Data Mining Using Business Analytics / Machine Learning Using Business Analytics

Data Descriptions

(for all editions: 3rd and 4th Editions, R, Python, RapidMiner)

[Organized by alphabetical order of dataset name]

Accidents

These data, from the U.S. Bureau of Transportation Statistics, can be used to predict whether an accident will result in injuries or fatalities, based on predictors such as alcohol involvement, time of day, road condition, etc. Such a prediction system could be used to prioritize responder resources at the time of the report.

Source: US Dept. of Transportation, Bureau of Transportation Statistics, "TranStats," (www.transtats.bts.gov -- select "databases" then "General Estimate System (GES))
http://www.transtats.bts.gov/Fields.asp?Table_ID=1158&SYS_Table_Name=T_GES_ACCIDENT&User_Table_Name=Accident&Year_Info=1&First_Year=1999&Last_Year=2001&Rate_Info=1&Frequency=Annual&Data_Frequency=Annual,Monthly&Map_Info=&Is_Survey=1&Univ_Filter=&Latest_Available_Data=2001

Note: TranStats reports both variables with missing data, and their derived counterparts with imputed values filled in, denoted by an "I" at the end. Only one variant (the original or the derived) is included here.
An "R" at the end of the variable name indicates that the Transtats variable has been collapsed into fewer categories for analysis purposes
Data are for the year 2001.

	Variables	
1	HOUR_I_R	1=rush hour, 0=not (rush = 6-9 am, 4-7 pm)
2	ALCOHOL_I	Alcohol involved = 1, not involved = 2
3	ALIGN_I	1 = straight, 2 = curve
4	STRATUM_R	1= NASS Crashes Involving At Least One Passenger Vehicle, i.e., A Passenger Car, Sport Utility Vehicle, Pickup Truck Or Van) Towed Due To Damage From The Crash Scene And No Medium Or Heavy Trucks Are Involved. 0=not
5	WRK_ZONE	1= yes, 0= no
6	WKDY_I_R	1=weekday, 0=weekend
7	INT_HWY	Interstate? 1=yes, 0= no
8	LGTCON_I_R	Light conditions - 1=day, 2=dark (including dawn/dusk), 3=dark, but lighted,4=dawn or dusk
9	MAN_COL_I	0=no collision, 1=head-on, 2=other form of collision
10	PED_ACC_R	1=pedestrian/cyclist involved, 0=not

11	REL_JCT_I_R	1=accident at intersection/interchange, 0=not at intersection
12	REL_RWY_R	1=accident on roadway, 0=not on roadway
13	PROFIL_I_R	1= level, 0=other
14	SPD_LIM	Speed limit, miles per hour
15	SUR_CON	Surface conditions (1=dry, 2=wet, 3=snow/slush, 4=ice, 5=sand/dirt/oil, 8=other, 9=unknown)
16	TRAF_CON_R	Traffic control device: 0=none, 1=signal, 2=other (sign, officer …)
17	TRAF_WAY	1=two-way traffic, 2=divided hwy, 3=one-way road
18	VEH_INVL	Number of vehicles involved
19	WEATHER_R	1=no adverse conditions, 2=rain, snow or other adverse condition
20	NO_INJ_I	Number of injuries
21	PRPTYDMG_CRASH	1=property damage, 2=no property damage
22	FATALITIES	1= yes, 0= no
23	MAX_SEV_IR	0=no injury, 1=non-fatal inj., 2=fatal inj.

AdSales

Hypothetical data about advertising expenditures in one time period and sales in a subsequent time period.

Airfares

1. S_CODE: starting airport's code
2. S_CITY: starting city
3. E_CODE: ending airport's code
4. E_CITY: ending city
5. COUPON: average number of coupons (a one-coupon flight is a non-stop flight, a two-coupon flight is a one stop flight, etc.) for that route
6. NEW: number of new carriers entering that route between Q3-96 and Q2-97
7. VACATION: whether a vacation route (Yes) or not (No); Florida and Las Vegas routes are generally considered vacation routes
8. SW: whether Southwest Airlines serves that route (Yes) or not (No)
9. HI: Herfindel Index – measure of market concentration (refer to BMGT 681)
10. S_INCOME: starting city's average personal income
11. E_INCOME: ending city's average personal income
12. S_POP: starting city's population
13. E_POP: ending city's population
14. SLOT: whether either endpoint airport is slot controlled or not; this is a measure of airport congestion
15. GATE: whether either endpoint airport has gate constraints or not; this is another measure of airport congestion

16. DISTANCE: distance between two endpoint airports in miles
17. PAX: number of passengers on that route during period of data collection
18. FARE: average fare on that route

Amtrak

Ridership= Amtrak Ridership Number of Passengers (in thousands)

ApplianceShipments

Source:Data courtesy Ken Black

The series of quarterly shipments (in millions of dollars) of US household appliances between 1985 and 1989.

