

Prepared by Dr. Wesam Ahmed

IT393
Business Intelligence

Lecture 2



Descriptive Analytics I: Nature of Data, Statistical Modeling

Lecture 2

IT393

Business Intelligence

Prepared by
Dr. Wesam Ahmed

01

Learning Objectives:

- 1.1 Understand the nature of data as it relates to business intelligence (BI) and analytics
- 1.2 Learn the methods used to make real world data analytics ready
- 1.3 Describe statistical modeling and its relationship to business analytics
- 1.4 Learn about descriptive and inferential statistics

02

A Simple Taxonomy of Data:

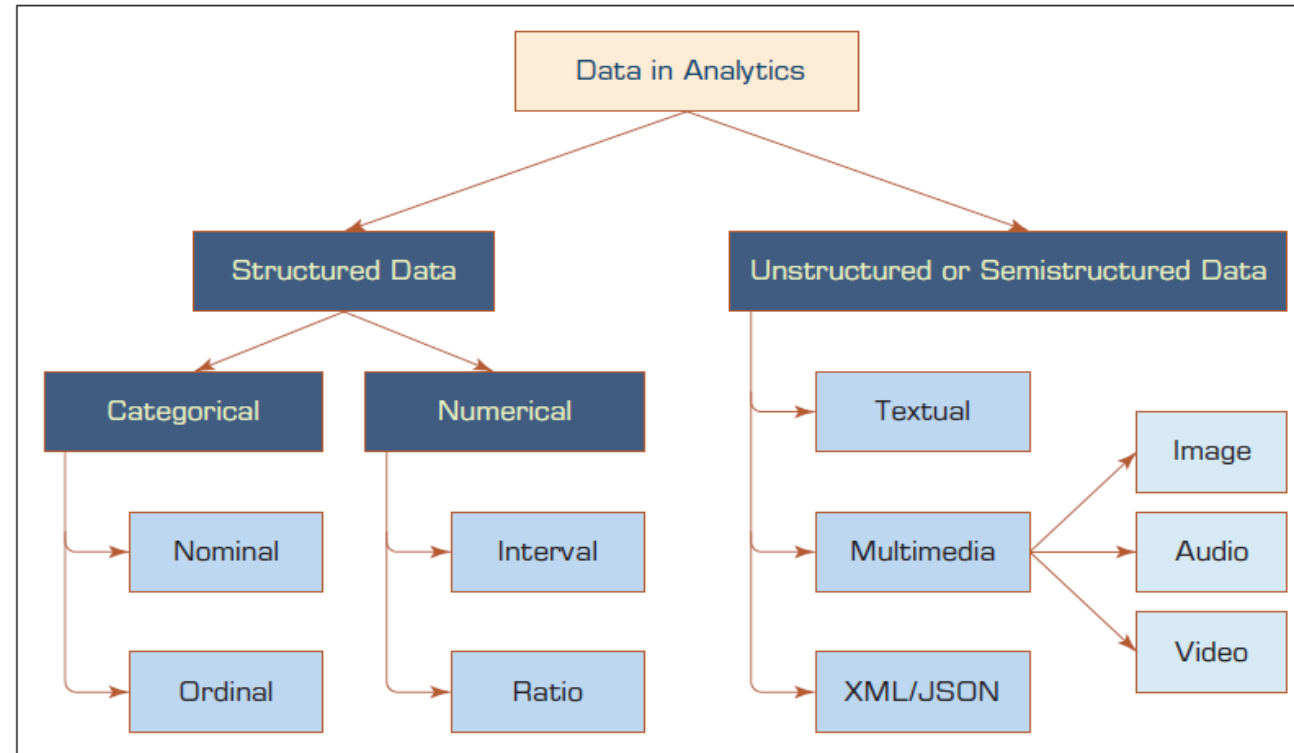


Figure 2.1 A Simple Taxonomy of Data

03

A Simple Taxonomy of Data:

- **Numeric data** represent the **numeric values of specific variables**. Examples of numerically valued variables include **age, number of children, total household income (in U.S. dollars), travel distance (in miles), and temperature (in Fahrenheit degrees)**.
- Numeric values representing a variable can be **integer (taking only whole numbers)** or **real (taking also the fractional number)**.
- The numeric data may also be called **continuous data** represents scalable measurements, and it is possible for **the data to contain an infinite number of fractional values**. Unlike **a discrete variable**, which represents finite, countable data.

04

A Simple Taxonomy of Data:

Other data types, including textual, spatial, imagery, video, and voice, need to be converted into some form of categorical or numeric representation before they can be processed by analytics methods .

analytics (i.e., data mining) methods and machine-learning algorithms are very selective about the type of data that they can handle. Providing them with incompatible data types may lead to incorrect models or (more often) halt the model development process.

For example, some data mining methods need all the variables (both input as well as output) represented as numerically valued variables (e.g., neural networks, support vector machines, logistic regression).

