, it is 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
- How many clusters?

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.

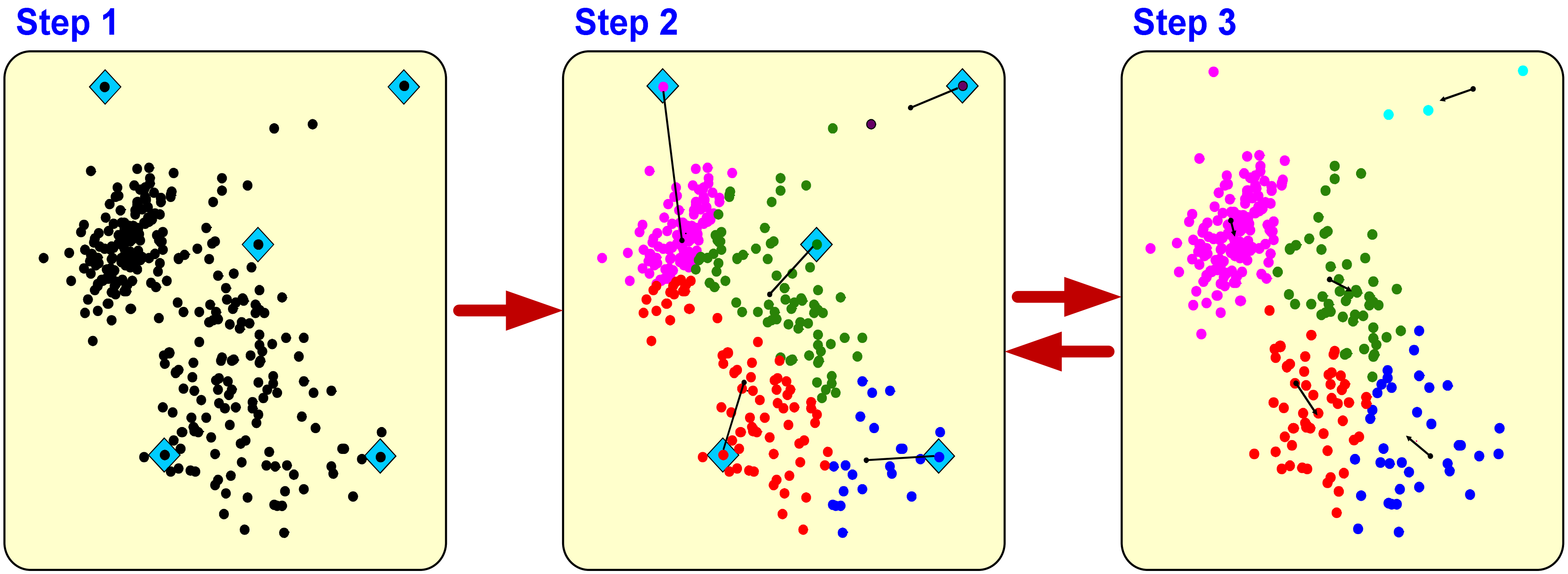
Step 2: Assign each point to the nearest cluster center.

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 - *k*-Means Clustering Algorithm

- FIGURE 4.13 A Graphical Illustration of the Steps in the *k*-Means Algorithm



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 lap-top 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
  - Promote the items as a package
  - Place items far apart from each other!

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)
  - …

Association Rule Mining

- Are all association rules interesting and useful?

A Generic Rule: **X ⇒ Y [S%, C%]**

X, Y: products and/or services

X: Left-hand-side (LHS)

Y: Right-hand-side (RHS)

S: Support: how often **X** and **Y** go together

C: Confidence: how often **Y** go together with the **X**

Example: {Laptop Computer, Antivirus Software} ⇒ {Extended Service Plan} [30%, 70%]

Association Rule Mining

- Several algorithms are developed for discovering (identifying) association rules
  - Apriori
  - Eclat
  - FP-Growth
  - + Derivatives and hybrids of the three
- The algorithms help identify the frequent itemsets, which are then converted to association rules

