Fundamentals of Business Intelligence

Univ.-Prof. Dr. Wilfried Grossmann
Univ. Prof. Dr. Stefanie Rinderle-Ma

Exercise Chapter 5: Data Mining for Cross-sectional Data

1. Used Cars Prices

The dataset `Toyota Corolla.csv` shows information about 1435 used Toyota cars. The following variables are available:

Price	Offer price in EURO
Age	in month
KM	Accumulate Km
Fuel Type	Fuel Type (petrol, Diesel, CNG)
HP	Horsepower
MetColor	Metallic Color (1 = yes, 0 = no)
Automatic	1= yes, 0 = no
Doors	Number of Doors
Weight	Weight in kilograms

Tasks:

a) Use descriptive methods and visualization techniques for analysis of the variables in the dataset.
b) Define a regression model which allows the prediction of the price in dependence of the other variables.
c) Split the data in a training set and test set and calculate the prediction error from the test set.

Note: This example is adapted from Leodolter: Data Mining and Business Analytics with R

2. Wholesale Customers 1

The dataset `WholesaleData.csv` shows for 440 customers of a wholesale distributor the following information:

FRESH	annual spending (monetary units) on fresh products (Continuous)
MILK	annual spending (monetary units) on milk products (Continuous)
GROCERY	annual spending (monetary units) on grocery products (Continuous)
FROZEN	annual spending (monetary units) on frozen products (Continuous)
DETERGENTS_PAPER	annual spending (monetary units) on detergents and paper products (Continuous)
DELICATESSEN (Continuous)	annual spending (monetary units) on and delicatessen products
CHANNEL	Horeca (Hotel/Restaurant/Café) or Retail channel (Nominal)
REGION	Lisbon, Oporto or Other (Nominal)

Tasks

a) Use descriptive statistics for business and data understanding and summarize the findings.
b) Find classification rules which allow the discrimination of the distribution channels and.
c) Find classification rules for discrimination of the regions.

Note: This dataset was taken from the UCI Machine Learning Repository:
https://archive.ics.uci.edu/ml/datasets/Wholesale+customers

3. Wholesale Customers 2

a) Apply for the dataset of exercise 1 different cluster analysis methods and select an appropriate solution for clustering.
b) Visualize the cluster solutions with different techniques and interpret the solutions.
c) Compare the cluster solution with the classification rules of exercise 1 b) and 1c). Are the results comparable?

4. Credit Card Default

The dataset CreditDefault.xlsx informs about the defaults of 3000 credit card clients. The following variables are given:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2: Gender (1 = male; 2 = female).
X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4: Marital status (1 = married; 2 = single; 3 = others).
X5: Age (year).
X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -2,-1, 0 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

Y Credit default (1 = default, 0 = no default=

Tasks:

a) Use descriptive statistics for business and data understanding and summarize the findings.
b) Split the dataset in a training and a test set (70%, 30%) and learn a classification rule from the training set using logistic regression. Find an appropriate threshold for the classification. Evaluate the solution for the test set.
c) Apply at least two other classification methods and compare the solutions with logistic regression from a numerical point of view (accuracy, sensitivity, specificity) and from a practitioners point of view (interpretation and understanding of the results).

Note: This example is adapted from the UCI Machine Learning Repository:
https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients