Package 'mlba'

December 16, 2022

Type Package

Title MLBA - Machine Learning for Business Analytics Concepts, Techniques, and Applications in R

Version 2.0.0

Author Peter Gedeck

Maintainer Peter Gedeck <peter.gedeck@gmail.com>

Description Source code and datasets for
``Machine Learning for Business Analytics Concepts, Techniques, and Applications in R"
by Galit Shmueli, Peter C. Bruce, Peter Gedeck, Inbal Yahav, Nitin R. Patel.
Publisher: Wiley; 2nd edition (February, 2023) ISBN: 978-1-118-83517-2

License GPL (>= 2)

Encoding UTF-8

LazyData true

Depends R (>= 2.10), caret, forecast

RoxygenNote 7.2.3

StagedInstall no

LazyDataCompression bzip2

R topics documented:

`Accidents` *Accidents*

Description

Cleaned up and reduced version of the `AccidentsFull` dataset to facilitate model building.

Usage

```
Accidents
```

Format

A data frame with 600 observations and 11 variables:

RushHour 1 = rush hour, 0 = not (rush = 6-9 am, 4-7 pm)

WRK_ZONE 1 = yes, 0 = no

WKDY 1 = weekday, 0 = weekend

INT_HWY Interstate? 1 = yes, 0 = no

LGTCON_day Light conditions - 1 = day, 0 = other

LEVEL 1 = level, 0 = other

SPD_LIM Speed limit, miles per hour

SUR_COND_dry Surface conditions (1 = dry, 0 = other)

TRAF_two_way 1 = two-way traffic, 0 = other

WEATHER_adverse 0 = no adverse conditions, 1 = adverse condition

MAX_SEV one of: "no-injury", "non-fatal", "fatal"

Details

Data

See Also

`AccidentsFull` for the original dataset

Examples

```
Accidents
```

`Accidents1000` *Accidents1000*

Description

Cleaned up and reduced version of the `AccidentsFull` dataset to facilitate model building.

Usage

```
Accidents1000
```

Format

A data frame with 1000 observations and 11 variables:

RushHour 1 = rush hour, 0 = not (rush = 6-9 am, 4-7 pm)

WRK_ZONE 1 = yes, 0 = no

WKDY 1 = weekday, 0 = weekend

INT_HWY Interstate? 1 = yes, 0 = no

LGTCON_day Light conditions - 1 = day, 0 = other

LEVEL 1 = level, 0 = other

SPD_LIM Speed limit, miles per hour

SUR_COND_dry Surface conditions (1 = dry, 0 = other)

TRAF_two_way 1 = two-way traffic, 0 = other

WEATHER_adverse 0 = no adverse conditions, 1 = adverse condition

MAX_SEV one of: "no-injury", "non-fatal", "fatal"

Details

Data

See Also

`AccidentsFull` for the original dataset

Examples

```
Accidents1000
```

`AccidentsFull` *Accidents*

Description

These data, from the U.S. Bureau of Transportation Statistics, can be used to predict whether an accident will results 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.

Usage

```
AccidentsFull
```

Format

A data frame with 42183 observations and 24 variables:

HOUR_I_R 1 = rush hour, 0 = not (rush = 6-9 am, 4-7 pm)

ALCHL_I 1 = Alcohol involved, 2 = not involved

ALIGN_I 1 = straight, 2 = curve

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

WRK_ZONE 1 = yes, 0 = no

WKDY_I_R 1 = weekday, 0 = weekend

INT_HWY Interstate? 1 = yes, 0 = no

LGTCON_I_R Light conditions - 1 = day, 2 = dark (including dawn/dusk), 3 = dark, but lighted, 4 = dawn or dusk

MANCOL_I_R 0 = no collision, 1 = head-on, 2 = other form of collision

PED_ACC_R 1 = pedestrian/cyclist involved, 0 = not

RELJCT_I_R 1 = accident at intersection/interchange, 0 = not at intersection

REL_RWY_R 1 = accident on roadway, 0 = not on roadway

PROFIL_I_R 1 = level, 0 = other

SPD_LIM Speed limit, miles per hour

SUR_COND Surface conditions (1 = dry, 2 = wet, 3 = snow/slush, 4 = ice, 5 = sand/dirt/oil, 8 = other, 9 = unknown)

TRAF_CON_R Traffic control device: 0 = none, 1 = signal, 2 = other (sign, officer . . .)

TRAF_WAY 1 = two-way traffic, 2 = divided hwy, 3 = one-way road

VEH_INVL Number of vehicles involved

WEATHER_R 1 = no adverse conditions, 2 = rain, snow or other adverse condition

INJURY_CRASH 1 = yes, 0 = no

NO_INJ_I Number of injuries

PRPTYDMG_CRASH 1 = property damage, 2 = no property damage

FATALITIES 1 = yes, 0 = no

MAX_SEV_IR 0 = no injury, 1 = non-fatal inj., 2 = fatal inj.

Details

Data are for the year 2001.

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.

Examples

```
AccidentsFull
```

`AccidentsNN` *AccidentsNN*

Description

Subset of the ‘Accidents‘ dataset

Usage

```
AccidentsNN
```

Format

A data frame with 999 observations and 5 variables:

ALCHL_I 1 = Alcohol involved, 2 = not involved

PROFIL_I_R 1 = level, 0 = other

SUR_COND Surface conditions (1 = dry, 2 = wet, 3 = snow/slush, 4 = ice, 5 = sand/dirt/oil, 8 = other, 9 = unknown)

VEH_INVL Number of vehicles involved

MAX_SEV_IR 0 = no injury, 1 = non-fatal inj., 2 = fatal inj.

Details

Data

See Also

`AccidentsFull` for the full dataset

Examples

```
head(AccidentsNN)
```

`Airfares` *Airfares*

Description

Airfares

Usage

```
Airfares
```

Format

A data frame with 638 observations and 18 variables:

S_CODE starting airport's code

S_CITY starting city

E_CODE ending airport's code

E_CITY ending city

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

NEW number of new carriers entering that route between Q3-96 and Q2-97

VACATION whether a vacation route (Yes) or not (No); Florida and Las Vegas routes are generally considered vacation routes

SW whether Southwest Airlines serves that route (Yes) or not (No)

HI Herfindel Index – measure of market concentration (refer to BMGT 681)

S_INCOME starting city's average personal income

E_INCOME ending city's average personal income

S_POP starting city's population

E_POP ending city's population

SLOT whether either endpoint airport is slot controlled or not; this is a measure of airport congestion

GATE whether either endpoint airport has gate constraints or not; this is another measure of airport congestion

DISTANCE distance between two endpoint airports in miles

PAX number of passengers on that route during period of data collection

FARE average fare on that route

Source

Examples

```
head(Airfares)
```

Amtrak *Amtrak*

Description

Data

Usage

```
Amtrak
```

Format

A data frame with 159 observations and 2 variables:

Month

Ridership Amtrak Ridership Number of Passengers (in thousands)

Examples

```
Amtrak
```

ApplianceShipments *Applicance Shipments*

Description

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

Usage

```
ApplianceShipments
```

Format

A data frame with 20 observations and 2 variables:

Quarter

Shipments

Source

Data courtesy Ken Black

Examples

```
head(ApplianceShipments)
```

AustralianWines *AustralianWines*

Description

Monthly Australian sales of wine Jan 1980 - Jul 1995

Usage

```
AustralianWines
```

Format

A data frame with 180 observations and 7 variables:

month

fortified

red

rose

sparkling

sweet.white

dry.white

Details

Data

Source

Website

Examples

```
AustralianWines
```

AutosElectronics *AutosElectronics*

Description

Filename of data set for a classification task—to classify Internet discussion posts as either auto-related or electronics-related.

The data set contains two sets of messages; one with 1000 auto-related posts and one with 1000 electronics-related posts.

Usage

```
AutosElectronics
```

Format

An object of class `character` of length 1.

Value

the filename of the dataset that is installed with the mlba package

Source

The posts are taken from Internet groups devoted to autos and electronics

Examples

```
## Not run: corpus <- Corpus(ZipSource(AutosElectronics, recursive=True))
```

BankBiasData *BankBiasData*

Description

Data

Usage

```
BankBiasData
```

Format

A data frame with 1878 observations and 17 variables:

age

job type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')

marital marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree'

default has credit in default? (categorical: 'no','yes','unknown')

balance account balance

housing has housing loan? (categorical: 'no','yes','unknown')

loan has personal loan? (categorical: 'no','yes','unknown')

contact related with the last contact of the current campaign: contact communication type (categorical: 'cellular','telephone')

day related with the last contact of the current campaign: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

month related with the last contact of the current campaign: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

duration related with the last contact of the current campaign: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

campaign number of contacts performed during this campaign and for this client (numeric, includes last contact)

pdays number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

previous number of contacts performed before this campaign and for this client (numeric)

poutcome outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

y has the client subscribed a term deposit? (binary: 'yes','no')

Source

https://archive.ics.uci.edu/ml/datasets/bank+marketing

Examples

```
head(BankBiasData)
```

`Banks` *Banks*

Description

Data

Usage

```
Banks
```

Format

A data frame with 20 observations and 5 variables:

Obs

Financial.Condition 1 = financially weak; 0 = financially strong

TotCap.Assets

TotExp.Assets

TotLns.Lses.Assets

Examples

```
Banks
```

BareggTunnel *BareggTunnel*

Description

Data

Usage

```
BareggTunnel
```

Format

A data frame with 747 observations and 2 variables:

Day

Number.of.vehicles

Examples

```
BareggTunnel
```

BathSoapHousehold *BathSoapHousehold*

Description

Data

Usage

```
BathSoapHousehold
```

Format

A data frame with 600 observations and 46 variables:

Member.id Unique identifier for each household

SEC Socioeconomic class (1 = high, 5 = low)

FEH Eating habits(1 = vegetarian, 2 = vegetarian but eat eggs, 3 = nonvegetarian, 0 = not specified)

MT Native language (20 different languages)

SEX Gender of homemaker (1 = male, 2 = female)

AGE Age of homemaker

EDU Education of homemaker (1 = minimum, 9 = maximum)

HS Number of members in household

CHILD Presence of children in household (4 categories)

CS Television availability (1 = available, 2 = unavailable)

Affluence.Index Weighted value of durables possessed

No..of.Brands Number of brands purchased

Brand.Runs Number of instances of consecutive purchase of brands

Total.Volume Sum of volume

No..of..Trans Number of purchase transactions (multiple brands purchased in a month are counted as separate transactions)

Value Sum of value

Trans...Brand.Runs Average transactions per brand run

Vol.Tran Average volume per transaction

Avg..Price Average price of purchase

Pur.Vol.No.Promo.... Percent of volume purchased under no promotion

Pur.Vol.Promo.6.. Percent of volume purchased under promotion code 6

Pur.Vol.Other.Promo.. Percent of volume purchased under other promotions

Br..Cd..57..144 Percent of volume purchased of the brand

Br..Cd..55

Br..Cd..272

Br..Cd..286

Br..Cd..24

Br..Cd..481

Br..Cd..352

Br..Cd..5

Others.999

Pr.Cat.1 Percent of volume purchased under the price category

Pr.Cat.2

Pr.Cat.3

Pr.Cat.4

PropCat.5 Percent of volume purchased under the product proposition category

PropCat.6

PropCat.7

PropCat.8

PropCat.9

PropCat.10

PropCat.11

PropCat.12

PropCat.13

PropCat.14

PropCat.15

Source

Examples

```
head(BathSoapHousehold)
```

Bicup2006 *Bicup2006*

Description

Data from a 2008 business intelligence competition, which focused on forecasting demand for public transportation.

Data

Usage

```
Bicup2006
```

Format

A data frame with 1512 observations and 3 variables:

The file contains the historic information with known demand for a 3-week period, separated into 15-minute intervals, and dates and times for a future 3-day period (DEMAND = NaN), for which forecasts should be generated.

DATE date information

TIME time information

DEMAND demand information; NA if missing

Source

Ken Black (used by permission)

Examples

```
Bicup2006
```

BostonHousing *Boston Housing*

Description

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).

Usage

```
BostonHousing
```

Format

A data frame with 506 observations and 14 variables:

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
CAT.MEDV

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.