AustralianWines

Source: Website
Monthly Australian sales of wine Jan 1980 - Jul 1995

Bankruptcy

Source: "Predicting Corporate Bankruptcy"
Darden Business Publishing
Case authors Mark E. Haskins (HASKINSM@Darden.virginia.edu) and Phillip E. Pfeifer (PFEIFERP@Darden.virginia.edu)

NO	Arbitrary ID number for each firm.
D	D=0 for failed firms, D=1 for healthy firms.
YR	Year of Bankruptcy for failed firm in matched pair
R1	CASH/CURDEBT
R2	CASH/SALES
R3	CASH/ASSETS

R4	CASH/DEBTS
R5	CFF0/SALES
R6	CFFO/ASSETS
R7	CFFO/DEBTS
R8	COGS/INV
R9	CURASS/CURDEBT
R10	CURASS/SALES
R11	CURRASS/ASSETS
R12	CURDEBT/DEBTS
R13	INC/SALES
R14	INC/ASSETS
R15	INC/DEBTS
R16	UBCDEP/SALES
R17	INCDEP/ASSETS
R18	INCDEP/DEBTS
R19	SALES/REC
R20	SALES/ASSETS
R21	ASSETS/DEBTS
R22	WCFO/SALES
R23	WCFO/ASSETS
R24	WCFO/DEBTS

banks

Financial Condition	1 = financially weak
	0 = financially strong

BathSoapHousehold

Demographic Data

MEM	Member ID
SEC	Socio economic class (1 = high, 4 = low)
1	A
2	B
3	C
4	D/E

FEH	Food Eating Habits
1	Pure Vegetarian
2	Veg.But Serve Eggs
3	Non Vegetarian
0	Not Specified

MT	Native Language (mother tongue)
1	Assamese
2	Bengali
3	English
4	Gujarati
5	Hindi
6	Kannada
7	Kashmiri
8	Konkani
9	Malayalam
10	Marathi
11	Oriya
12	Punjabi
13	Rajasthani
14	Sindhi
15	Tamil
16	Telugu
17	Urdu
18	Sanskrit
19	Other
0	Not Specified

SEX	Sex of homemaker
1	Male
2	Female

AGE	Age of homemaker
1	Up to 24
2	25-34
3	35-44
4	45+

EDU	Education of homemaker
1	Illiterate
2	Literate, but no formal schooling
3	Up to 4 years of school
4	5-9 years of school

5	10-12 years of school
6	Some college
7	College graduate
8	Some graduate school
9	Graduate or professional school degree
0	Not specified

HS	Household size
	Number of people in the household

CHILD	Presence of children in household
1	Children up to age 6 present (only)
2	Children 7-14 present (only)
3	Both
4	None
5	Not specified

CS	Television
1	Cable or broadcast TV available
2	Unavailable

Affluence Index
Calculated from Durables sheet.

Purchase Summary Data

Labels	What they stand for
No. Brands	Number of brands purchased
Brand Runs	Number of runs (streaks) of purchasing the same brand
Total volume	Volume of product purchased (grams)
No. of trans.	Number of transactions
Value	Value in paise (100 paise = 1 rupee)
Avg. Price	Avg. price (rupees per 100 gram cake); computed from total volume and value
Purch. Vol. no promo	Percent of volume purchased not on promotion
Purch Vol. promo 6	Percent of volume purchased on promo code 6
Purch. Vol other promo	Percent of volume purchased on promo code other than 6

Brand Codelist (click here)

Price Codelist

1	ANY PREMIUM SOAPS
2	ANY POPULAR SOAP
3	ANY ECONOMY/CARBOLIC
4	ANY SUB-POPULAR

Promotion Codelist

1	Price off
2	Exchange Offer
3	Coupons
4	Extra grammage
5	Value added Pack
6	Banded Offer
7	Free gift
8	Others

Proposition Codelist

5	ANY BEAUTY
6	ANY HEALTH
7	ANY HERBAL
8	ANY FRESHNESS
9	ANY HAIR
10	ANY SKIN CARE
11	ANY FAIRNESS
12	ANY BABY
13	ANY GLYCERINE
14	ANY CARBOLIC
15	ANY OTHERS

Durable Ownership

Code	Durables	Affluence Weights
1	Radio/Transistor with FM	1
2	Radio/Transistor without FM	1
3	Stereo/Mono Tape Recorder	1
4	Two-in-one	2
5	Hi-Fi System/Music System without Compact disk	3

6	Hi-Fi System/Music System with Comapct disk	4
7	Walkman with FM	2
8	Walkman without FM	2
9	Discman with FM	3
10	Discman without FM	3
11	Video (VCP/VCR)	3
12	Laser Discs VCD/LD/DVD	5
13	TV - Black & White	2
14	Colour TV with remote	3
15	Colour TV without remote	3
16	Bicycle	1
17	Moped	2
18	Motorcycle	8
19	Scooter	5
20	Electric/Immersion Water heater	1
21	LPG/Bio-Gas stove	1
22	Mixer/Grinder	2
23	Pressure Cooker	1
24	Toaster	1
25	Cooking Range	4
26	Refrigerator - Non Frost free	3
27	Refrigerator - Frost free	5
28	Automatic dish washer	6
29	Oven - Electric	4
30	Electric Pressure Cooker	2
31	Microwave Oven	5
32	Rice Cooker	2
33	Electric Irons	1
34	Geyser	1
35	Cameras (still)	2
36	Telephones (with NSD/STD/ISD)	3
37	Telephones (Local only)	2
38	"Air Coolers"	2
39	Vacuum cleaner	2
40	Air Conditioners	5
41	Water purifier (Aquaguard etc.)	1
42	Washing Machines (Rs.5000+) Semi Automatic	4
43	Washing Machines (Rs.5000+) Fully Automatic	5
44	Washing Machines (Rs.5000+) Front Loading	6