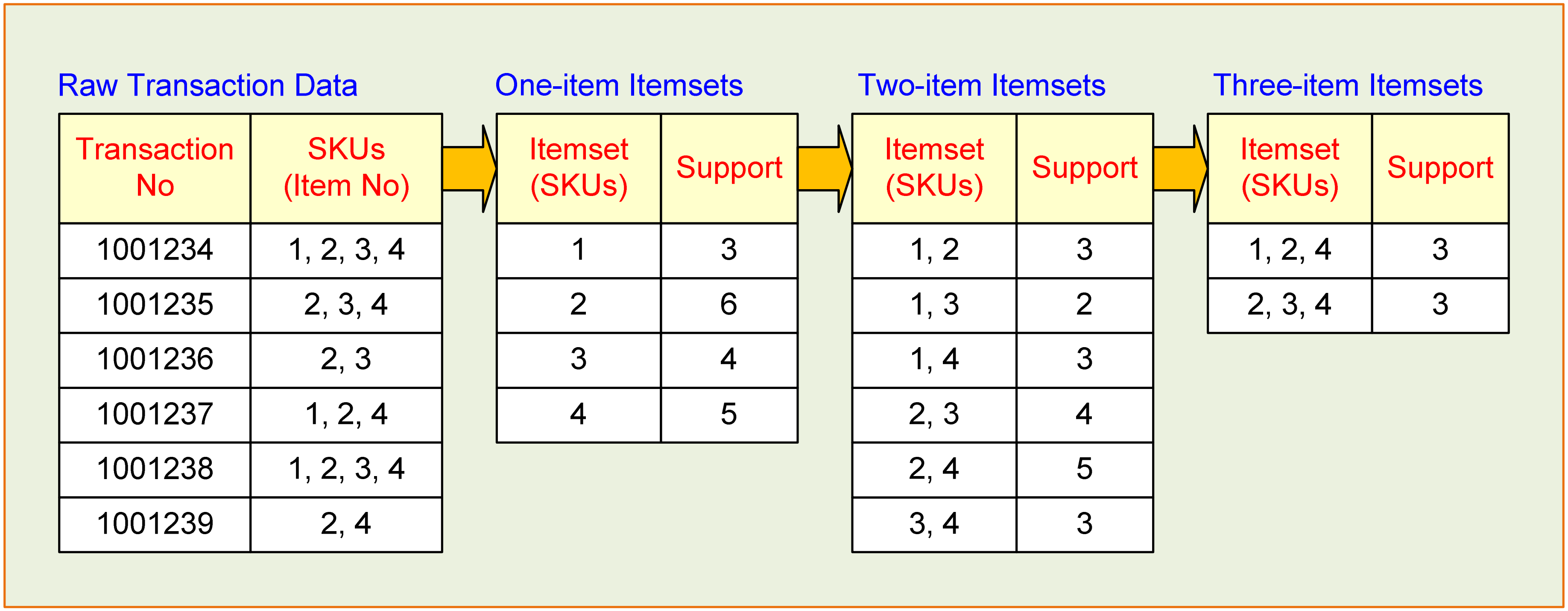
Association Rule Mining

- Apriori Algorithm
  - Finds subsets that are common to at least a minimum number of the itemsets
  - Uses a bottom-up approach
    - frequent subsets are extended one item at a time (the size of frequent subsets increases from one-item subsets to two-item subsets, then three-item subsets, and so on), and
    - groups of candidates at each level are tested against the data for minimum support

      (*see the figure*) → --

Association Rule Mining
Apriori Algorithm

- FIGURE 4.13 A Graphical Illustration of the Steps in the *k*-Means Algorithm



Raw Transaction Data

Transaction No	SKUs (Item No)
1001234	1, 2, 3, 4
1001235	2, 3, 4
1001236	2, 3
1001237	1, 2, 4
1001238	1, 2, 3, 4
1001239	2, 4