Examples

```
BostonHousing
```

CanadianWorkHours *CanadianWorkHours*

Description

Data

Usage

```
CanadianWorkHours
```

Format

A data frame with 35 observations and 2 variables:

Year
Hours average annual number of weekly hours spent by Canadian manufacturing workers

Source

Ken Black (used by permission)

Examples

```
CanadianWorkHours
```

CatalogCrossSell *CatalogCrossSell*

Description

CatalogCrossSell

Usage

```
CatalogCrossSell
```

Format

A data frame with 4998 observations and 10 variables:

Customer.Number Customer IDs. For each customer, the following columns describe if the customer made a purchase in the division

Clothing.Division

Housewares.Division

Health.Products.Division

Automotive.Division

Personal.Electronics.Division

Computers.Division

Garden.Division

Novelty.Gift.Division

Jewelry.Division

Source

The data for this case have been adapted from the data in a set of cases provided for educational purposes by the Direct Marketing Education Foundation ("DMEF Academic Data Set Two, Multi Division Catalog Company, Code: 02DMEF"); used with permission.

Examples

```
head(CatalogCrossSell)
```

Cereals *Cereals*

Description

Cereals

Usage

```
Cereals
```

Format

A data frame with 77 observations and 16 variables:

name Name of cereal

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

type cold or hot

calories calories per serving

protein grams of protein

fat grams of fat

sodium milligrams of sodium

fiber grams of dietary fiber

carbo grams of complex carbohydrates

sugars grams of sugars

potass milligrams of potassium

vitamins vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended

shelf display shelf (1, 2, or 3, counting from the floor)

weight weight in ounces of one serving

cups number of cups in one serving

rating a rating of the cereals calculated by Consumer Reports

Source

Data analysis for student learning (DASL)

Examples

```
head(Cereals)
```

`CharlesBookClub` *Charles Book Club*

Description

Data

Usage

```
CharlesBookClub
```

Format

A data frame with 4000 observations and 24 variables:

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
ChildBks this and following - book categories
YouthBks
CookBks
DoItYBks
RefBks
ArtBks
GeogBks
ItalCook
ItalAtlas
ItalArt
Florence
Related.Purchase Number of related books purchased
Mcode Recoding of M - see case description in DMBA and MLBA
Rcode Recoding of R - see case description in DMBA and MLBA
Fcode Recoding of F - see case description in DMBA and MLBA
Yes_Florence
No_Florence

Source

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

Examples

```
head(CharlesBookClub)
```

COMPAS_clean *COMPAS clean*

Description

Dataset derived from the ProPublica story "Machine Bias." See https://github.com/propublica/compas-analysis for details

Usage

```
COMPAS_clean
```

Format

A data frame with 5304 observations and 7 variables

id

age_cat categories: Less than 25, 25 - 45, Greater than 45

c_charge_degree

race

sex

priors_count

two_year_recid

Details

Data

Source

https://github.com/propublica/compas-analysis/blob/master/compas-scores-two-years.csv

Examples

```
head(COMPAS_clean)
```

Cosmetics *Cosmetics*

Description

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.

Usage

```
Cosmetics
```

Format

A data frame with 1000 observations and 15 variables:

Trans. Transaction #

Bag

Blush

Nail.Polish

Brushes

Concealer

Eyebrow.Pencils

Bronzer

Lip.liner

Mascara

Eye.shadow

Foundation

Lip.Gloss

Lipstick

Eyeliner

Details

The data (synthetic) are in the form of a matrix in which each column represents a product group, and each row a customer transaction.

Data

Source

statistics.com; Copyright 2016 Galit Shmueli and Peter Bruce

Examples

```
head(Cosmetics)
```

`CourseRating` *CourseRating*

Description

Student ratings of online statistics courses at Statistics.com

Usage

```
CourseRating
```

Format

A data frame with 15 observations and 10 variables:

X

SQL

Spatial

PA1

DM.in.R

Python

Forecast

R.Prog

Hadoop

Regression

Source

Copyright 2016 statistics.com

Examples

```
head(CourseRating)
```

CourseTopics *CourseTopics*

Description

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

Usage

```
CourseTopics
```

Format

A data frame with 365 observations and 8 variables:

Intro

DataMining

Survey

Cat.Data

Regression

Forecast

DOE

SW

Source

Copyright 2016 Galit Shmueli and Peter Bruce

Examples

```
head(CourseTopics)
```

`DepartmentStoreSales` *DepartmentStoreSales*

Description

Data on the quarterly sales for a department store over a 6-year period.

Usage

```
DepartmentStoreSales
```

Format

A data frame with 24 observations and 2 variables:

Quarter

Sales quarterly sales for the department store

Examples

```
DepartmentStoreSales
```

`Drug` *Drug*

Description

Data

Usage

```
Drug
```

Format

A data frame with 60 observations and 7 variables:

Entity

Related.Entity

Relationship

Descrption

Related.Entity.Address.1

Related.Entity.Address.2

Related.Address.3

Examples

```
Drug
```

`EastWestAirlinesCluster`
EastWestAirlinesCluster

Description

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.

Usage

```
EastWestAirlinesCluster
```

Format

A data frame with 3999 observations and 12 variables:

ID. Unique ID

Balance Number of miles eligible for award travel

Qual_miles Number of miles counted as qualifying for Topflight status

cc1_miles Number of miles earned with freq. flyer credit card in the past 12 months: 1 = under 5,000; 2 = 5,000 - 10,000; 3 = 10,001 - 25,000; 4 = 25,001 - 50,000; 5 = over 50,000

cc2_miles Number of miles earned with Rewards credit card in the past 12 months (see cc1_miles)

cc3_miles Number of miles earned with Small Business credit card in the past 12 months (see cc1_miles)

Bonus_miles Number of miles earned from non-flight bonus transactions in the past 12 months

Bonus_trans Number of non-flight bonus transactions in the past 12 months

Flight_miles_12mo Number of flight miles in the past 12 months

Flight_trans_12 Number of flight transactions in the past 12 months

Days_since_enroll Number of days since Enroll_date

Award. Dummy variable for Last_award (1=not null, 0=null)

Source

Based upon real business data; company names have been changed. Copyright 2016 Galit Shmueli and Peter Bruce

Examples

```
head(EastWestAirlinesCluster)
```

EastWestAirlinesNN *EastWestAirlinesNN*

Description

East-West Airlines has entered into a partnership with the wireless phone company Telcom 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').

Usage

```
EastWestAirlinesNN
```

Format

A data frame with 4987 observations and 16 variables:

ID. Unique ID

Topflight

Balance Number of miles eligible for award travel

Qual_miles Number of miles counted as qualifying for Topflight status

cc1_miles Number of miles earned with freq. flyer credit card in the past 12 months: 1 = under 5,000; 2 = 5,000 - 10,000; 3 = 10,001 - 25,000; 4 = 25,001 - 50,000; 5 = over 50,000

cc2_miles Number of miles earned with Rewards credit card in the past 12 months (see cc1_miles)

cc3_miles Number of miles earned with Small Business credit card in the past 12 months (see cc1_miles)

Bonus_miles Number of miles earned from non-flight bonus transactions in the past 12 months

Bonus_trans Number of non-flight bonus transactions in the past 12 months

Flight_miles_12mo Number of flight miles in the past 12 months

Flight_trans_12 Number of flight transactions in the past 12 months

Online_12

Email E-mail address on file. 1= yes, 0 =no?

Club_member Member of the airline's club (paid membership), 1=yes, 0=no

Any_cc_miles_12mo Dummy variable indicating whether member added miles on any creditcard type within the past 12 months (1='Y', 0='N')

Phone_sale Dummy variable indicating whether member purchased Telcom service as a result of the direct mail campaign (1=sale, 0=no sale)

Source

Based upon real business data; company names have been changed.

Examples

```
head(EastWestAirlinesNN)
```

eBayAuctions *ebay Auctions*

Description

Data

Usage

```
eBayAuctions
```

Format

A data frame with 1972 observations and 8 variables:

Category Category of the auctioned item.

currency

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)

Source

Examples

```
eBayAuctions
```

eBayNetwork *ebay Network*

Description

Data

Usage

```
eBayNetwork
```

Format

A data frame with 200 observations and 5 variables:

Seller

Bidder

Weight

Bidder.Volume

Seller.Volume

Examples

```
eBayNetwork
```

`eBayTreemap` *ebay Treemap*

Description

Data

Usage

```
eBayTreemap
```

Format

A data frame with 10078 observations and 5 variables:

High.Bid

Seller.Feedback

Category

Sub.Category

Brand

Examples

```
head(eBayTreemap)
```

EmailABtest *EmailABtest*

Description

Data

Usage

```
EmailABtest
```

Format

A data frame with 426 observations and 3 variables:

Customer

Message

Open.

Examples

```
head(EmailABtest)
```

Faceplate *Faceplate*

Description

Synthetic Data on Purchases of Phone Faceplates.

Usage

```
Faceplate
```

Format

A data frame with 10 observations and 7 variables:

Transaction

Red

White

Blue

Orange

Green

Yellow

Source

Examples

```
Faceplate
```

`FarmAds` *FarmAds*

Description

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.

Usage

```
FarmAds
```

Format

A data frame with 4143 observations and 2 variables:

label 1: ad is relevant; -1 ad is not relevant

text text of ad pre-processed into a list of words. The location of the word in the ad is in some cases indicated by a prefix.

Details

Data

Examples

```
FarmAds
```

`FlightDelays` *Flight delays*

Description

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

Usage

```
FlightDelays
```

Format

A data frame with 2201 observations and 13 variables:

CRS_DEP_TIME scheduled departure time

CARRIER The airline. 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.

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)

Source

Bureau of Transportation Statistics

Examples

```
head(FlightDelays)
```

`Fundraising` *Fundraising*

Description

Data

Usage

```
Fundraising
```

Format

A data frame with 3120 observations and 24 variables:

Row.Id

Row.Id.

zipconvert_2 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

zipconvert_3

zipconvert_4

zipconvert_5

homeowner.dummy 1 = homeowner, 0 = not a homeowner

NUMCHLD Number of children

INCOME Household income

gender.dummy 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 lifetime gifts 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.

Examples

```
head(Fundraising)
```

FutureFundraising *FutureFundraising*

Description

Data

Usage

```
FutureFundraising
```

Format

A data frame with 2000 observations and 24 variables:

Row.Id
Row.Id.
zipconvert_2
zipconvert_3
zipconvert_4
zipconvert_5
homeowner.dummy
NUMCHLD
INCOME
gender.dummy
WEALTH
HV
Icmed
Icavg
IC15
NUMPROM
RAMNTALL
MAXRAMNT
LASTGIFT
totalmonths
TIMELAG
AVGGIFT
TARGET_B
TARGET_D

See Also

See Fundraising for a description of the dataset

Examples

```
head(FutureFundraising)
```

GDP *GDP*

Description

Data

Usage

```
GDP
```

Format

A data frame with 264 observations and 5 variables:

Country.Name

Country.Code

Indicator.Name

Indicator.Code

GDP2015 Gross domestic product of the countries.

Source

Data from Veenhoven's world database of happiness. http://data.worldbank.org/indicator/NY.GDP.MKTP.CD. World Development Indicators.

Examples

```
head(GDP)
```

GermanCredit *German credit*

Description

Data

Usage

```
GermanCredit
```

Format

A data frame with 1000 observations and 32 variables:

OBS. Observation No.

CHK_ACCT Checking account status. 0: < 0 DM; 1: 0 <...< 200 DM; 2: => 200 DM; 3: no checking account

DURATION Duration of credit in months

HISTORY Credit history. 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. car (new) 0: No, 1: Yes

USED_CAR Purpose of credit. car (used) 0: No, 1: Yes

FURNITURE Purpose of credit. furniture/equipment 0: No, 1: Yes

RADIO.TV Purpose of credit. radio/television 0: No, 1: Yes

EDUCATION Purpose of credit. education 0: No, 1: Yes

RETRAINING Purpose of credit. retraining 0: No, 1: Yes

AMOUNT Credit amount

SAV_ACCT Average balance in savings account. 0 : < 100 DM; 1 : 100 <=...< 500 DM; 2 : 500 <=...< 1000 DM; 3 : =>1000 DM; 4 : unknown/ no savings account

EMPLOYMENT Present employment since. 0 : unemployed; 1 : < 1 year; 2 : 1 <= ... < 4 years 3 : 4 <=... < 7 years 4 : >= 7 years

