The Insurance Company Benchmark (COIL 2000)

Data Type

multivariate

Abstract

This data set used in the CoIL 2000 Challenge contains information on customers of an insurance company. The data consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes. The data was collected to answer the following question: Can you predict who would be interested in buying a caravan insurance policy and give an explanation why?

Sources

Original Owner and Donor

Peter van der Putten
Sentient Machine Research
Baarsjesweg 224
1058 AA Amsterdam
The Netherlands
+31 20 6186927
pvdputten@hotmail.com, putten@liacs.nl

TIC Benchmark Homepage

Date Donated: March 7, 2000

Data Characteristics

Information about customers consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes. The data was supplied by the Dutch data mining company Sentient Machine Research and is based on a real world business problem. The training set contains over 5000 descriptions of customers, including the information of whether or not they have a caravan insurance policy. A test set contains 4000 customers of whom only the organisers know if they have a caravan insurance policy.

The data dictionary describes the variables used and their values.

Note: All the variables starting with M are zipcode variables. They give information on the distribution of that variable, e.g. Rented house, in the zipcode area of the customer.

Data Format

One instance per line with tab delimited fields.

TICDATA2000.txt: Dataset to train and validate prediction models and build a description (5822 customer records). Each record consists of 86 attributes, containing sociodemographic data (attribute 1-43) and product ownership (attributes 44-86).The sociodemographic data is derived from zip codes. All customers living in areas with the same zip code have the same sociodemographic attributes. Attribute 86, "CARAVAN:Number of mobile home policies", is the target variable.

TICEVAL2000.txt: Dataset for predictions (4000 customer records). It has the same format as TICDATA2000.txt, only the target is missing. Participants are supposed to return the list of predicted targets only. All datasets are in tab delimited format. The meaning of the attributes and attribute values is given below.

TICTGTS2000.txt Targets for the evaluation set.

Past Usage

P. van der Putten and M. van Someren (eds). CoIL Challenge 2000: The Insurance Company Case. Published by Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report 2000-09. June 22, 2000.

In this report you will find 29 short papers and extended abstracts on this problem.

Acknowledgements

Data is (c) Sentient Machine Research 2000

This dataset is owned and supplied by the Dutch datamining company Sentient Machine Research, and is based on real world business data. You are allowed to use this dataset and accompanying information for non commercial research and education purposes only. It is explicitly not allowed to use this dataset for commercial education or demonstration purposes.

Please cite/acknowledge:

P. van der Putten and M. van Someren (eds) . CoIL Challenge 2000: The Insurance Company Case. Published by Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report 2000-09. June 22, 2000.

References and Further Information

There is a special website for this benchmark at http://www.liacs.nl/~putten/library/cc2000/. On the website you can find an online report featuring 29 papers written by participants in the CoIL Challenge 2000 and further background information. In future more papers will be added to the website. If you have any submissions, please send them to putten@liacs.nl.

The UCI KDD Archive
Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425

Last modified: October 12, 2000.