The nominal or ordinal variables are converted into numeric representations using some type of 1-of-N pseudo variables (e.g., a categorical variable with three unique values can be transformed into three pseudo variables with binary values 1 or 0).

05

A Simple Taxonomy of Data:

Name	Age	City	Education Level	Satisfaction Rating
Alice	25	New York	Bachelor's	4
Bob	30	Los Angeles	Master's	5
Charlie	22	Chicago	High School	3
Diana	28	Houston	PhD	4

Figure 2.2 Example of dataset

06

A Simple Taxonomy of Data:

Explanation of Data Types:

Name: Categorical (nominal)

Age: Numerical (continuous)

City: Categorical (nominal)

Education Level: Categorical (ordinal)

Satisfaction Rating: Numerical
(discrete)

This dataset is small (4 rows, 5 columns) .

Name	Age	City	Education Level	Satisfaction Rating
Alice	25	New York	Bachelor's	4
Bob	30	Los Angeles	Master's	5
Charlie	22	Chicago	High School	3
Diana	28	Houston	PhD	4

07

The Art and Science of Data Preprocessing:

Data in its original form (i.e., the real-world data) is not usually ready to be used in analytics tasks. It is often dirty, misaligned, overly complex, and inaccurate.

A tedious and time-demanding process (so-called data preprocessing) is necessary to convert the raw real-world data into a well-refined form for analytics algorithms .

Many analytics professionals would testify that the time spent on data preprocessing is significantly longer than the time spent on the rest of the analytics tasks (the fun of analytics model building and assessment).

08

The Art and Science of Data Preprocessing:

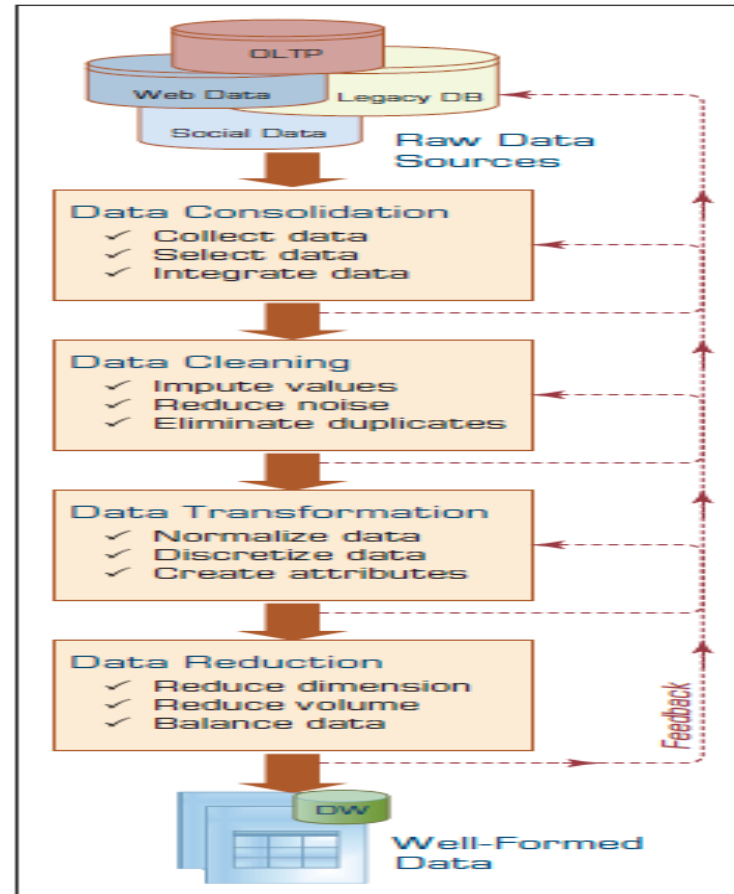


Figure 2.3 shows the main steps in the data preprocessing

09

The Art and Science of Data Preprocessing:

In the first phase of data preprocessing, the relevant data is collected from the identified sources, the necessary records and variables are selected (based on an intimate understanding of the data, the unnecessary information is filtered out), and the records coming from multiple data sources are integrated/merged .

In the second phase of data preprocessing, the data is cleaned (this step is also known as data cleaning). Data in its original/raw/real-world form is usually dirty . In this step, the values in the data set are identified and dealt with. In some cases, missing values are an anomaly in the data set, in which case they need to be imputed (filled with a most probable value) or ignored; in other cases, the missing values are a natural part of the data set (e.g., the household income field is often left unanswered by people who are in the top income tier). In this step, the analyst should also identify noisy values in the data (i.e., the outliers) and smooth them out.

10

The Art and Science of Data Preprocessing:

In addition, **inconsistencies** (unusual values within a variable) in the data should be handled **using domain knowledge and/or expert opinion**.