45	Washing Machines (Rs.5000+)	
	Top Loading	5
46	Mobil/Cellular phone	4
47	Pager	2
48	Personal/Home Computers	8
49	Computer Printers	6
50	Fax Machine	6
51	Video camera/Handycam	6
52	Radio Clock	2
53	Deep Freezer	5
54	Electirc Kettle	1
55	Dish Washing Machine	5
56	Kitchen Sink	1
57	Floor Polisher	1
58	Cars/Jeeps/Vans	8
59	Auto Rickshaw	3
60	Tractors	5
61	Oven-In Built Range	5
62	Oven Ordinary Box (Gas)	3
63	Electric Table Fan	1
64	Electric Ceiling Fan	1
65	Torch	1
66	Sewing Machine	2
67	Generator	5
68	Pump Set/Water Pump	5

Not used:
Product Codelist

02	Toilet Soaps
05	Tooth Paste/Powder
01	Washing Soaps/Detergents
21	Washing Powder
45	Skin Creams
20	Edible Oils/Ghee/Vanaspati

Bicup2006

Source: Oct. 2006 public business intelligence competition
http://www.tis.cl/2007/futurosTalleres/_2006/Taller_1/BICUP2006-ENGLISH/

Data are the number of customers appearing at a bus terminal during 15 minute periods beginning at the specified time periods

Book Purchases

Columns indicate book categories, cells indicate whether a book in that category was purchased.

BostonHousing

This dataset contains information collected by the US Census Service concerning housing in the area of Boston Massachusetts. It was obtained from the StatLib archive (http://lib.stat.cmu.edu/datasets/boston). The dataset has 506 cases.
Source: The data was originally published by Harrison, D. and Rubinfeld, D.L. `Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.
There are 14 attributes in each case of the dataset. They are:

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town.
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000

CanadianWorkHours

average annual number of weekly hours spent by Canadian manufacturing workers

Source: Ken Black (used by permission)

CatalogCrossSell

Multi-Division Catalog Company
Scenario - A random sample of customers is shown in the Data sheet. A "1" indicates a purchase has been made from a catalog in that division, a "0" indicates no purchase.
Source: Adapted from a set of cases provided for educational purposes by the Direct Marketing Education Foundation; used with permission.

Cereals

Source: DATA ANALYSIS FOR STUDENT LEARNING (DASL)

1. Name: Name of cereal
2. mfr: Manufacturer of cereal where A = American Home Food Products; G = General Mills; K = Kelloggs; N = Nabisco; P = Post; Q = Quaker Oats; R = Ralston Purina
3. type: cold or hot
4. calories: calories per serving
5. protein: grams of protein
6. fat: grams of fat
7. sodium: milligrams of sodium
8. fiber: grams of dietary fiber
9. carbo: grams of complex carbohydrates
10. sugars: grams of sugars
11. potass: milligrams of potassium
12. vitamins: vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended
13. shelf: display shelf (1, 2, or 3, counting from the floor)
14. weight: weight in ounces of one serving
15. cups: number of cups in one serving
16. rating: a rating of the cereals calculated by Consumer Reports

CharlesBookClub

Source: Adapted with permission from The Bookbinders Club, prepared by Nissan Levin and Jacob Zahavi.

Variable	Description

Seq#	Sequence number in the sample
ID#	ID# in the full dataset
Gender	0=male, 1=female
M	Monetary - total money spent on books
R	Recency - Months since last purchase
F	Frequency - Total number of purchases
FirstPurch	Months since first purchase
Col H - R	book categories
Related Purchase	Number of related books purchased
Mcode, Rcode, Fcode	Recoding of M, R and F - see case description in DMBA

Cosmetics

Source: Statistics.com
A drug store chain wants to learn more about cosmetics buyers purchase patterns. Specifically, they want to know what items are purchased in conjunction with each other, for purposes of display, point of sale special offers, and to eventually implement a real time recommender system to cross-sell items at time of purchase.
The data (synthetic) are in the form of a matrix in which each column represents a product group, and each row a customer transaction.

Cosmetics-small

Source: Statistics.com
A drug store chain wants to learn more about cosmetics buyers purchase patterns. Specifically, they want to know what items are purchased in conjunction with each other, for purposes of display, point of sale special offers, and to eventually implement a real time recommender system to cross-sell items at time of purchase.
The data are in the form of a matrix in which each column represents a product group, and each row a customer transaction.
Note: Data are from Peter Bruce, partially drawn from a real source unrelated to cosmetics and partially generated.

Courserating

Source: Statistics.com

Student ratings of online statistics courses at Statistics.com

.

Coursetopics

Source: Statistics.com
Course topics at statistics.com (each row is a customer, column heads are topics taken [1] or not taken [0] by that customer)

DepartmentStoreSales

Data on the quarterly sales for a department store over a 6-year period.
Source = Chris Albright, used with permission

drug

EastWestAirlines or EastWestAirlinesCluster

East-West Airlines is trying to learn more about its customers. Key issues are their flying patterns, earning and use of frequent flyer rewards, and use of the airline credit card. The task is to identify customer segments via clustering.
Source: Based upon real business data; company names have been changed.

Field Name	Data Type	Max Data Length	Raw Data or Telcom Created Field?	Description
ID#	NUMBER		Telcom	Unique ID
Balance	NUMBER	8	Raw	Number of miles eligible for award travel
Qual_miles	NUMBER	8	Raw	Number of miles counted as qualifying for Topflight status
cc1_miles	CHAR	1	Raw	Number of miles earned with freq. flyer credit card in the past 12 months:
cc2_miles	CHAR	1	Raw	Number of miles earned with Rewards credit card in the past 12 months:
cc3_miles	CHAR	1	Raw	Number of miles earned with Small Business credit card in the past 12 months:
note: miles bins:				1 = under 5,000 2 = 5,000 - 10,000 3 = 10,001 - 25,000 4 = 25,001 - 50,000

				5 = over 50,000
Bonus_miles	NUMBER		Raw	Number of miles earned from non-flight bonus transactions in the past 12 months
Bonus_trans	NUMBER		Raw	Number of non-flight bonus transactions in the past 12 months
Flight_miles_12mo	NUMBER		Raw	Number of flight miles in the past 12 months
Flight_trans_12	NUMBER		Raw	Number of flight transactions in the past 12 months
Days_since_enroll	NUMBER		Telcom	Number of days since Enroll_date
Award?	NUMBER	Telcom		Dummy variable for Last_award (1=not null, 0=null)

EastWestAirlineNN

East-West Airlines has entered into a partnership with the wireless phone company Telcon to sell the latter's service via direct mail. These are a sample of data, provided so that the analyst can develop a model to classify East-West customers as to whether they purchase a wireless phone service contract (target variable Phone_sale).
Source: Based upon a real business case and real data; company names have been changed.

Field Name	Data Type	Max Data Length	Raw Data or Telcom Created Field?	Description
ID#	NUMBER		Telcom	Unique ID
Balance	NUMBER	8	Raw	Number of miles eligible for award travel
Qual_miles	NUMBER	8	Raw	Number of miles counted as qualifying for Topflight status
cc1_miles	CHAR	1	Raw	Number of miles earned with freq. flyer credit card in the past 12 months:
cc2_miles	CHAR	1	Raw	Number of miles earned with Rewards credit card in the past 12 months:
cc3_miles	CHAR	1	Raw	Number of miles earned with Small Business credit card in the past 12 months:
note: miles bins:				1 = under 5,000 2 = 5,000 - 10,000 3 = 10,001 - 25,000 4 = 25,001 - 50,000 5 = over 50,000
Bonus_miles	NUMBER		Raw	Number of miles earned from non-flight bonus transactions in the past 12 months
Bonus_trans	NUMBER		Raw	Number of non-flight bonus transactions in the past 12 months
Flight_miles_12mo	NUMBER		Raw	Number of flight miles in the past 12 months
Flight_trans_12	NUMBER		Raw	Number of flight transactions in the past 12 months
Email	CHAR	1	Raw	E-mail address on file. 1= yes, 0 =no?
Club_member	NUMBER		Telcom	Member of the airline's club (paid membership), 1=yes, 0=no
Any_cc_miles_12mo	NUMBER		Telcom	Dummy variable indicating whether member added miles on any credit card type within the past 12 months (1='Y', 0='N')
Phone_sale	NUMBER		Telcom	Dummy variable indicating whether member purchased Telcom service as a result of the direct mail campaign (1=sale, 0=no sale)

eBayAuctions

Source: Compiled from eBay.com for the period May-June 2004.

Variable descriptions

Category :	Category of the auctioned item.
currancy:	
sellerRating:	a rating by eBay, as a function of the number of "good" and "bad" transactions the seller had on eBay.
Duration :	Number of days the auction lasted (set by seller at auction start)
endDay :	Day of week that the auction closed
ClosePrice :	Price item sold at (converted into USD)
OpenPrice :	Initial price set by the seller (converted into USD)
Competitive? :	whether the auction had a single bid (0) or more (1)

EuropeanJobs

Data labels

1. Country:	Name of country
2. Agr:	Percentage employed in agriculture
3. Min:	Percentage employed in mining
4. Man:	Percentage employed in manufacturing
5. PS:	Percentage employed in power supply industries
6. Con:	Percentage employed in construction
7. SI:	Percentage employed in service industries
8. Fin:	Percentage employed in finance
9. SPS:	Percentage employed in social and personal services
10. TC:	Percentage employed in transport and communications

Faceplate

Synthetic Data on Purchases of Phone Faceplates.

