Statistical Learning

Introduction to Statistics

PGPBDML

Outline

1. Why Statistics?
2. Business Statistics-Tools
3. Types of Statistics - Descriptive and Inferential Statistics
4. Data Sources and Types of Datasets
5. Attributes of Datasets
6. Key Takeaways

Why Statistics is So Important?

Three significant events triggered the current meteoric growth in the use of analytics in decision making and *Statistics is the Heart of Analytics*

Event1

- Technological developments, Revolution of Internet and social networks, data generated from mobile phones and other electronic devices, produce large amount of data from which insights will have to be sifted.

- The discovery of pattern and trends from these data for organizations will pave the way for improving profitability, understanding customer expectations, and appropriately pricing their products so that they can gain competitive advantage in the marketplace.

Why Statistics is So Important?

Event 2

- Advances in enormous computing power to effectively process and analyze massive amounts of data
- Sophisticated and faster algorithms for solving problems
- Data Visualization for Business Intelligence

Why Statistics is So Important?

Event 3

- Large data storage capability
- Parallel computing, and cloud computing coupled with better computer hardware have enabled businesses to solve large scale problems faster than ever before without sacrificing

Big Data

Big data

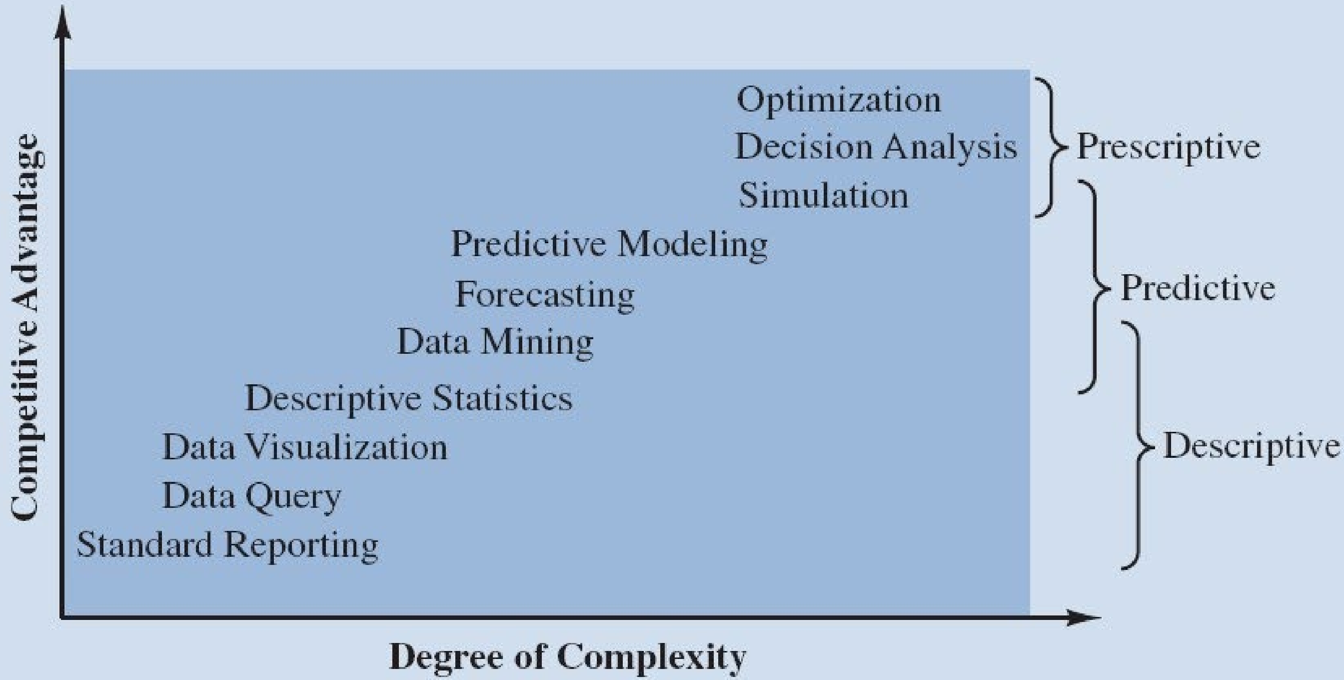
- A set of data that cannot be managed, processed, or analyzed with traditional software/algorithms within a reasonable amount of time.
- Big data revolves around

Volume Velocity Variety Value Veracity

Walmart handles over one million purchase transactions per hour.

Facebook processes more than 250 million picture uploads per day.

Statistics in The Spectrum of Business Analytics



Source: Adapted from SAS.

Business Statistics-Tools

Classification

- *Classification* techniques helps in segmenting the customers into appropriate groups based on key characteristics.

- For example, using *appropriate statistical model*, an organization could easily segment the customers into Long Term Customers, Medium Term Customers, and Brand Switchers.

- Another application in this context is classifying customers into "Buyers and Non-Buyers."

- Classification will help professionals understand the customer behavior and position their products and brands using appropriate strategies.