Inconsistencies in data refer to **unusual or unexpected values within a variable** that do not align with what is typically expected in a given context.

These could be:

- **Outliers:** Values that are significantly different from most others (e.g., a **human height of 10 meters** in a dataset).
- **Errors:** Mistakes in data entry or collection (e.g., **typing 100 instead of 10**).

Domain knowledge is the understanding of the specific field or area the data relates to. It provides a foundation for recognizing what values are realistic or expected.

For example:

In a dataset of human heights, domain knowledge from biology and anthropology tells us that heights typically range from **about 1.5 to 2 meters**. **A value of 10 meters would be flagged as inconsistent** because it's far beyond what's biologically possible.

11

The Art and Science of Data Preprocessing:

In the third phase of data preprocessing, **the data is transformed for better processing**. For instance, in many cases the data is normalized between a certain minimum and maximum for all variables to mitigate the potential bias of one variable (having large numeric values, such as for household income) dominating other variables (such as years in service, which may potentially be more important) having smaller values.

Imagine a dataset with one variable:

Household Income: Ranges from \$10,000 to \$1,000,000.

After Normalization: For income: A \$10,000 income becomes 0, and \$1,000,000 becomes 1.

Another transformation that takes place is **discretization and/or aggregation**.

In some cases, the numeric variables are converted to categorical values Example: Consider a dataset with exact ages: 23, 45, 67, etc. Instead of using these continuous values, we discretize them into categories:

Low: 0-30 years (e.g., 23 becomes "Low").

Medium: 31-60 years (e.g., 45 becomes "Medium").

High: 61+ years (e.g., 67 becomes "High").

12

The Art and Science of Data Preprocessing:

in other cases, a nominal variable's unique value range is reduced to a smaller set using concept hierarchies to have a data set that is more amenable to computer processing.

Example:

Suppose you have a "location" variable listing all 50 U.S. states (e.g., New York, Texas, California). Aggregating these into regions simplifies the data: **This reduces 50 unique values to 4.**

Northeast: New York, Massachusetts, etc.

South: Texas, Florida, etc.

Midwest: Ohio, Illinois, etc.

West: California, Washington, etc.

Still, in other cases one might choose **to create new variables based on the existing ones** to magnify the information found in a collection of variables in the data set.

13

The Art and Science of Data Preprocessing:

The final phase of data preprocessing is **data reduction**. Even though data scientists like to **have large data sets**, **too much data may also be a problem**. In the simplest sense, one can visualize the data commonly used in predictive analytics projects as a flat file consisting of two dimensions: variables (the number of columns) and cases/records (the number of rows). In some cases (**e.g., image processing and genome projects with complex microarray data**), the number of variables can be rather large, and **the analyst must reduce the number down to a manageable size**. Because the variables are **treated as different dimensions** that describe the phenomenon from different perspectives, in predictive analytics and data mining this process **is commonly called dimensional reduction (or variable selection)**. Even though there is not a single best way to accomplish this task, run appropriate statistical tests (e.g., principal component analysis or independent component analysis); use a combination of these techniques to reduce the dimensions in the data into most relevant subset.

14

The Art and Science of Data Preprocessing:

With respect to the other dimension, some data sets may include millions or billions of records. Even though computing power is increasing exponentially. In such cases, one may need to sample a subset of the data for analysis. The underlying assumption of sampling is that the subset of the data will contain all relevant patterns of the complete data set.

15

The Art and Science of Data Preprocessing:

For skewed data, the stratified sampling (a proportional representation of different subgroups in the data is represented in the sample data set) may be required.