Farm-ads

Data on advertisements posted at a website that caters to the needs of a specific farming community. Each ad is in a row, and each ad labeled as either −1 (not relevant) or 1 (relevant). The goal is to develop a predictive model that can classify ads automatically.

fiftytransactions

A small database of 50 transactions, where each of the nine items is assigned randomly to each transaction.

FlightDelays

Source: Bureau of Transportation Statistics
Variable explanations are in comments appended to column heads.
Note the data has both scheduled and actual departure time - pay attention to which you use!

All flights out of 3 DC airports (WAS)
into 3 NYC airports
not cancelled
flights in January 2004

Data labels:

CRS_DEP_TIME	scheduled departure time
CARRIER	The airline
DEP_TIME	Actual departure time
DEST	Destination airport in NY: Kennedy (JFK), LaGuardia (LGA), Newark (EWR)
DISTANCE	Flight distance in miles
FL_DATE	Flight date
FL_NUM	Flight number
ORIGIN	Departure airport in Washingon DC: National (DCA), Baltimore-Washington (BWI), Dulles (IAD)
Weather	Whether the weather was inclement (1) or not (0)
DAY_WEEK	Day of week. 1=Mon, 2=Tues...
DAY_OF_MONTH	
TAIL_NUM	This number is airplane specific

Flight Status	Whether the flight was delayed or on time (defined as arriving within 15 min of scheduled time)

Carrier Code	Carrier Name
AA	American Airlines, Inc.
CO	Continental Air Lines, Inc.
DH	Atlantic Coast Airlines
DL	Delta Air Lines, Inc.
EV	Atlantic Southeast Airlines
FL	Airtran Airways Corporation
MQ	American Eagle Airlines,inc
OH	Comair, Inc.
RU	Continental Express Airline
UA	United Air Lines, Inc.
US	US Airways, Inc.

Fundraising

ZIP: Zipcode group (zipcodes were grouped into 5 groups; only 4 are needed for analysis since if a potential donor falls into none of the four he or she must be in the other group. Inclusion of all five variables would be redundant and cause some modeling techniques to fail. A "1" indicates the potential donor belongs to this zip group.)

- 00000-19999 => 1 (omitted for above reason)
- 20000-39999 => zipconvert_2
- 40000-59999 => zipconvert_3
- 60000-79999 => zipconvert_4
- 80000-99999 => zipconvert_5

HOMEOWNER	1 = homeowner, 0 = not a homeowner
NUMCHLD	Number of children
INCOME	Household income
GENDER	Gender: 0 = Male 1 = Female
WEALTH	Wealth Rating (Wealth rating uses median family income and population statistics from each area to index relative wealth within each state. The segments are denoted 0-9, with 9 being the highest wealth group and zero being the lowest. Each rating has a different meaning within each state.)
HV	Average Home Value in potential donor's neighborhood in $ hundreds
ICmed	Median Family Income in potential donor's neighborhood in $ hundreds

ICavg	Average Family Income in potential donor's neighborhood in hundreds
IC15	Percent earning less than 15K in potential donor's neighborhood
NUMPROM	Lifetime number of promotions received to date
RAMNTALL	Dollar amount of lifetime gifts to date
MAXRAMNT	Dollar amount of largest gift to date
LASTGIFT	Dollar amount of most recent gift
TOTALMONTHS	Number of months from last donation to July 1998 (the last time the case was updated)
TIMELAG	Number of months between first and second gift
AVGGIFT	Average dollar amount of gifts to date
TARGET_B	
1 = Donor	
0 = Non-donor	
TARGET_D	Target Variable: Donation Amount (in $). We will NOT use it.

gdp

DATA FROM VEENHOVEN'S WORLD DATABASE OF HAPPINESS.
http://data.worldbank.org/indicator/NY.GDP.MKTP.CD
World Development Indicators.
Gross domestic product of the countries.

GermanCredit

Codelist (available in the textbook)

Variable Name	Description	Variable Type	Code Description
OBS#	Observation No.	Categorical	
CHK_ACCT	Checking account status	Categorical	0: < 0 DM 1: 0 < ...< 200 DM 2: => 200 DM 3: no checking account
DURATION	Duration of credit in months	Numerical	
HISTORY	Credit history	Categorical	0: no credits taken 1: all credits at this bank paid back duly 2: existing credits paid back duly till now 3: delay in paying off in the past 4: critical account
NEW_CAR	Purpose of credit	Binary	car (new) 0: No, 1: Yes
USED_CAR	Purpose of credit	Binary	car (used) 0: No, 1: Yes
FURNITURE	Purpose of credit	Binary	furniture/equipment 0: No, 1: Yes
RADIO/TV	Purpose of credit	Binary	radio/television 0: No, 1: Yes