INSTALL_RATE Installment rate as % of disposable income

MALE_DIV Applicant is male and divorced. 0: No, 1: Yes

MALE_SINGLE Applicant is male and single. 0: No, 1: Yes

MALE_MAR_or_WID Applicant is male and married or a widower. 0: No, 1: Yes

CO.APPLICANT Application has a co-applicant. 0: No, 1: Yes

GUARANTOR Applicant has a guarantor. 0: No, 1: Yes

PRESENT_RESIDENT Present resident since-years. 0: <= 1 year; 1: <. . . <= 2 years; 2: <. . . <= 3 years; 3: >4years

REAL_ESTATE Applicant owns real estate. 0: No, 1: Yes

PROP_UNKN_NONE Applicant owns no property (or unknown). 0: No, 1: Yes

AGE Age in years

OTHER_INSTALL Applicant has other installment plan credit. 0: No, 1: Yes

RENT Applicant rents. 0: No, 1: Yes

OWN_RES Applicant owns residence. 0: No, 1: Yes

NUM_CREDITS Number of existing credits at this bank

JOB Nature of job. 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

TELEPHONE Applicant has phone in his or her name. 0: No, 1: Yes

FOREIGN Foreign worker. 0: No, 1: Yes

RESPONSE Credit rating is good. 0: No, 1: Yes

Examples

```
head(GermanCredit)
```

HairCareProduct *Hair Care Product*

Description

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.

Usage

```
HairCareProduct
```

Format

A data frame with 10000 observations and 8 variables:

Purchase 1: purchased; 0: not purchased

Age age of customer

Hair.Color one of 'Black', 'Red', 'Blond', or 'Brown'

U.S..Region one of 'Southwest', 'Northwest', 'Northeast', or 'Southeast'

Validation 0: training set; 1: validation set

Promotion_ord 1: customer received a promotion; 0: did'n receive promotional material

Gender_ord 1: male; 0: female

Residence_ord 1: urban; 0: rural

Source

SAS Institute, used by permission; sample of 10,000 observations from full 126,184 dataset

Examples

```
head(HairCareProduct)
```

IMDBdataset10K *IMDBdataset10K*

Description

Collection of movie reviews from the Internet Movie Database (IMDB)

The dataset contains 5,000 positive and 5,000 negative movie reviews collected from the Internet Movie Database (IMDB) (Maas et al. 2011). The original large movie dataset 25K positive and 25K negative movie reviews. This dataset is a stratified sample of 10K records from this dataset.

Data

Usage

```
IMDBdataset10K
```

Format

A data frame with 10,000 observations and 2 variables:

review Movie review

sentiment Sentiment of the movie review categorized as "positive" or "negative"

Source

Original large movie dataset is published at http://ai.stanford.edu/~amaas/data/sentiment/

Examples

```
head(IMDBdataset10K)
```

LaptopSales *Laptop Sales*

Description

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

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

Data

Usage

```
LaptopSales
```

Format

A data frame with 297,572 observations and 16 variables:

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..Inches. screen size of laptop (Inches)

Battery.Life..Hours. battery life of laptop (Hours)

RAM..GB. RAM size of laptop(GB)

Processor.Speeds..GHz. processor speed of laptop (GHz)

Integrated.Wireless. whether the laptop has integrated wireless or not

HD.Size..GB. 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

Source

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

Examples

```
head(LaptopSales)
```

`LaptopSalesJanuary2008`

Laptop Sales January 2008

Description

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

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

Data

Usage

```
LaptopSalesJanuary2008
```

Format

A data frame with 7956 observations and 17 variables:

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..Inches. screen size of laptop (Inches)

Battery.Life..Hours. battery life of laptop (Hours)

RAM..GB. RAM size of laptop(GB)

Processor.Speeds..GHz. processor speed of laptop (GHz)

Integrated.Wireless. whether the laptop has integrated wireless or not

HD.Size..GB. HD size of laptop (GB)

Bundled.Applications. whether the laptop comes with bundled applications or not

OS.X.Customer X geo coordinates for customer location.

OS.Y.Customer Y geo coordinates for customer location.

OS.X.Store X geo coordinates for store location

OS.Y.Store Y geo coordinates for store location

CustomerStoreDistance

Source

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

See Also

`LaptopSales` for the full dataset

Examples

```
head(LaptopSalesJanuary2008)
```

`LiftExample` *LiftExample*

Description

Synthetic dataset to demonstrate lift

Usage

```
LiftExample
```

Format

A data frame with 24 observations and 2 variables:

prob

actual

Examples

```
head(LiftExample)
```

MovieLensMovies *MovieLensMovies*

Description

Movies from the Movielens 10k dataset

Usage

```
MovieLensMovies
```

Format

A data frame with 9,742 observations and 3 variables:

movieId Same as ratings data

title Movie titles, including the year of release in parentheses

genres Pipe-separated list of genres for corresponding movie title

Source

The Movielens dataset is downloaded from https://grouplens.org/datasets/movielens/. Dataset creation date: September 26, 2018). See https://files.grouplens.org/datasets/movielens/ml-latest-small-README.html for the usage license.

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4: 19:1–19:19. https://doi.org/10.1145/2827872

Examples

```
head(MovieLensMovies)
```

MovieLensRatings *MovieLensRatings*

Description

Ratings from the Movielens 10k dataset

Usage

```
MovieLensRatings
```

Format

A data frame with 100,836 observations and 4 variables:

userId Anonymized MovieLens user ID. Each user has rated at least 20 movies.

movieId Movie ID consistent with that used on the MovieLens website (e.g., information for movie id 1 is available at https://movielens.org/movies/1).

rating Movie rating provided by user on a 5-star scale, with half-star increments (0.5 star - 5 stars)

timestamp Timestamp when rating was recorded. Time is in seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970

Source

The Movielens dataset is downloaded from https://grouplens.org/datasets/movielens/. Dataset creation date: September 26, 2018). See https://files.grouplens.org/datasets/movielens/ml-latest-small-README.html for the usage license.