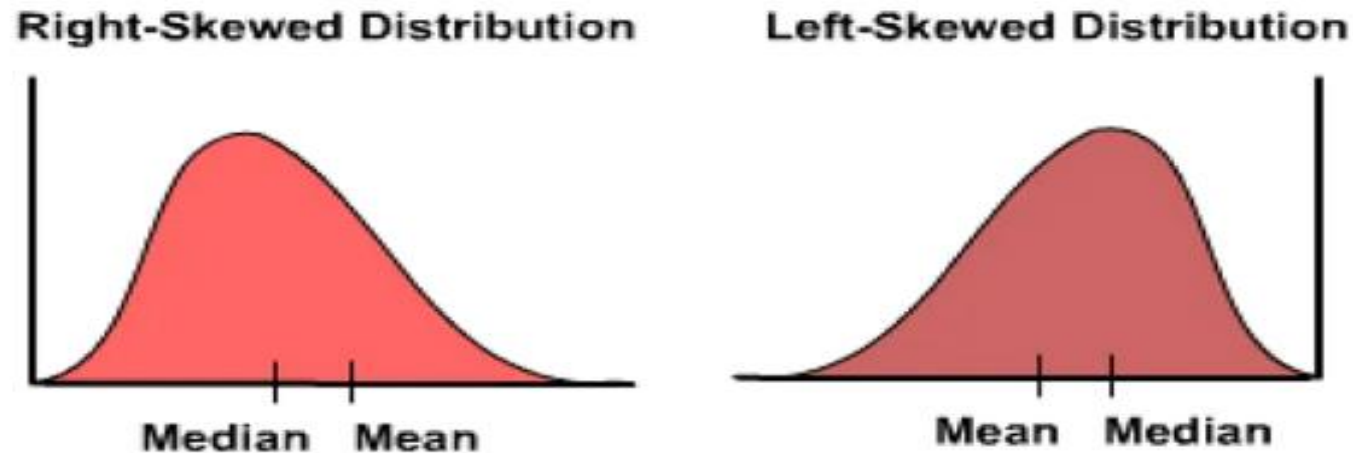


Figure 2.4 There are 2 different types of skews in data, left(negative) or right(positive) skew.

16

The Art and Science of Data Preprocessing:

For skewed data, the stratified sampling (a proportional representation of different subgroups in the data is represented in the sample data set) may be required.

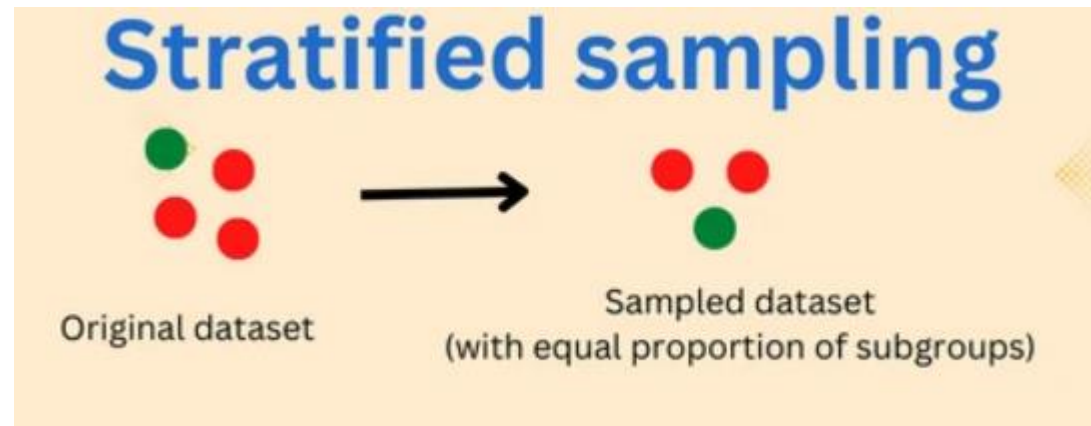


Figure 2.5 The stratified sampling.

17

The Art and Science of Data Preprocessing:

It is a good practice to **balance the highly skewed data** by either **oversampling** the less represented or **undersampling** the more represented classes. Research has shown that balanced data sets tend to produce better prediction models than unbalanced ones.

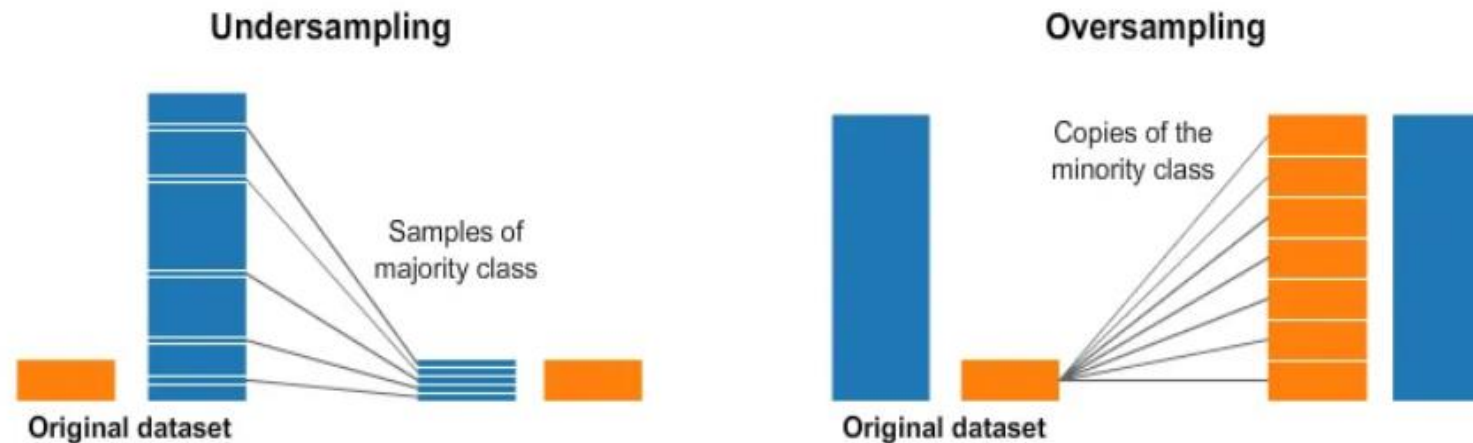


Figure 2.6 The oversampling and undersampling.