EDUCATION	Purpose of credit	Binary	education 0: No, 1: Yes
RETRAINING	Purpose of credit	Binary	retraining 0: No, 1: Yes
AMOUNT	Credit amount	Numerical	
SAV_ACCT	Average balance in savings account	Categorical	0 : < 100 DM 1 : 100<= ... < 500 DM 2 : 500<= ... < 1000 DM 3 : =>1000 DM 4 : unknown/ no savings account
EMPLOYMENT	Present employment since	Categorical	0 : unemployed 1 : < 1 year 2 : 1 <= ... < 4 years 3 : 4 <=... < 7 years 4 : >= 7 years
INSTALL_RATE	Installment rate as % of disposable income	Numerical	
MALE_DIV	Applicant is male and divorced	Binary	0: No, 1: Yes
MALE_SINGLE	Applicant is male and single	Binary	0: No, 1: Yes
MALE_MAR_WID	Applicant is male and married or a widower	Binary	0: No, 1: Yes
CO-APPLICANT	Application has a co-applicant	Binary	0: No, 1: Yes
GUARANTOR	Applicant has a guarantor	Binary	0: No, 1: Yes
PRESENT_RESIDENT	Present resident since-years	Categorical	0: <= 1 year 1<…<=2 years 2<…<=3 years 3:>4years
REAL_ESTATE	Applicant owns real estate	Binary	0: No, 1: Yes
PROP_UNKN_NONE	Applicant owns no property (or unknown)	Binary	0: No, 1: Yes
AGE	Age in years	Numerical	
OTHER_INSTALL	Applicant has other installment plan credit	Binary	0: No, 1: Yes
RENT	Applicant rents	Binary	0: No, 1: Yes
OWN_RES	Applicant owns residence	Binary	0: No, 1: Yes
NUM_CREDITS	Number of existing credits at this bank	Numerical	
JOB	Nature of job	Categorical	0: unemployed/ unskilled - non-resident 1: unskilled - resident 2: skilled employee / official 3: management/ self-employed/highly qualified employee/ officer
NUM_DEPENDENTS	Number of people for whom liable to provide maintenance	Numerical	
TELEPHONE	Applicant has phone in his or her name	Binary	0: No, 1: Yes
FOREIGN	Foreign worker	Binary	0: No, 1: Yes
RESPONSE	Credit rating is good	Binary	0: No, 1: Yes

Hair-Care-Product

Fictional data representing an uplift study. A promotion for a hair color product was sent out to a sample of potential customers.

Promotional literature about a hair care product was sent to members of a buyers club. The goal is to determine which groups are most likely to make increased purchases as a result of receiving the promotion.
Source: SAS Institute, used by permission.
Worksheets:
Hair Care Product_original - This worksheet contains original hair care product data of size 1,26,184.
Hair Care Product_sample - This worksheet contains a sample dataset of size 10,000, sampled (without replacement) from the original dataset of size 1,26,184.
Data_for_analysis - This worksheet contains the sample dataset of size 10,000, but with variables Promotion(Yes/No), Gender(Male/Female) and Residence(Urban/Rural) recoded as Promotion(1/0), Gender(1/0) and Residence(1/0) respectively.

LaptopSales

Date	purchase date
Configuration	A numerical code representing a combination of screen size, battery life, RAM, etc. Each code corresponds to a particular combination.
Customer Postcode	postcode in London of the customer
Store Postcode	postcode in London of the store
Retail Price	price of laptop in GBP
Screen Size	screen size of laptop (Inches)
Battery Life	battery life of laptop (Hours)
RAM	RAM size of laptop(GB)
Processor Speeds	processor speed of laptop (GHz)
Integrated Wireless?	whether the laptop has integrated wireless or not
HD Size	HD size of laptop (GB)
Bundled Applications?	whether the laptop comes with bundled applications or not
customer X	X geo coordinates for customer location.
customer Y	Y geo coordinates for customer location.
store X	X geo coordinates for store location
store Y -	Y geo coordinates for store location

LaptopSalesJanuary2008

This is a subset of the Laptop sales dataset. It includes only the Jan 2008 sales (the complete dataset includes the entire 2008 sales).

Source: The laptop sales data were part of the ENBIS 2009 Challenge in Industrial Statistics

MortgageDefaulters

This data set contains data on mortgages that have been approved by bank underwriters.

Variable	Explanation
Bo_Age	Borrower age
Ln_Orig	Value of loan, USD
Orig_LTV_Ratio_Pct	Ratio of loan to home purchase price
Credit_score	Borrower's credit score
First_home	First time home buyer? (Y/N)
Tot_mthly_debt_exp	Borrower's total monthly debt expense
Tot_mthly_incm	Borrower's total monthly income
orig_apprd_val_amt	Appraised value of home at origination
pur_prc_amt	Purchase price for house
DTI_ratio	Borrower debt to income ratio (Tot_mthly_debt_exp/Tot_mthly_incm)
Status	Current loan status
OUTCOME	Binary version of "Status" (either default or non-default)
State	US state in which home is located
Median_state_inc	Median household income by state 2002-2004
UPB>Appraisal	Loan amount (Ln_Orig) greater than appraisal (orig_apprd_val_amt) 0-no, 1=yes

Note that some of the above variables were derived from combinations of two others.

