Business Analytics

Introduction & R

Dr. Jiun-Yu Yu
BA, NTU
11 Sep 2018

Outlines

- Introduction
  - Analytics and Business
    - Three levels of analytics
  - What is Statistics
    - Some rules for implementing statistical data analysis
    - Data types
    - Data analyzing tools
- Introducing R

Analytics? Big Data?

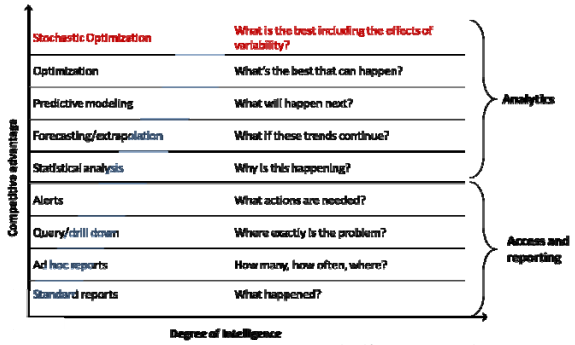
- Analytics is the scientific process of transforming data into insight for making better decisions.
  - Data
  - Transforming data
  - Insight
  - Decisions → Making better decisions
  - Scientific process
- Information-based strategy / Data-driven decision-making
  - Make better business decisions with analytics – data collection and analysis

Analytics and Business

- To provide management with a better appreciation of the value of data analytics.
- To communicate concepts of statistics and data mining in plain language.
- To illustrate how data can add value to the everyday life of business executives.

- Includes:
  - Statistics
  - Management Science / Operational Research (MS/OR)
    - Optimization, Simulation, …

Analytics and Business Intelligence



Adopted from "Competing on Analytics", Figure 1-2, p. 8

Three Levels of Analytics

Descriptive Analytics

What is happening

Past data only, no model

Predictive Analytics

What will happen

Model fitting past data, to see what's going to be in the future

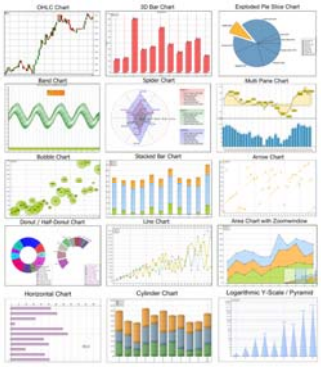
Prescriptive Analytics

What is the best strategy

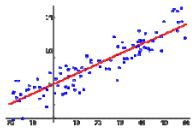
Tell you what to do next: optimization, randomized control, design of experiments, simulation

Descriptive Analytics

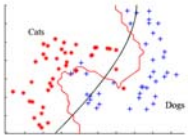
- KPIs
- Dashboards
- Visualizations
- …



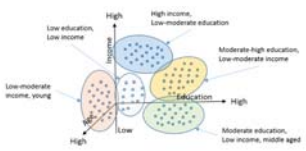
Predictive Analytics



Regression



Classification



Clustering

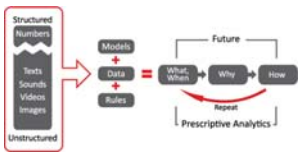
Anomaly detection

Predictive Analytics

- Enabling businesses to use predictive models to exploit patterns found in historical data to identify potential risks and opportunities before they occur.
- Predictive analytics can use ALL available data
  - Descriptive data
    - attributes, characteristics, self-declared info, (geo)demographics…
  - Behavioral data
    - orders, transactions, payment history, usage history…
  - Attitudinal data
    - opinions, preferences, needs & desired, survey results, social media…
  - Interaction data
    - email/chat transcripts, call center notes, web click-streams, in-person dialogues…

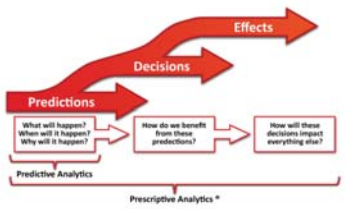
Prescriptive Analytics

- Prescriptive analytics synergistically combines hybrid data and business rules with mathematical and computational models to make predictions and then suggests decision options to take advantage of the predictions.



Prescriptive Analytics

- Prescriptive analytics can continually take in new data to re-predict and re-prescribe, thus automatically improving prediction accuracy and prescribing better decision options.



Prescriptive Analytics

Major tools:

- Design of Experiment / Experimental Design
- Optimization
- Simulation
- Decision Analysis

What is Statistics

- uncertainty & variability → outcomes → observations → data
- Statistics is the science of *collecting*, *displaying*, *analyzing*, and *interpreting* data
- Do *decision-making* based on the results
- Statistics as problem solving
  - Formulate a real problem in statistical terms
  - Collect data efficiently
  - Analyze data to extract the maximum amount of information
  - Interpret (*statistical inference*)
  - Repot results

Chatfield (1995), *Problem Solving – A Statistician's Guide*, Chapmen & Hall

Some Rules about Data Analysis

- Do not attempt to analyze the data until you understand what is being measured and why. Find out whether there is any prior information about likely effects.
- Find out how the data were collected.
- Look at the structure of the data.
- The data then needs to be carefully examined in an exploratory way, before attempting a more sophisticated analysis.
- Use your common sense at all times.
- Report the results in a clear, self-explanatory way.