One-item Itemsets

Itemset (SKUs)	Support
1	3
2	6
3	4
4	5

Two-item Itemsets

Itemset (SKUs)	Support
1, 2	3
1, 3	2
1, 4	3
2, 3	4
2, 4	5
3, 4	3

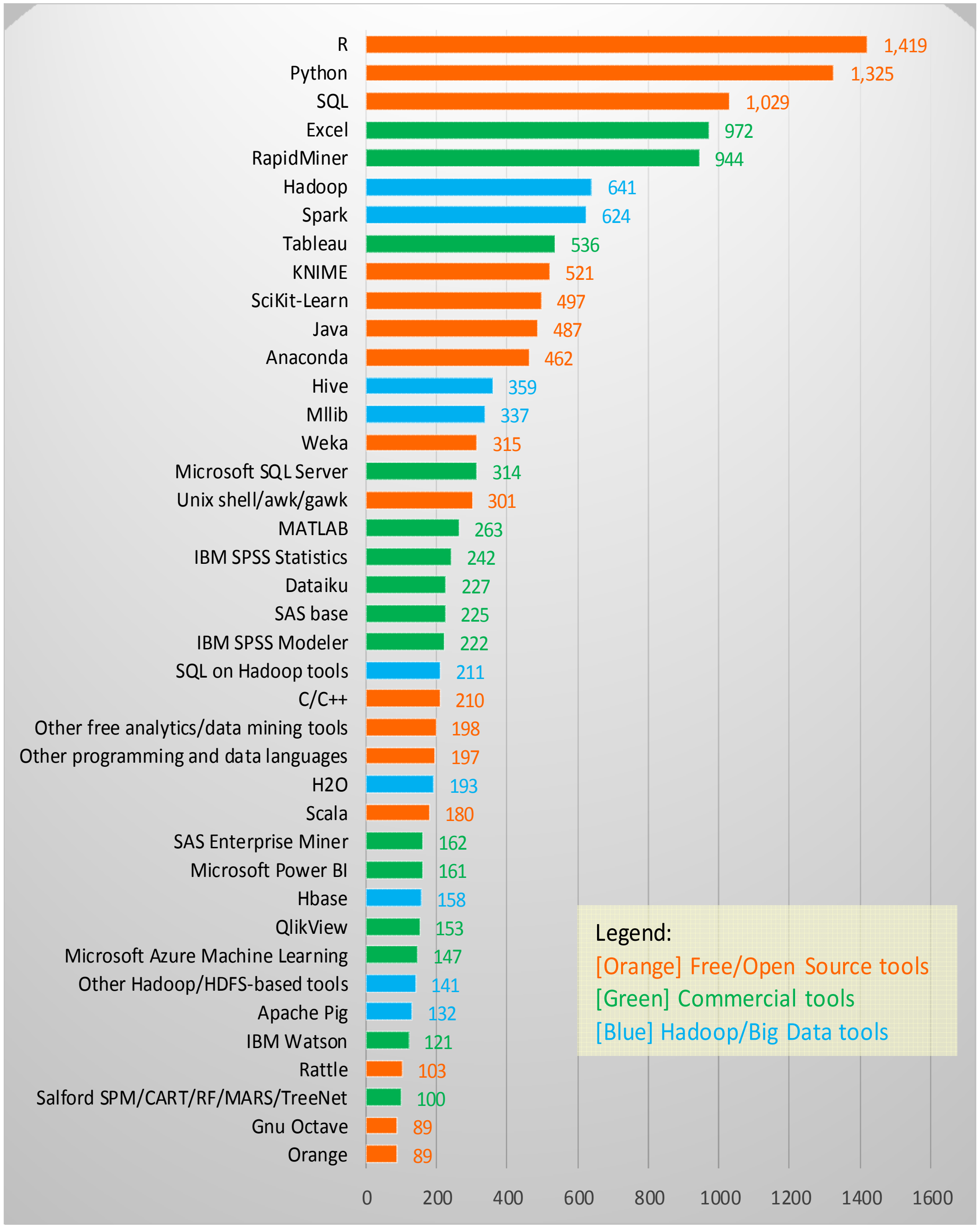
Three-item Itemsets

Itemset (SKUs)	Support
1, 2, 4	3
2, 3, 4	3

Pearson

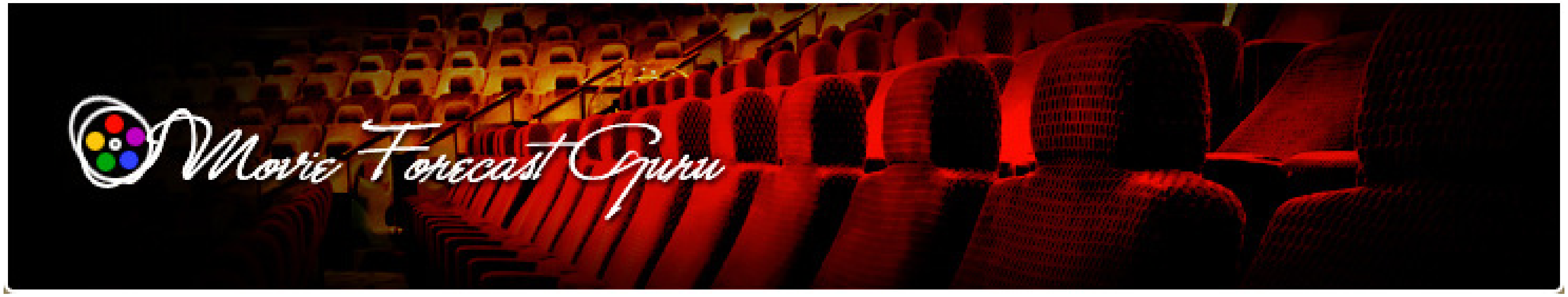
Data Mining Software Tools

- Commercial
  - IBM SPSS Modeler (formerly Clementine)
  - SAS Enterprise Miner
  - Statistica - Dell/Statsoft
  - … many more
- Free and/or Open Source
  - KNIME
  - RapidMiner
  - Weka
  - R, …



Application Case 4.6

Data Mining Goes to Hollywood: Predicting Financial Success of Movies



- Goal: Predicting financial success of Hollywood movies before the start of their production process
- How: Use of advanced predictive analytics methods
- Results: promising

Application Case 4.6

Data Mining Goes to Hollywood: Predicting Financial Success of Movies

Class No.	1	2	3	4	5	6	7	8	9
Range (in $Millions)	< 1 (Flop)	> 1 < 10	> 10 < 20	> 20 < 40	> 40 < 65	> 65 < 100	> 100 < 150	> 150 < 200	> 200 (Blockbuster)