18

The Art and Science of Data Preprocessing:

Main Task	Subtasks
Data consolidation	Access and collect the data Select and filter the data Integrate and unify the data
Data cleaning	Handle missing values in the data Identify and reduce noise in the data Find and eliminate erroneous data

Figure 2.7 A Summary of Data Preprocessing Tasks and Potential Methods

19

The Art and Science of Data Preprocessing:

Data transformation

Normalize the data

Discretize or aggregate the data

Construct new attributes

Data reduction

Reduce number of attributes

Reduce number of records

Balance skewed data

20

Application Case 2.1:

Student attrition has become one of the most challenging problems for decision makers in academic institutions. **Student attrition refers to the phenomenon of students leaving a course or degree program before completion.** Despite all the programs and services that are put in place to help retain students, according to the U.S. Department of Education, Center for Educational Statistics (nces.ed.gov), **only about half of those who enter higher education actually earn a bachelor's degree.** Enrollment management and the retention of students has become a top priority for administrators of colleges and universities in the United States and other countries around the world. **High dropout of students usually results in overall financial loss, lower graduation rates, and inferior school reputation in the eyes of all stakeholders.** The legislators and policy makers who oversee higher education and allocate funds, the parents who pay for their children's education to prepare them for a better future, and the students who make college choices look for evidence of institutional quality and reputation to guide their decision-making processes.

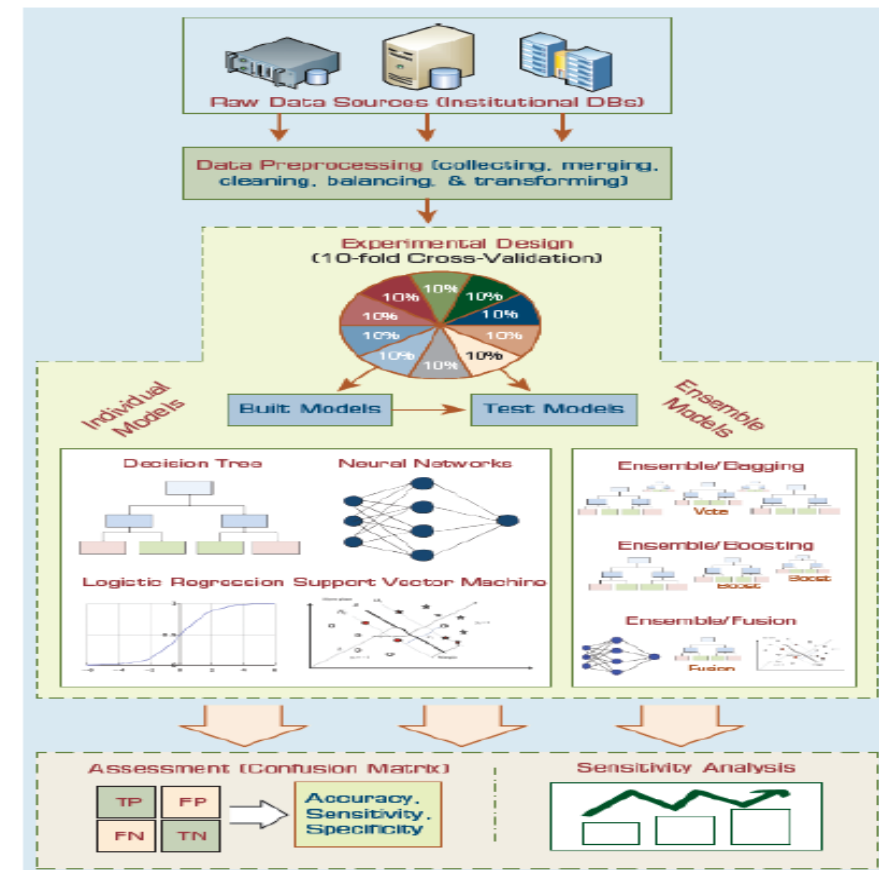
21

Application Case 2.1:

Proposed Solution: To improve student retention, one should try to understand the nontrivial reasons behind the attrition. To be successful, one should also be able to accurately identify those students that are at risk of dropping out. The project summarized in this case study proposed a quantitative research approach where the historical institutional data from student databases could be used to develop models that are capable of predicting as well as explaining the institution-specific nature of the attrition problem. **The solution predicts which students are at risk of dropping out before completing their degree.** Using historical data, the models identify individuals who show signs of potential attrition based on patterns like poor academic performance, irregular attendance, or other institution-specific indicators.