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4: 19:1–19:19. https://doi.org/10.1145/2827872

Examples

```
head(MovieLensRatings)
```

`NaturalGasSales` *NaturalGasSales*

Description

Data

Usage

```
NaturalGasSales
```

Format

A data frame with 16 observations and 2 variables:

Quarter

Gas.Sales

Examples

```
NaturalGasSales
```

`NYPDMotorVehicleCollisions`
NYPD Motor Vehicle Collisions

Description

The Motor Vehicle Collisions crash table contains details on the crash event. Each row represents a crash event. The Motor Vehicle Collisions data tables contain information from all police reported motor vehicle collisions in NYC. This dataset is a random subset of 1000 entries of the NYPD motor vehicle collisions dataset

Usage

```
NYPDMotorVehicleCollisions
```

Format

A data frame with 1000 observations and 29 variables:

DATE Occurrence date of collision

TIME Occurrence time of collision

BOROUGH Borough where collision occurred

ZIP.CODE Postal code of incident occurrence

LATITUDE Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

LONGITUDE Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

LOCATION Latitude , Longitude pair

ON.STREET.NAME Street on which the collision occurred

CROSS.STREET.NAME Nearest cross street to the collision

OFF.STREET.NAME Street address if known

NUMBER.OF.PERSONS.INJURED Number of persons injured

NUMBER.OF.PERSONS.KILLED Number of persons killed

NUMBER.OF.PEDESTRIANS.INJURED Number of pedestrians injured

NUMBER.OF.PEDESTRIANS.KILLED Number of pedestrians killed

NUMBER.OF.CYCLIST.INJURED Number of cyclists injured

NUMBER.OF.CYCLIST.KILLED Number of cyclists killed

NUMBER.OF.MOTORIST.INJURED Number of vehicle occupants injured

NUMBER.OF.MOTORIST.KILLED Number of vehicle occupants killed

CONTRIBUTING.FACTOR.VEHICLE.1 Factors contributing to the collision for designated vehicle

CONTRIBUTING.FACTOR.VEHICLE.2 Factors contributing to the collision for designated vehicle

CONTRIBUTING.FACTOR.VEHICLE.3 Factors contributing to the collision for designated vehicle

CONTRIBUTING.FACTOR.VEHICLE.4 Factors contributing to the collision for designated vehicle

CONTRIBUTING.FACTOR.VEHICLE.5 Factors contributing to the collision for designated vehicle

UNIQUE.KEY Unique record code generated by system. Primary Key for Crash table.

VEHICLE.TYPE.CODE.1 Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)

VEHICLE.TYPE.CODE.2 Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)

VEHICLE.TYPE.CODE.3 Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)

VEHICLE.TYPE.CODE.4 Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)

VEHICLE.TYPE.CODE.5 Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)

Source

Random sample of the NYPD Motor Vehicle Collisions Dataset. (https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95)

Examples

```
head(NYPDMotorVehicleCollisions)
```

`OwnerExample` *OwnerExample*

Description

Data

Usage

```
OwnerExample
```

Format

A data frame with 24 observations and 2 variables:

Class

Probability

Examples

```
head(OwnerExample)
```

`Pharmaceuticals` *Pharmaceuticals*

Description

Data

Usage

```
Pharmaceuticals
```

Format

A data frame with 21 observations and 14 variables:

Symbol

Name

Market_Cap

Beta

PE_Ratio

ROE

ROA

Asset_Turnover

Leverage

Rev_Growth

Net_Profit_Margin

Median_Recommendation

Location

Exchange

Source

Compiled from various web sources. Copyright 2016 Galit Shmueli and Peter Bruce

Examples

```
head(Pharmaceuticals)
```

regressionSummary *regressionSummary*

Description

Convenience function to calculate a variety of regression metrics

Usage

```
regressionSummary(predicted, actual)
```

Arguments

predicted	predicted values
actual	actual values

Value

list containing the regression metrics ME, RMSE and MAE

Examples

```
lm.SR <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings)
regressionSummary(predict(lm.SR, LifeCycleSavings), LifeCycleSavings$sr)
```

RidingMowers *Riding Mowers*

Description

Data

Usage

```
RidingMowers
```

Format

A data frame with 24 observations and 3 variables:

Income Annual income in $000

Lot_Size In thousands of sq. feet

Ownership Whether the resident owns a riding mower or not

Source

Data courtesy of Dean Wichern

Examples

```
head(RidingMowers)
```

SCstudents *statistics.com Students*

Description

Data

Usage

```
SCstudents
```

Format

A data frame with 1696 observations and 2 variables:

latitude

longitude

Examples

```
head(SCstudents)
```

Sept11Travel *Sept11Travel*

Description

Estimated Impacts of September 11th on US Travel

Usage

```
Sept11Travel
```

Format

A data frame with 172 observations and 4 variables:

Month

Air.RPM..000s. Air revenue passenger miles (1 RMP is one revenue passenger carried for one mile)

Rail.PM Rail passenger miles

VMT..billions. Vehicle miles traveled

Source

Bureau of Transportation Statistics: https://www.bts.gov/archive/publications/estimated_impacts_of_9_11_on_us_travel/index

Examples

```
Sept11Travel
```

ShampooSales *ShampooSales*

Description

Data on the monthly sales of a certain shampoo over a 3-year period.

Usage

```
ShampooSales
```

Format

A data frame with 36 observations and 2 variables:

Month

Shampoo.Sales

Source

Time Series Data Library, https://pkg.yangzhuoranyang.com/tsdl/

Examples

```
ShampooSales
```

SouvenirSales *SouvenirSales*

Description

Monthly sales for a souvenir shop at a beach resort town in Queensland, Australia, between 1995–2001.

Usage

```
SouvenirSales
```

Format

A data frame with 84 observations and 2 variables:

Date

Sales

Source

Time Series Data Library, https://pkg.yangzhuoranyang.com/tsdl/

Examples

```
SouvenirSales
```

SP500 *SP500*

Description

Monthly closing prices of S&P500

Usage

```
SP500
```

Format

A data frame with 100 observations and 2 variables:

Date

Close Monthly closing prices of S&P500

Examples

```
head(SP500)
```

Spambase *Spambase*

Description

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".

Usage

```
Spambase
```

Format

A data frame with 4601 observations and 58 variables:

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..1 C(

C..2 C[

C..3 C!

C..4 C$

C..5 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

Examples

```
head(Spambase)
```

`SystemAdministrators` *System Administrator*

Description

Data

Usage

```
SystemAdministrators
```

Format

A data frame with 75 observations and 3 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

Source

Samprit Chatterjee

Examples

```
head(SystemAdministrators)
```

TaxiCancellationCase *Taxi Cancellation Case*

Description

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.

