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?

Acknowledgements

DISCLAIMER

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. For any other use, please contact Peter van der Putten, info@smr.nl.

This dataset has been used in the CoIL Challenge 2000 datamining competition. For papers describing results on this dataset, see the TIC 2000 homepage: <http://www.wi.leidenuniv.nl/~putten/library/cc2000/>

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.

The Data

Originally, this dataset was broken into two parts: the training set and the evaluation set. As this was a competition, the responses to the evaluation set were not given as part of the original release; they were, however, released after the end of the competition in a separate file. This dataset contains all three of these files, combined into one.

The field ORIGIN in the caravan-insurance-challenge.csv file has the values *train* and *test*, corresponding to the training and evaluation sets, respectively. To simulate the original challenge, you can ignore the *test* rows, and test your model's prediction on those observations once you've trained only on the training set.

Each observation corresponds to a postal code. Variables beginning with M refer to demographic statistics of the postal code, while variables beginning with P and A (as well as CARAVAN, the target variable) refer to product ownership and insurance statistics in the postal code.

The data file contains the following fields:

* ORIGIN: *train* or *test*, as described above
* MOSTYPE: Customer Subtype; see L0
* MAANTHUI: Number of houses 1 - 10
* MGEMOMV: Avg size household 1 - 6
* MGEMLEEF: Avg age; see L1
* MOSHOOFD: Customer main type; see L2

\*\* Percentages in each group, per postal code (see L3)\*\*:

* MGODRK: Roman catholic
* MGODPR: Protestant ...
* MGODOV: Other religion
* MGODGE: No religion
* MRELGE: Married
* MRELSA: Living together
* MRELOV: Other relation
* MFALLEEN: Singles
* MFGEKIND: Household without children
* MFWEKIND: Household with children
* MOPLHOOG: High level education
* MOPLMIDD: Medium level education
* MOPLLAAG: Lower level education
* MBERHOOG: High status
* MBERZELF: Entrepreneur
* MBERBOER: Farmer
* MBERMIDD: Middle management
* MBERARBG: Skilled labourers
* MBERARBO: Unskilled labourers
* MSKA: Social class A
* MSKB1: Social class B1
* MSKB2: Social class B2
* MSKC: Social class C
* MSKD: Social class D
* MHHUUR: Rented house
* MHKOOP: Home owners
* MAUT1: 1 car
* MAUT2: 2 cars
* MAUT0: No car
* MZFONDS: National Health Service
* MZPART: Private health insurance
* MINKM30: Income < 30.000
* MINK3045: Income 30-45.000
* MINK4575: Income 45-75.000
* MINK7512: Income 75-122.000
* MINK123M: Income >123.000
* MINKGEM: Average income
* MKOOPKLA: Purchasing power class

\*\* Total number of variable in postal code (see L4)\*\*:

* PWAPART: Contribution private third party insurance
* PWABEDR: Contribution third party insurance (firms) ...
* PWALAND: Contribution third party insurane (agriculture)
* PPERSAUT: Contribution car policies
* PBESAUT: Contribution delivery van policies
* PMOTSCO: Contribution motorcycle/scooter policies
* PVRAAUT: Contribution lorry policies
* PAANHANG: Contribution trailer policies
* PTRACTOR: Contribution tractor policies
* PWERKT: Contribution agricultural machines policies
* PBROM: Contribution moped policies
* PLEVEN: Contribution life insurances
* PPERSONG: Contribution private accident insurance policies
* PGEZONG: Contribution family accidents insurance policies
* PWAOREG: Contribution disability insurance policies
* PBRAND: Contribution fire policies
* PZEILPL: Contribution surfboard policies
* PPLEZIER: Contribution boat policies
* PFIETS: Contribution bicycle policies
* PINBOED: Contribution property insurance policies
* PBYSTAND: Contribution social security insurance policies
* AWAPART: Number of private third party insurance 1 - 12
* AWABEDR: Number of third party insurance (firms) ...
* AWALAND: Number of third party insurance (agriculture)
* APERSAUT: Number of car policies
* ABESAUT: Number of delivery van policies
* AMOTSCO: Number of motorcycle/scooter policies
* AVRAAUT: Number of lorry policies
* AAANHANG: Number of trailer policies
* ATRACTOR: Number of tractor policies
* AWERKT: Number of agricultural machines policies
* ABROM: Number of moped policies
* ALEVEN: Number of life insurances
* APERSONG: Number of private accident insurance policies
* AGEZONG: Number of family accidents insurance policies
* AWAOREG: Number of disability insurance policies
* ABRAND: Number of fire policies
* AZEILPL: Number of surfboard policies
* APLEZIER: Number of boat policies
* AFIETS: Number of bicycle policies
* AINBOED: Number of property insurance policies
* ABYSTAND: Number of social security insurance policies
* CARAVAN: Number of mobile home policies 0 - 1

Keys (L1 - L4)

L0: Customer subtype

* *1*: High Income, expensive child
* *2*: Very Important Provincials
* *3*: High status seniors
* *4*: Affluent senior apartments
* *5*: Mixed seniors
* *6*: Career and childcare
* *7*: Dinki's (double income no kids)
* *8*: Middle class families
* *9*: Modern, complete families
* *10*: Stable family
* *11*: Family starters
* *12*: Affluent young families
* *13*: Young all american family
* *14*: Junior cosmopolitan
* *15*: Senior cosmopolitans
* *16*: Students in apartments
* *17*: Fresh masters in the city
* *18*: Single youth
* *19*: Suburban youth
* *20*: Etnically diverse
* *21*: Young urban have-nots
* *22*: Mixed apartment dwellers
* *23*: Young and rising
* *24*: Young, low educated
* *25*: Young seniors in the city
* *26*: Own home elderly
* *27*: Seniors in apartments
* *28*: Residential elderly
* *29*: Porchless seniors: no front yard
* *30*: Religious elderly singles
* *31*: Low income catholics
* *32*: Mixed seniors
* *33*: Lower class large families
* *34*: Large family, employed child
* *35*: Village families
* *36*: Couples with teens 'Married with children'
* *37*: Mixed small town dwellers
* *38*: Traditional families
* *39*: Large religous families
* *40*: Large family farms
* *41*: Mixed rurals

L1: average age keys:

*1*: 20-30 years *2*: 30-40 years *3*: 40-50 years *4*: 50-60 years *5*: 60-70 years *6*: 70-80 years

L2: customer main type keys:

* *1*: Successful hedonists
* *2*: Driven Growers
* *3*: Average Family
* *4*: Career Loners
* *5*: Living well
* *6*: Cruising Seniors
* *7*: Retired and Religeous
* *8*: Family with grown ups
* *9*: Conservative families
* *10*: Farmers

L3: percentage keys:

* *0*: 0%
* *1*: 1 - 10%
* *2*: 11 - 23%
* *3*: 24 - 36%
* *4*: 37 - 49%
* *5*: 50 - 62%
* *6*: 63 - 75%
* *7*: 76 - 88%
* *8*: 89 - 99%
* *9*: 100%

L4: total number keys:

* *0*: 0
* *1*: 1 - 49
* *2*: 50 - 99
* *3*: 100 - 199
* *4*: 200 - 499
* *5*: 500 - 999
* *6*: 1000 - 4999
* *7*: 5000 - 9999
* *8*: 10,000 - 19,999
* *9*: >= 20,000