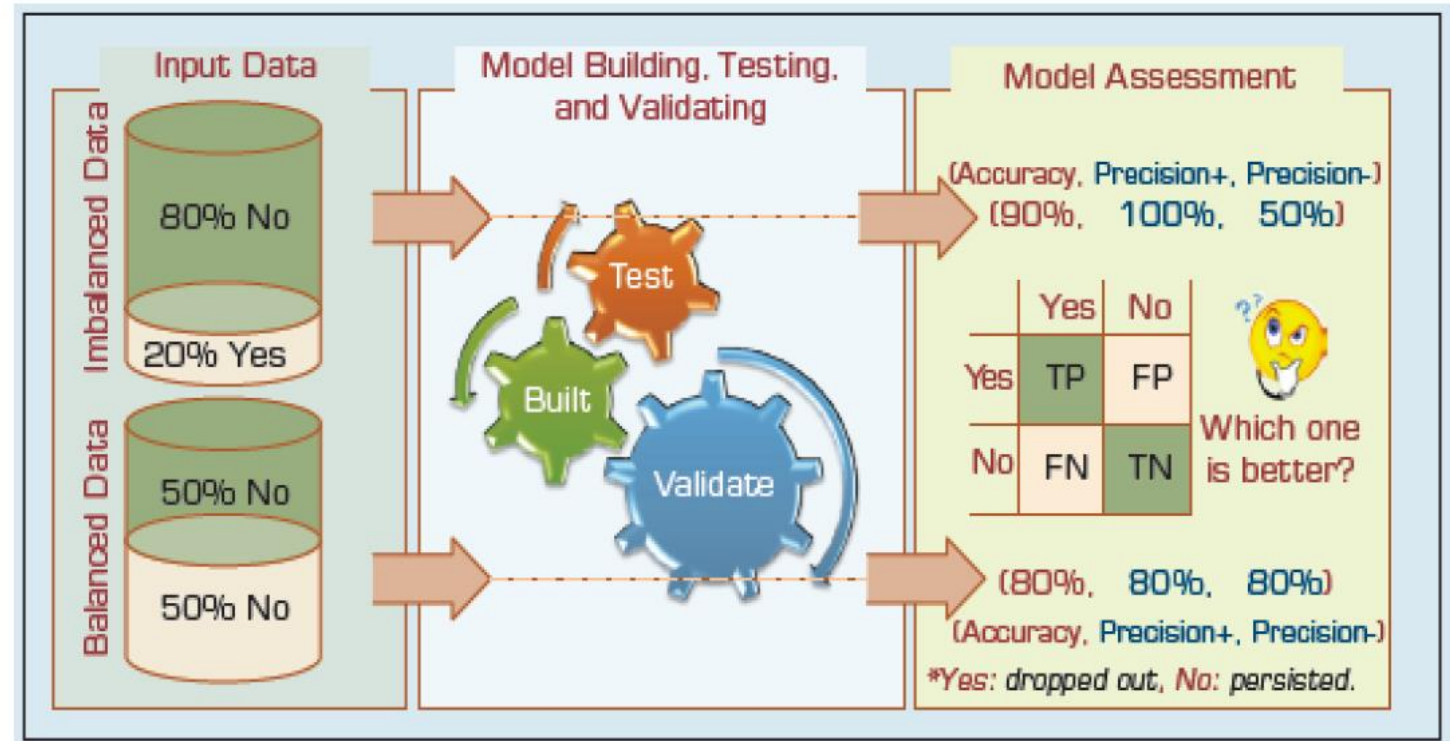
Application Case 2.1: Improving Student Retention with Data-Driven Analytics

The proposed analytics approach is shown in Figure 2.8.



Application Case 2.1: Improving Student Retention with Data-Driven Analytics

Research shows that having such an **imbalanced data** has a **negative impact on model performance**. Therefore, the study experimented with the options of using and comparing the results of the same type of models built **with the original imbalanced data** and the **well balanced data**. Figure 2.9 A Graphical Depiction of the Class Imbalance Problem.



Application Case 2.1: Improving Student Retention with Data-Driven Analytics

	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	1494	384	1518	304	1478	255	1438	376
Yes	1596	11142	1572	11222	1612	11271	1652	11150
SUM	3090	11526	3090	11526	3090	11526	3090	11526
Per-Class Accuracy	48.35%	96.67%	49.13%	97.36%	47.83%	97.79%	46.54%	96.74%
Overall Accuracy	86.45%		87.16%		87.23%		86.12%	

Figure 2.10 Prediction Results for the Original/Unbalanced Dataset

25

Application Case 2.1: Improving Student Retention with Data-Driven Analytics

Confusion Matrix	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	2309	464	2311	417	2313	386	2125	626
Yes	781	2626	779	2673	777	2704	965	2464
SUM	3090	3090	3090	3090	3090	3090	3090	3090
Per-class Accuracy	74.72%	84.98%	74.79%	86.50%	74.85%	87.51%	68.77%	79.74%
Overall Accuracy	79.85%		79.85%		81.18%		74.26%	