Business Statistics-Tools

Pattern Recognition

- "A picture is worth thousand words" and it reveals hidden pattern in the data that could be leveraged by retail professionals. Pattern recognition techniques include *Histogram*, *Box Plot*, *Scatter Plot* and other *Visual Analytics*.
- For example, histogram drawn for income of a particular class of customers may reveal a symmetrical bell curve pattern or may be left or right skewed.
- Relationship between age and expenditure could be captured using a scatter plot.
- *Box Plot* enables Analytics Professionals to sift outliers (extreme points) apart from providing the distribution pattern.

Business Statistics-Tools

greatlearning

Association

- *Association* Analysis helps in determining which of the items go together. Association rules include a set of analytics that focuses on discovering relationships that exist among specific objects.

- In this context, market basket analysis refers to an association rule that generates the probability for an outcome.

- For example, market basket analysis may lead to a finding that if customers buy coffee, there is a 40% probability that they also buy bread.

- Association rules can be adapted by organizations to store lay out, items bundling, discount and sales promotion decisions, and cross selling among others.

Business Statistics-Tools

Predictive Modeling

- Both customer segmentation as well as identifying and targeting most profitable customers can be facilitated by predictive models.

- *Regression* can be used for predicting the amount of expenditure on a particular product based on input variables income, age, and gender.

- Organizations can leverage on other advanced models that comprise *Logistic Regression*, and *Neural Networks* for predicting a target variable as well as classifying and predicting into which group the consumer belongs to.

- For example, these models can classify and predict buyers and non-buyers, and defaulters and non-defaulters on credit card loan.

Classic Definition of Statistics

" By Statistics, we mean methods specially adopted to the elucidation of quantitative data affected to a marked extent by multiplicity of causes".

Yule and Kendal

It is interesting to see what *Thomas Davenport* means by Business Analytics and note the similarities and dissimilarities between the two.

"Business Analytics (BA) can be defined as the broad use of data and quantitative analysis for decision making within organizations".

Types of Statistics

Descriptive Statistics is concerned with Data Summarization, Graphs/Charts, and Tables

Inferential Statistics is a method used to talk about a Population Parameter from a Sample.

Some Key Terms Used in Statistics

Population is the collection of all possible observations of a specified characteristic of interest. An example is all the students in the Quantitative Methods course in an MBA program.

Sample is a subset of the population. Suppose you want to select a team of 20 students from 200 students in an MBA program for participating in a management quiz. The total 200 students is the population. The 20 students selected for the quiz is the sample.

Parameter is the population characteristic of interest. For example, you are interested in the average income of a particular class of people. The average income of this entire class of people is called a parameter.

Statistic is based on a sample to make inferences about the population parameter. If you look at the previous example, the average income in the population can be estimated by the average income based on the sample. This sample average is called a statistic.

Data Sources

Primary Data are collected by the organization itself for a particular purpose. The benefits of primary data are that they fit the needs exactly, are up to date, and reliable.

Secondary Data are collected by other organizations or for other purposes. Any data, which are not collected by the organization for the specified purpose, are secondary data. These may be published by other organizations, available from research studies, published by the government, web, social media and so on.

Types of Data

Qualitative Data are nonnumeric in nature and can't be measured. Examples are gender, religion, and place of birth.

Quantitative Data are numerical in nature and can be measured. Examples are balance in your savings bank account, and number of members in your family.

Quantitative data can be classified into discrete type or continuous type. **Discrete type** can take only certain values, and there are discontinuities between values, such as the number of rooms in a hotel, which cannot be in fraction. **Continuous type** can take any value within a specific interval, such as the production quantity of a particular type of paper (measured in kilograms).

Types of Data Sets

- **Record**
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- **Graph and network**
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- **Ordered**
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- **Spatial, image and multimedia:**
  - Spatial data: maps
  - Image data
  - Video data

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Data Objects

- Data sets are made up of data objects.
- A data object represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples* , *examples, instances, data points, objects, tuples*.
- Data objects are described by attributes.
- Database rows -> data objects; columns ->attributes.

Attributes

- Attribute (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Binary
  - Ordinal
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types

- Nominal: categories, states, or “names of things”
  - *Hair_color = {auburn, black, blond, brown, grey, red, white}*
  - marital status, occupation, ID numbers, zip codes
- Binary
  - Nominal attribute with only 2 states (0 and 1)
  - <u>Symmetric binary</u>: both outcomes equally important
    - e.g., gender
  - <u>Asymmetric binary</u>: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- Ordinal
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size = {small, medium, large}*, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- Interval
    - Measured on a scale of equal-sized units
    - Values have order
        - E.g., *temperature in C˚or F˚, calendar dates*
    - No true zero-point
- Ratio
    - Inherent zero-point
    - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
        - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Take-aways

- Statistics has become a very important field now.
- We briefly discussed about various business stats tools like classification, pattern recognition, association, and predictive modeling.
- Statistics now has much more focus on predictive and prescriptive aspects.
- Types of datasets, Data types.

References: Data Types and Statistical Measures

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009