Pharmaceuticals

Source: compiled from various web sources

RidingMowers

Source: Data courtesy of Dean Wichern.

Income:	Annual income in $000
Lot Size:	In thousands of sq. feet
Ownership:	Whether the resident owns a riding mower or not

Sept11Travel

Source: Bureau of Transportation Statistics - https://goo.gl/w2lJPV

AirRMP	Air revenue passenger miles (1 RMP is one revenue passenger carried for one mile)
RailPM	Rail passenger miles
VMT	Vehicle miles traveled

ShampooSales

Data on the monthly sales of a certain shampoo over a 3-year period.
Source: Time Series Data Library, http://data.is/TSDLdemo

SouvenirSales

Monthly sales for a souvenir shop at a beach resort town in Queensland, Australia, between 1995–2001.
Source: Time Series Data Library, http://data.is/TSDLdemo

SP500

Close=Monthly closing prices of S&P500

Spambase

Source: UCI Machine Learning Repository, HP database of emails

Each of the words below are columns in the data and the values represent % of words in the e-mail that match that particular word. For example, make represent % of words in the e-mail that match "make".

make address all W_3d our over remove internet order mail receive will people report addresses free business email you credit your font W_000 money hp hpl george W_650 lab labs telnet W_857 data W_415 W_85 technology W_1999 parts pm direct cs meeting original project re: edu table conference C; C(C[C! C$ C#

CAP_avg - average length of uninterrupted sequences of capital letters
CAP_long - length of longest uninterrupted sequence of capital letters
CAP_tot - total number of capital letters in the e-mail
Spam - 1 = spam, 0 = not spam

SystemAdministrators

Source: Samprit Chatterjee

Variables

Experience	-	measures months of full-time system administrator experience
Training	-	measures the number of relevant training credits
Completed task	-	either Yes or No, according to whether or not the administrator completed the tasks

Taxi-cancellation-case

The data are a randomly selected subset of the original data, with 10,000 rows, one row for each booking of a taxi. There are 17 input variables, including user (customer) ID, vehicle model, whether the booking was made online or via a mobile app, type of travel, type of booking package, geographic information, and the date and time of the scheduled trip. The target variable of interest is the binary indicator of whether a ride was canceled.

tinydata

Data includes information on a tasting score for a certain processed cheese. The two predictors are scores for fat and salt, indicating the relative presence of fat and salt in the particular cheese sample (where 0 is the minimum amount possible in the manufacturing process, and 1 the maximum). The outcome variable is the cheese sample's consumer taste preference, where like or dislike indicate whether the consumer likes the cheese or not.

Tayko

Codelist

Var. #	Variable Name	Description	Variable Type	Code Description
1.	US	Is it a US address?	binary	1: yes 0: no
2 - 16	Source_*	Source catalog for the record (15 identified sources plus one "other source" category; 15 dummies created with "other" as the reference, hence omitted.)	binary	1: yes 0: no
17.	Freq.	Number of transactions in last year at source catalog	numeric	
18.	last_update_days_ago	How many days ago was last update to cust. record	numeric	
19.	1st_update_days_ago	How many days ago was 1st update to cust. record	numeric	
20.	Web_order	Customer placed at least 1 order via web	binary	1: yes 0: no
21.	Gender=mal	Customer is male	binary	1: yes 0: no
22.	Address_is_res	Address is a residence	binary	1: yes 0: no
23.	Purchase	Person made purchase in test mailing	binary	1: yes 0: no
24.	Spending	Amount spent by customer in test mailing ($)	numeric	

Textiles

Codelist

ID	case number	
SILKWT	silk weight	
ZARIWT	zari weight	
SILKWT_Cat		categorical version of SILKWT
ZARIWT_Cat		categorical version of ZARIWT
BODYCOL	body color	
BODYCOL_*	body color	series of binary variables, 1 = body is that color
BRDCOL	border color	
BRDCOL_*	border color	series of binary variables, 1 = border is that color
BODYSHD_*	body shade	1 = pale, 4 = bright
BRDSHD	border shade	
BRDSHD_*	border shade	1 = pale, 4 = bright
SARSIDE	1 or 2 sided sari	1 = 1-sided, 2 = 2-sided
BODYDES	body design	
BODYDES_*	body design	series of binary variables, 1 = body is that design
BRDDES	border design	
BRDDES_*	border design	series of binary variables, 1 = border is that design
PALDES	pallav design	
PALDES_*	pallav design	series of binary variables, 1 = border is that design
BRDSZ	border size	
PALSZ	pallav size	
SALE	1 = sale, 0 = no sale	

Note: The colors and designs selected for the binary variables were those that were most common.