Figure 2.11 Prediction Results for the Balanced Data Set

Application Case 2.1: Improving Student Retention with Data-Driven Analytics

	Boosting		Bagging		Information Fusion	
	(Boosted Trees)		(Random Forest)		(Weighted Average)	
	No	Yes	No	Yes	No	Yes
No	2242	375	2327	362	2335	351
Yes	848	2715	763	2728	755	2739
SUM	3090	3090	3090	3090	3090	3090
Per-Class Accuracy	72.56%	87.86%	75.31%	88.28%	75.57%	88.64%
Overall Accuracy	80.21%		81.80%		82.10%	

Figure 2.12 Prediction Results for the Three Ensemble Models

27

Statistical Modeling for Business Analytics:

Because of the increasing popularity of business analytics, the traditional statistical methods and underlying techniques are also regaining their attractiveness as enabling tools to support evidence-based managerial decision making. Statistics (statistical methods and underlying techniques) is usually considered as part of descriptive analytics .

Descriptive analytics has two main branches: statistics and online analytics processing (OLAP). OLAP is the term used for analyzing, characterizing, and summarizing structured data stored in organizational databases (often stored in a data warehouse) The OLAP branch of descriptive analytics has also been called Business Intelligence.

Statistical Modeling for Business Analytics:

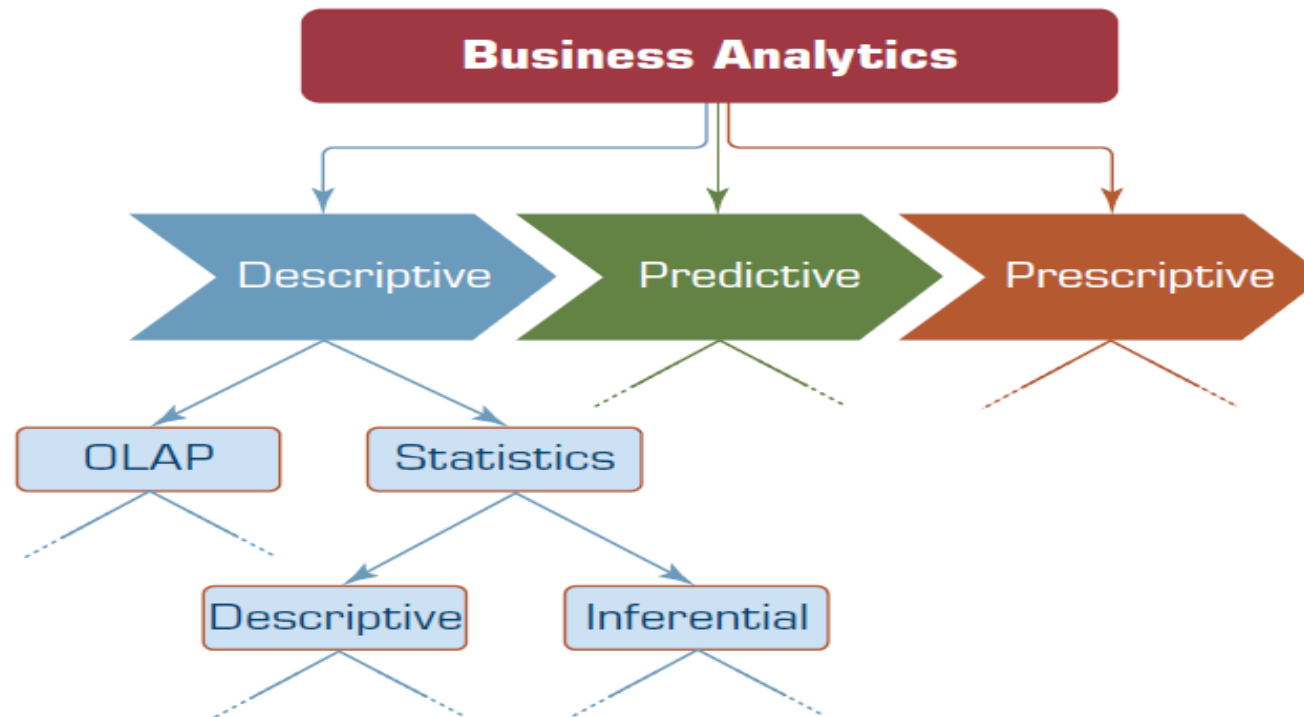


Figure 2.13 Relationship between statistics and descriptive analytics.

29

Statistical Modeling for Business Analytics:

- **Statistics:** a collection of mathematical techniques to characterize and interpret data. at the highest level, statistical methods can be classified as either **descriptive or inferential**.
- **The descriptive statistics** is all about describing the sample data on hand.
- **Inferential statistics** is about drawing inferences or conclusions about the characteristics of the population.