Dependent Variable

Independent Variables

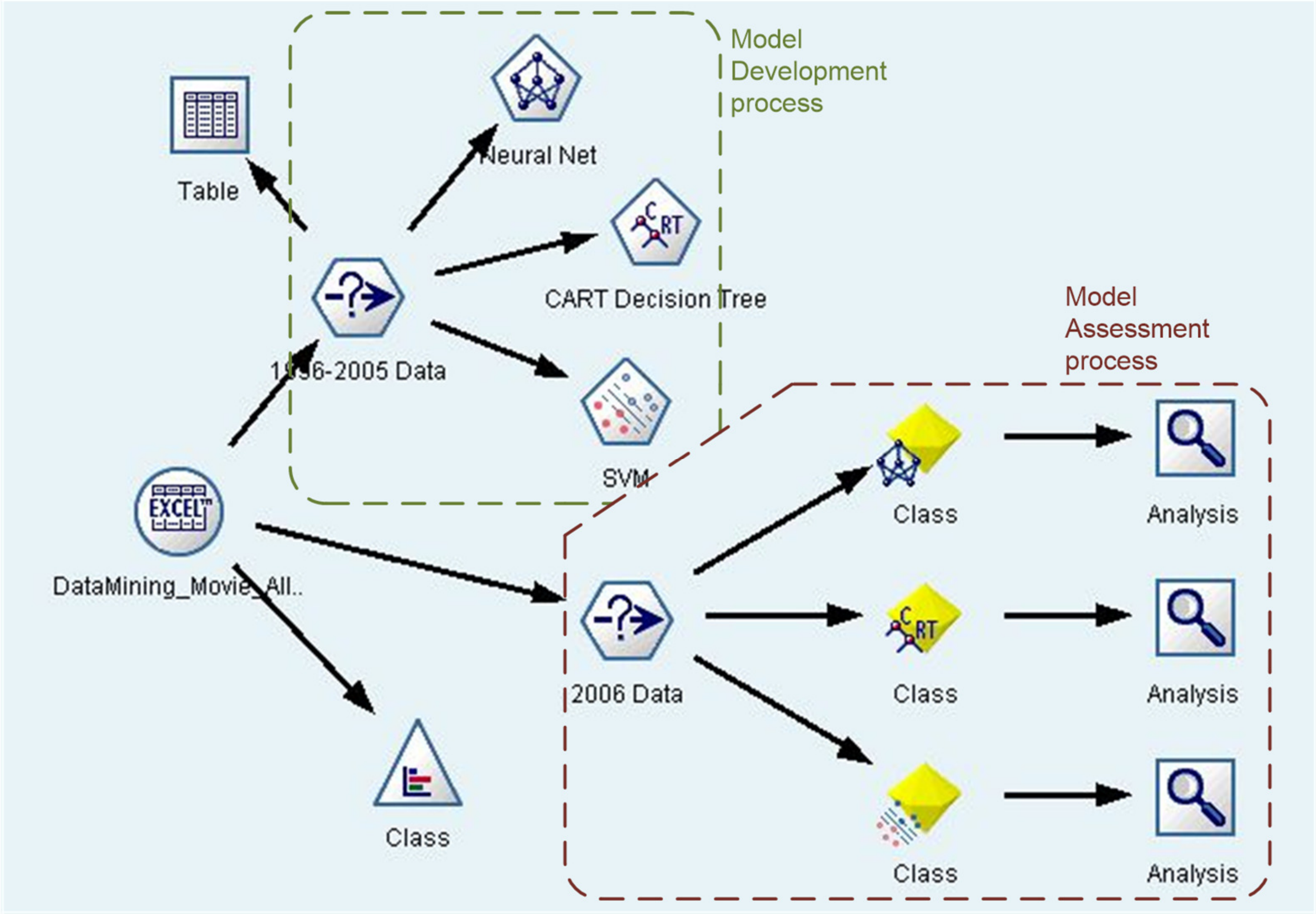
Independent Variable	Number of Values	Possible Values
MPAA Rating	5	G, PG, PG-13, R, NR
Competition	3	High, Medium, Low
Star value	3	High, Medium, Low
Genre	10	Sci-Fi, Historic Epic Drama, Modern Drama, Politically Related, Thriller, Horror, Comedy, Cartoon, Action, Documentary
Special effects	3	High, Medium, Low
Sequel	1	Yes, No
Number of screens	1	Positive integer

A Typical Classification Problem

Application Case 4.6

Data Mining Goes to Hollywood: Predicting Financial Success of Movies

The DM Process Map in IBM SPSS Modeler



Application Case 4.6

Data Mining Goes to Hollywood: Predicting Financial Success of Movies

	Prediction Models					
	Individual Models			Ensemble Models		
Performance Measure	SVM	ANN	C&RT	Random Forest	Boosted Tree	Fusion (Average)
Count (*Bingo*)	192	182	140	189	187	**194**
Count (*1-Away*)	104	120	126	121	104	**120**
Accuracy (*% Bingo*)	55.49%	52.60%	40.46%	54.62%	54.05%	**56.07%**
Accuracy (*% 1-Away*)	85.55%	87.28%	76.88%	89.60%	84.10%	**90.75%**
Standard deviation	0.93	0.87	1.05	0.76	0.84	**0.63**