ToyotaCorolla

Variable	Description
Id	Record_ID
Model	Model Description
Price	Offer Price in EUROs
Age_08_04	Age in months as in August 2004
Mfg_Month	Manufacturing month (1-12)
Mfg_Year	Manufacturing Year

KM	Accumulated Kilometers on odometer
Fuel_Type	Fuel Type (Petrol, Diesel, CNG)
HP	Horse Power
Met_Color	Metallic Color? (Yes=1, No=0)
Color	Color (Blue, Red, Grey, Silver, Black, etc.)
Automatic	Automatic ((Yes=1, No=0)
CC	Cylinder Volume in cubic centimeters
Doors	Number of doors
Cylinders	Number of cylinders
Gears	Number of gear positions
Quarterly_Tax	Quarterly road tax in EUROs
Weight	Weight in Kilograms
Mfr_Guarantee	Within Manufacturer's Guarantee period (Yes=1, No=0)
BOVAG_Guarantee	BOVAG (Dutch dealer network) Guarantee (Yes=1, No=0)
Guarantee_Period	Guarantee period in months
ABS	Anti-Lock Brake System (Yes=1, No=0)
Airbag_1	Driver_Airbag (Yes=1, No=0)
Airbag_2	Passenger Airbag (Yes=1, No=0)
Airco	Airconditioning (Yes=1, No=0)
Automatic_airco	Automatic Airconditioning (Yes=1, No=0)
Boardcomputer	Boardcomputer (Yes=1, No=0)
CD_Player	CD Player (Yes=1, No=0)
Central_Lock	Central Lock (Yes=1, No=0)
Powered_Windows	Powered Windows (Yes=1, No=0)
Power_Steering	Power Steering (Yes=1, No=0)
Radio	Radio (Yes=1, No=0)
Mistlamps	Mistlamps (Yes=1, No=0)
Sport_Model	Sport Model (Yes=1, No=0)
Backseat_Divider	Backseat Divider (Yes=1, No=0)
Metallic_Rim	Metallic Rim (Yes=1, No=0)
Radio_cassette	Radio Cassette (Yes=1, No=0)
Parking_Assistant	Parking assistance system (Yes=1, No=0)
Tow_Bar	Tow Bar (Yes=1, No=0)

ToysRUsRevenues

The quarterly revenues of Toys “R” Us between 1992 and 1995
Source: Chris Albright

UniversalBank

Courtesy - Statistics.com

Data Description:

ID	Customer ID
Age	Customer's age in completed years
Experience	#years of professional experience
Income	Annual income of the customer ($000)
ZIPCode	Home Address ZIP code.
Family	Family size of the customer
CCAvg	Avg. spending on credit cards per month ($000)
Education	Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
Mortgage	Value of house mortgage if any. ($000)
Personal Loan	Did this customer accept the personal loan offered in the last campaign?
Securities Account	Does the customer have a securities account with the bank?
CD Account	Does the customer have a certificate of deposit (CD) account with the bank?
Online	Does the customer use internet banking facilities?
CreditCard	Does the customer use a credit card issued by UniversalBank?

Note: Data are synthetic

Universities

The dataset on American college and university rankings (available from www.dataminingbook.com) contains information on 1302 American colleges and universities offering an undergraduate program. For each university, there are 17 measurements that include continuous measurements (such as tuition and graduation rate) and categorical measurements (such as location by state and whether it is a private or a public school).

Source: Compiled from US News and World Report rankings on 1302 American Colleges and Universities

Utilities

Variable	Description
Company	Company name
Fixed_charge	Fixed-charge coverage ratio (income/debt)
RoR	Percent rate of return on capital
Cost	Cost per KW capacity in place
Load_factor	Annual load factor
Demand_growth	Percent demand growth
Sales	Sales (KWH use per year)
Nuclear	Percent nuclear
Fuel_Cost	Total fuel costs (cents per KWH)

Veerhoven

Data measuring happiness of countries. according to a 2006 Gallup survey.

Voter-Persuasion

Source: Ken Strasma and HaystaqDNA
See separate dictionary sheet for variable descriptions.
These data and this method are used in the Uplift Case in the Cases chapter.

WalMartStock

The series of Walmart daily closing prices between February 2001 and February 2002.publicly available, for example, at http: //finance.yahoo.com.
These data are also used in "Data Analysis for Managers" by Albright, Winston & Zappe.

West Roxbury

Variable	Description
TOTAL VALUE	Total assessed value for property, in thousands of USD
TAX	Tax bill amount based on total assessed value multiplied by the tax rate
LOT SQFT	Total lot size of parcel in square feet

YR BUILT	Year property was built
GROSS AREA	Gross floor area
LIVING AREA	Total living area for residential properties (ft2)
FLOORS	Number of floors
ROOMS	Total number of rooms
BEDROOMS	Total number of bedrooms
FULL BATH	Total number of full baths
HALF BATH	Total number of half baths
KITCHEN	Total number of kitchens
FIREPLACE	Total number of fireplaces
REMODEL	When house was remodeled (Recent/Old/None)

Wine

Wine dataset contains properties of wine captured from three different wineries in the same region. There are 13 variables describing various properties of wine and 3 classes. This dataset can be used for classification with Type as a output variable OR can be used to perform clustering to without using Type variable to see the accuracy of prediction.

This data set can be found in the UCI Machine Learning Repository
(http://www.ics.uci.edu/~mlearn/MLSummary.html or
ftp://ftp.ics.uci.edu/pub/machine-learning-databases/wine/)