30

Descriptive Statistics for Descriptive Analytics:

In business analytics, **descriptive statistics** plays a critical role:

- It allows us to understand and **explain/present our data in a meaningful manner** using aggregated numbers **data tables, or charts/graphs**. Such an understanding not only **helps business users in their decision-making processes**, but also helps **analytics professionals and data scientists to characterize and validate the data for other more sophisticated analytics tasks**.
- Descriptive statistics allows analysts to **identify data concertation**, unusually large or small values (i.e., **outliers**), and unexpectedly distributed data values for numeric variables.
- Therefore, the methods in descriptive statistics can be classified as either **measures for central tendency or measures of dispersion**.

31

Measures of centrality tendency (May Also Be called measures of location or centrality)

A measure of central tendency is a single numerical value that aims to describe a set of data by simply identifying or estimating the central position within the data.

The mean (often called the arithmetic mean or the simple average) is the most commonly used measure of central tendency.

In addition to mean, you could also see median or mode being used to describe the centrality of a given variable.

Although, the mean, median, and mode are all valid measures of central tendency, under different circumstances, one of these measures of centrality becomes more appropriate than the others.

32

Arithmetic mean

The arithmetic mean (or simply mean or average) is the sum of all the values/observations divided by the number of observations in the data set.

It is by far the most popular and most commonly used measure of central tendency.

It is used with continuous or discrete numeric data.

33

Arithmetic mean

For a given variable x , if we happen to have n values/observations $(x_1, x_2, \dots, x_n)$, we can write the **arithmetic mean of the data sample** ($\bar{x}$, pronounced as **x-bar**) as follows in Equation 1:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

34

Arithmetic mean

For a given variable x , if we happen to have n values/observations ($x_1, x_2, \dots, x_n$), we can write the **arithmetic mean of the data sample ($\bar{x}$, pronounced as x-bar)** as follows in Equation 1:

One major downside is that the mean can be **affected by outliers** (observations that are considerably larger or smaller than the rest of the data points).

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

35

Median:

The median is **the measure of center value in a given data set.**

It is the number in the middle of a given **set of data that has been arranged/sorted in order of magnitude** (either ascending or descending).

If the number of observation is **an odd number**, identifying the median is very easy—just sort the observations based on their values and pick the value right in the middle.

If the number of observations is an **even number**, then **identify the two middle values, and then take the simple average of these two values.**

In contrary to the mean, the median is not affected by outliers or skewed data.

Mode:

The mode is the observation that occurs most frequently (the most frequent value in our data set).

On a histogram it represents the highest bar in a bar chart, and hence, it may be considered as being the most popular option/value.

The mode is most useful for data sets that contain a relatively small number of unique values.

That is, it may be useless if the data have too many unique values .

Although it is a useful measure (especially for nominal data), mode is not a very good representation of centrality, and therefore, it should not be used as the only measure of central tendency for a given data set.

Mode:

In summary, which central tendency measure is the best?

Use the mean when the data is not prone to outliers and there is no significant level of skewness.

Use the median when the data has outliers and/or it is ordinal in nature.

Use the mode when the data is nominal.

Perhaps the best practice is to use all three together so that the central tendency of the data set can be captured and represented from three perspectives.

38

Example of Measures of centrality tendency:

Sample Dataset

Let's consider a dataset representing the number of visitors to a website each day over a week:

- Day 1: 120
- Day 2: 150
- Day 3: 130
- Day 4: 150
- Day 5: 140
- Day 6: 170
- Day 7: 155

39

Example of Measures of centrality tendency:

1. Mean (Average)

The mean is calculated by summing all the values in the dataset and then dividing by the number of values.

Sum of the values:

$$120 + 150 + 130 + 150 + 140 + 170 + 155 = 1015$$

Number of values: 7

Mean:

$$1015 \div 7 = 145$$

So, the mean number of visitors is **145**

2. Median

Ordered dataset:

[120, 130, 140, 150, 150, 155, 170]

Median:

The 4th value, which is 150

So, the median number of visitors is 150.

40

Example of Measures of centrality tendency:

3. Mode

In this dataset, 150 appears twice, while all other values (120, 130, 140, 155, 170) appear only once.

So, the mode is 150.

Summary

1. Data refers to a collection of facts usually obtained as the result of experiments, observations.
2. At the highest level of abstraction, data can be classified as structured and unstructured.
3. Data in its original/raw form is not usually ready to be useful in analytics tasks.
4. Data preprocessing is a tedious, time-demanding, yet crucial task in business analytics.
5. Statistics is a collection of mathematical techniques to characterize and interpret data.
6. Statistical methods can be classified as either descriptive or inferential.
7. Statistics in general, and descriptive statistics in particular, is a critical part of BI and business analytics.
8. Descriptive statistics methods can be used to measure central tendency.



Thank You