* *Training set: 1998 – 2005 movies; Test set: 2006 movies*

Data Mining Myths

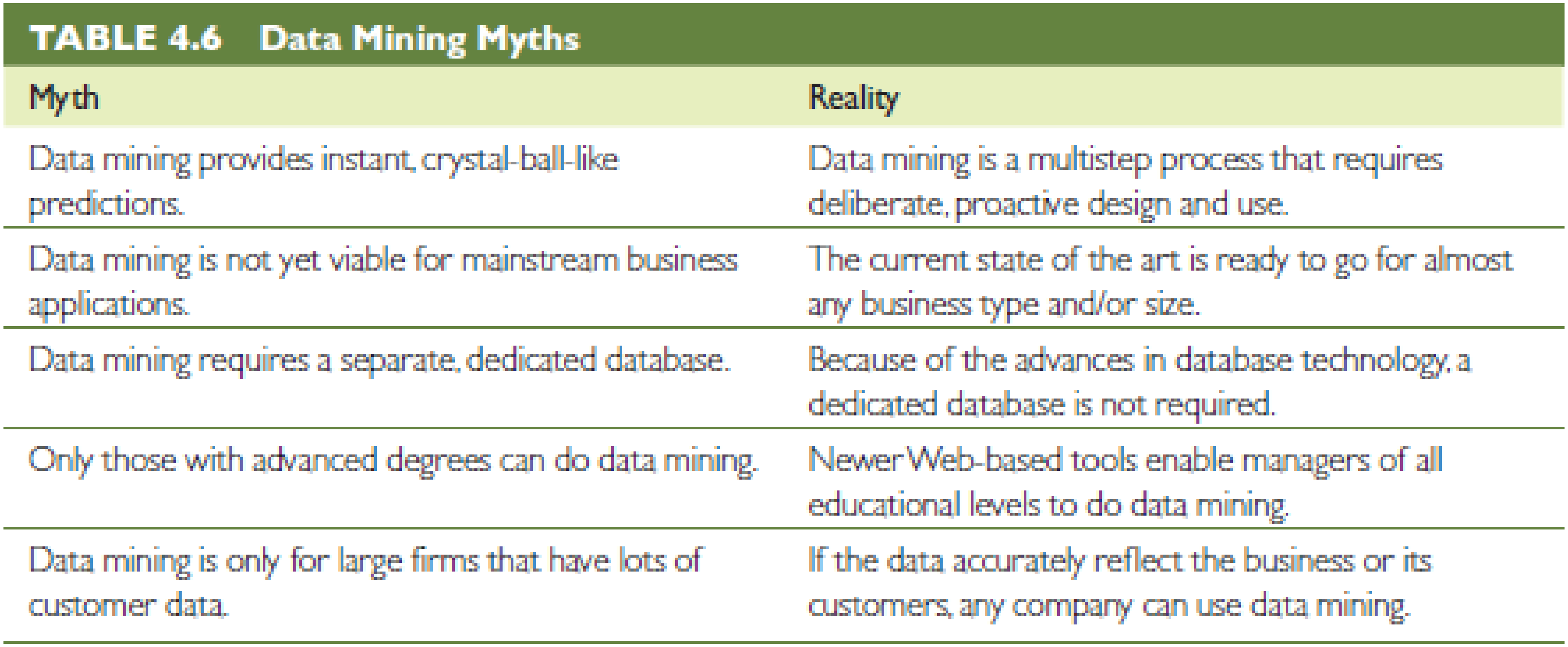


TABLE 4.6 Data Mining Myths

Myth	Reality
Data mining provides instant, crystal-ball-like predictions.	Data mining is a multistep process that requires deliberate, proactive design and use.
Data mining is not yet viable for mainstream business applications.	The current state of the art is ready to go for almost any business type and/or size.
Data mining requires a separate, dedicated database.	Because of the advances in database technology, a dedicated database is not required.
Only those with advanced degrees can do data mining.	Newer Web-based tools enable managers of all educational levels to do data mining.
Data mining is only for large firms that have lots of customer data.	If the data accurately reflect the business or its customers, any company can use data mining.

Data Mining Mistakes

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. Beginning without the end in mind
4. Not leaving sufficient time for data acquisition, selection, and preparation
5. Looking only at aggregated results and not at individual records/predictions
6. … 10 more mistakes… in your book