Usage

```
TaxiCancellationCase
```

Format

A data frame with 10000 observations and 19 variables:

row.

user_id the ID of the customer (based on mobile number)

vehicle_model_id vehicle model type.

package_id type of package (1=4hrs & 40kms, 2=8hrs & 80kms, 3=6hrs & 60kms, 4= 10hrs & 100kms, 5=5hrs & 50kms, 6=3hrs & 30kms, 7=12hrs & 120kms)

travel_type_id type of travel (1=long distance, 2= point to point, 3= hourly rental).

from_area_id unique identifier of area. Applicable only for point-to-point travel and packages

to_area_id unique identifier of area. Applicable only for point-to-point travel

from_city_id unique identifier of city

to_city_id unique identifier of city (only for intercity)

from_date time stamp of requested trip start

to_date time stamp of trip end

online_booking if booking was done on desktop website

mobile_site_booking if booking was done on mobile website

booking_created time stamp of booking

from_lat latitude of from area

from_long longitude of from area

to_lat latitude of to area

to_long longitude of to area

Car_Cancellation 1=trip cancelled; 0=trip not cancelled

Source

Examples

```
head(TaxiCancellationCase)
```

Tayko *Tayko*

Description

Data

Usage

```
Tayko
```

Format

A data frame with 2000 observations and 25 variables:

sequence_number Unique identifier

US Is it a US address? 1=yes; 0=no

source_a Source catalog for the record. 1=yes; 0=no
(15 identified sources plus one "other source" category; 15 dummies created with "other" as the reference, hence omitted.)

source_c

source_b

source_d

source_e

source_m

source_o

source_h

source_r

source_s

source_t

source_u

source_p

source_x

source_w

Freq Number of transactions in last year at source catalog

last_update_days_ago How many days ago was last update to customer record

X1st_update_days_ago How many days ago was 1st update to customer record

Web.order Customer placed at least 1 order via web. 1=yes; 0=no

Gender.male Customer is male. 1=yes; 0=no

Address_is_res Address is a residence. 1=yes; 0=no

Purchase Person made purchase in test mailing. 1=yes; 0=no

Spending Amount spent by customer intest mailing ($)

Source

Copyright 2016 statistics.com

Examples

```
head(Tayko)
```

`TinyData` *TinyData*

Description

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. Data

Usage

```
TinyData
```

Format

A data frame with 6 observations and 4 variables:

Obs. unique identifier

Fat relative presence of fat in cheese. (where 0 is the minimum amount possible in the manufacturing process, and 1 the maximum)

Salt relative presence of salt in cheese. (where 0 is the minimum amount possible in the manufacturing process, and 1 the maximum)

Acceptance Consumer taste preference; one of 'like' or 'dislike'

Examples

```
head(TinyData)
```

`ToyotaCorolla` *Toyoto Corolla*

Description

Data

Usage

```
ToyotaCorolla
```

Format

A data frame with 1436 observations and 39 variables:

Id Record_ID; unique identifier

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. one of 'Petrol', 'Diesel', or '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)

Source

Copyright 2016 Nitin Patel, Galit Shmueli and Peter Bruce

Examples

```
head(ToyotaCorolla)
```

`ToysRUsRevenues` *ToysRUsRevenues*

Description

The quarterly revenues of Toys "R" Us between 1992 and 1995

Usage

```
ToysRUsRevenues
```

Format

A data frame with 16 observations and 4 variables:

Index

QuarterYear

Revenue.in.million...

Quarter

Source

Chris Albright

Examples

```
ToysRUsRevenues
```

`UniversalBank` *UniversalBank*

Description

Synthetic dataset courtesy of statistics.com

Usage

```
UniversalBank
```

Format

A data frame with 500 observations and 14 variables:

ID Customer ID

Age Customer's age in completed years

Experience #years of professional experience

Income Annual income of the customer ($000)

ZIP.Code 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?

Source

Copyright Cytel, Inc. 2005

Examples

```
UniversalBank
```

`UniversalBankCase` *UniversalBankCase*

Description

A bank of small to moderate size is thinking of automating its loan approval process to compete with internet-based financial service companies. A consultant has been engaged to develop a predictive model, based on historical human-reviewed loans and their approval rates. The model would be used to approve or reject loans automatically.

Usage

```
UniversalBankCase
```

Format

A data frame with 5000 observations and 17 variables:

ID Customer ID

Age Customer's age in completed years

Experience #years of professional experience

Income Annual income of the customer ($000)

ZIP.Code 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?

Race 0 = White/other (non-hisp), 1 = Black, 2 = Hispanic

Gender 0 = Male, 1 = Female

Neighborhood 1 = Urban, 0 = Other

Source

Synthetic dataset courtesy of statistics.com. Copyright Cytel, Inc. 2005, modifications Copyright Peter Bruce, 2022

Examples

```
UniversalBankCase
```

`Universities` *Universities*

Description

The dataset on American college and university rankings 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).

Usage

```
Universities
```

Format

A data frame with 1302 observations and 20 variables:

College.Name

State

Public..1...Private..2.

X..appli..rec.d

X..appl..accepted

X..new.stud..enrolled

X..new.stud..from.top.10.

X..new.stud..from.top.25.

X..FT.undergrad

X..PT.undergrad

in.state.tuition

out.of.state.tuition

room

board

add..fees

estim..book.costs

estim..personal..

X..fac..w.PHD

stud..fac..ratio

Graduation.rate

Source

Copyright 2016 Galit Shmueli and Peter Bruce. Compiled from US News and World Report rankings on 1302American Colleges and Universities

Examples

```
head(Universities)
```

`Utilities` *Utilities*

Description

Data

Usage

```
Utilities
```

Format

A data frame with 22 observations and 9 variables:

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)

Examples

```
Utilities
```

`Veerhoven` *Veerhoven.*

Description

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

Usage

```
Veerhoven
```

Format

A data frame with 159 observations and 5 variables:

Serial

Code

Nation

Score

X..surveys

Examples

```
Veerhoven
```

`VoterPersuasion` *VoterPersuasion*

Description

Data

Usage

```
VoterPersuasion
```

Format

A data frame with 10000 observations and 79 variables:

VOTER_ID Unique ID for each voter

SET_NO Set number, assigned at random. Can be used to divide development and test sets. Build models using only sets 1 & 2. Validate on the hold-out sample of set_no=3 voters

OPP_SEX % of people with the same first name who have a different gender (Derived)

AGE Age (Voterfile)

HH_ND Number of Democrats in this household (Derived)