Chatfield (1995), *Problem Solving – A Statistician's Guide*, Chapmen & Hall

Data Types

- Dataset is mainly stored and displayed in tabular form in which columns represent variables and rows are observations (*spreadsheet / Excel format*)

	Var 1	*Var 2*	*Var 3*	….	*Var m*
1					
2					
3					
…					
n					

- We are interested in finding some interesting patterns or structures that are hidden in the data
- Patterns and structures may exist among variables or among observations
- Usually we must ensure $n > m$

Data Types

- Discrete / Qualitative / Categorical
  - Nominal
    - No particular ordering to the possible values: *post code, ID number*
  - Ordinal
    - Natural ordering but no implication of distance between scale positions: *exam rank, education level*
- Continuous / Quantitative / Numerical
  - Interval
    - Equal differences between successive integers but where the zero point is arbitrary: *temperature (°C), longitude, time interval*
  - Ratio
    - Can compare relative magnitude of scores and differences in scores: *income, weight, length*

Analyzing Tools

- If we believe that one of the variables can be "*explained*" by other variables, then the *regression-type / function-based* models can be applied.
- The variable selected to be explained:
  - *y*, response variable, dependent variable, or random component.
- Other variables:
  - *x*, explanatory variables, independent variables, predictor, covariate, or systematic component
- If no variable is suitable to serve the role of response variable, then the traditional *Multivariate Analysis* (*MVA*) methods would be better choices.
  - MVA is not discussed in this course.

Analyzing Tools

Regression type

- $y = f(x_1, x_2, \ldots, x_p)$
  - C C → Linear Model (LM)
  - C C+D → Analysis of Co-variance (ANCOVA)
  - C D → Analysis of Variance (ANOVA), *Experimental Design*
  - D (D,C) → Generalized Linear Model (GLM)

Multivariate Analysis

- Structures among variables
  - Principal Component Analysis (PCA), Factor Analysis (FA)
- Groups among observations
  - Clustering, Multi-Dimensional Scaling (MDS)

R

- Free, flexible, fast developing, frequently updated
- Both a statistics package and a programming language
- www.r-project.org
  - Download → CRAN → Taiwan → NTU
  - Download R for Windows (also available for Mac OS X & Linux)
  - Base
  - Current version: R 3.4.1 for Windows
  - Installation

R

- Documentation
  - Manuals
    - *"An Introduction to R"*
    - *"R Data Import/Export"*
    - …
  - Contributed
    - *"R for Beginners"*
    - Reference cards
    - …
    - Some are available in *simplified* Chinese version
- NTU Statistics Education Center
  - http://www.statedu.ntu.edu.tw/chinese/download.asp
    - The first two tutorials are sufficient for this course

Some codes on R (1)

- Value assign:
  - `n <- 15`
- Case sensitive:
  - `x <- 1`
  - `X <- 10`
- Value replace:
  - `n <- 10 + rnorm(1)`
- Need some help?
  - `help(rnorm)`
  - `?rnorm`
  - `help.start()`

Some codes on R (2)

- Change directory
  - Change manually on GUI, or
  - `setwd("D:/R_work")`
- Load package:
  - `library(MASS)`
- Install package:
  - Install manually on GUI, or
  - `install.packages("fBasics")`
    - You will be asked to highlight the mirror nearest to you for downloading (e.g. Taipei), then everything else is automatic.
- Inspect packages currently loaded:
  - `search()`

Some codes on R (3)

- List the objects in memory:
  - `ls()`
    - `name <- "Carmen"`
    - `n1 <- 10;  n2 <- 100;  m <- 0.5`
    - `ls();  ls(pattern="m");  ls(pat="^m")`
    - `ls.str()`
- Delete objects:
  - `rm()`
    - `rm(n1)`
    - `rm(list=ls())`

Some codes on R (4)

- Vector:
  - `x1 <- c(0:7)` `# c(1:3, 7:9)`
  - `x2 <- rep(2, 8)`
  - `x2 <- rep(1:4, 2)`
  - `x2 <- rep(1:4, each = 2)`
  - `x3 <- seq(-1, 1, 0.2)` `# evenly spaced sequence`
- Matrix:
  - `x4 <- rnorm(12);  length(x4)`
  - `dim(x4) <- c(4,3)`
  - `x4`

Some codes on R (5)

- Aggregate – `cbind / rbind`:
  - `u <- c(1:3);  v <- c(-1:-3)`
  - `( m <- cbind(u, v) )`
- Matrix indexing:
  - `m[2,2];  x4[1,3]`
- Measures of Association for Continuous Variables
  - `cov(x1, x2)`
  - `cor(x1, x2)`
  - `var(x4)`
  - `cor(x4)`
  - `sd(x4)^2`
  - `cov(x4[,1], x4[,2])`

Lecture Summary

- Introduction
  - Analytics and Business
    - Three levels of analytics
  - What is Statistics
    - Some rules for implementing statistical data analysis
    - Data types
    - Data analyzing tools
- Introducing R

Reading & Assignment

- Paradis (2005), *R for Beginners*
- Assignment 0
  - Install R onto your personal computer and practice script '`s01_Intro.R`', downloadable from course website.