HH_NR Number of Republicans in this household (Derived)

HH_NI Number of independents in this household (Derived)

MED_AGE Census Block Group median age (Census)

NH_WHITE % Non-Hispanic Caucasian (Census)

NH_AA % Non-Hispanic African-American (Census)

NH_ASIAN % Non-Hispanic Asian (Census)

NH_MULT % Non-Hispanic multi-race (Census)

HISP % Hispanic (Census)

COMM_LT10 % of workers who commute less than 10 minutes each way (Census)

COMM_609P % of workers who commute 60+ minutes each way (Census)

MED_HH_INC Median household income (Census)

COMM_CAR % of workers who commute by themselves by car (Census)

COMM_CP % of workers who carpool (Census)

COMM_PT % of workers who take public transportation (Census)

COMM_WALK % of workers who walk to work (Census)

KIDS % of families with children under 18 (Census)

M_MAR % of adult men who are married (Census)

F_MAR % of adult females who are married (Census)

ED_4COL % of adult population with at least 4 years of college (Census)

GENDER_F Flag - is female (Voterfile)

GENDER_M Flag - is male (Voterfile)

H_AFDLN3P Flag - household all-female different last names, 3+ members (Derived)

H_F1 Flag - household single female (Derived)

H_M1 Flag - Single male (Derived)

H_MFDLN3P Flag - household male & female, different last names, 3+ members (Derived)

PARTY_D Flag - Democrat (Voterfile)

PARTY_I Flag - Independent (Voterfile)

PARTY_R Flag - Republican (Voterfile)

VPP_08 Flag - vote history - voted presidential primary 2008 (Derived)

VPP_12 Flag - vote history - voted presidential primary 2012 (Derived)

VPR_08 Flag - vote history - voted primary 2008 (Derived)

VPR_10 Flag - vote history - voted primary 2010 (Derived)

VPR_12 Flag - vote history - voted primary 2012 (Derived)

VG_04 Flag - vote history - voted general election 2004 (Derived)

VG_06 Flag - vote history - voted general election 2006 (Derived)

VG_08 Flag - vote history - voted general election 2008 (Derived)

VG_10 Flag - vote history - voted general election 2010 (Derived)

VG_12 Flag - vote history - voted general election 2012 (Derived)

PP_PELIG Voted in % of presidential primaries in which they were eligible (Derived)

PR_PELIG Voted in % of non-presidential primaries in which they were eligible (Derived)

AP_PELIG Voted in % of any kind of primary in which they were eligible (Derived)

G_PELIG Voted in % of general elections in which they were eligible (Derived)

E_PELIG Voted in % of any kind of election in which they were eligible (Derived)

NL5G # of the last 5 elections in which the voter voted (Derived)

NL3PR # of last 3 primaries in which the voter voted (Derived)

NL5AP # of last 5 primaries of any kind in which the voter voted (Derived)

NL2PP # of last 2 presidential primaries in which the voter voted (Derived)

REG_DAYS Days since the voter registered to vote at their current address (Derived)

UPSCALEBUY Upscale Buyer In Home (Commercial Data)

UPSCALEMAL Upscale Male Buyer In Home (Commercial Data)

UPSCALEFEM Upscale Female Buyer In Home (Commercial Data)

BOOKBUYERI Book Buyer In Home (Commercial Data)

FAMILYMAGA Family Magazine In Home (Commercial Data)

FEMALEORIE Female Oriented Magazine In Home (Commercial Data)

RELIGIOUSM Religious Magazine In Home (Commercial Data)

GARDENINGM Gardening Magazine In Home (Commercial Data)

CULINARYIN Culinary Interest Magazine In Home (Commercial Data)

HEALTHFITN Health Fitness Magazine In Home (Commercial Data)

DOITYOURSE Do It Yourselfer Magazine In Home (Commercial Data)

FINANCIALM Financial Magazine In Home (Commercial Data)

RELIGIOUSC Religious Contributor In Home (Commercial Data)

POLITICALC Political Contributer In Home (Commercial Data)

MEDIANEDUC Median Education Years (Commercial Data)

CAND1S Wave 1 candidate ID with strenght of support. SD=Strong Democrat, LD=Lean Democrat, U=Undecided, LR=Lean Republican, SR=Strong Republican (Synthetic)

CAND2S Wave 2 candidate ID with strenght of support (Synthetic)

MESSAGE_A Flag indicating if the voter received message A (Synthetic)

MESSAGE_A_REV Flag indicating if the voter received message B (Synthetic)

I3 Independent 3-way. Y if voter is an independent or minor party member. N if Democrat or Republican

CAND1_UND Undecided in wave 1 IDs

CAND2_UND Undecided in wave 2 IDs

MOVED_AD Moved to be more supportive of the Democratic candidate. Stronger support for the Democrat or weaker support for the Republican in wave 2 than in wave 1.

MOVED_A 0/1 encoding of 'MOVED_AD'

opposite reverse of 'MOVED_AD'

Partition 'V'=validation, 'T'= test set

Source

Copyright 2016 Ken Strasma and statistics.com; Ken Strasma and HaystaqDNA

Examples

```
VoterPersuasion
```

`WalMartStock` *WalMartStock*

Description

The series of Walmart daily closing prices between February 2001 and February 2002. Data

Usage

```
WalMartStock
```

Format

A data frame with 248 observations and 2 variables:

Date

Close

Source

publicly available, for example, at https://finance.yahoo.com. These data are also used in "Data Analysis for Managers" by Albright, Winston & Zappe.

Examples

```
WalMartStock
```

`WestRoxbury` *West Roxbury.*

Description

Data

Usage

```
WestRoxbury
```

Format

A data frame with 5802 observations and 14 variables:

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 (in square feet)

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)

Examples

```
WestRoxbury
```

`Wine` *Wine.*

Description

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.

Data

Usage

```
Wine
```

Format

A data frame with 178 observations and 14 variables:

Type Type of wine; one of 'A', 'B', or 'C'

Alcohol

Malic_Acid

Ash

Ash_Alcalinity Alcalinity of ash

Magnesium

Total_Phenols

Flavanoids

Nonflavanoid_Phenols

Proanthocyanins

Color_Intensity

Hue

OD280_OD315 OD280/OD315 of diluted wines

Proline

Source

This data set can be found in the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/wine)

Examples

```
Wine
```

Index