LAB 3: Caravan-Insurance Problem

1. Intro: Caravan-insurance problem

Direct mailings to a company's potential customers ('junk mail' to many) can be a very effective way for them to market a product or a service. However, as we all know, much of this junk mail is really of no interest to the people that receive it. Most of it ends up thrown away, not only wasting the money that the company spent on it, but also filling up landfill waste sites or needing to be recycled.

If the company had a better understanding of who their potential customers were, they would know more accurately who to send it to, so some of this waste and expense could be reduced. We will study this problem in a context of a Dutch insurance company that among others sells insurances for customers that own their own caravans. We have two questions posed by the company:

1) *Can you describe a potential customer interested in buying a caravan insurance?*

2) *Can you predict who would be interested in buying a caravan insurance policy?*

2. Assignments

The company's questions result in two assignments:

- **Assignment 1:** Describe the actual or potential customers and possibly explain why these customers buy a caravan policy.
- **Assignment 2:** Select customers from a test file to send information to. The file with those customers will be provided on *the day of lab's deadline*.

Build classification models for these assignments. Note that these assignments may be conflicting in the sense that some models are better suited for correct classification, while others give clearer models. Therefore, you may need to apply different models to these two assignments.

2.1 Data

The data about customers is represented by 86 variables and includes product usage data and socio-demographic data derived from zip area codes. The training data contains over 5000 descriptions of customers, including the information of whether or not they have a caravan insurance policy. The test data contains 4000 customers of whom only company's supervisors know if they have a caravan insurance policy.

The data sets are given in two formats: csv format and arff format (Weka). In the csv format the input variables are considered as numeric. In the arff format the input variables are considered as nominal. In **Appendix A** more details are provided.

2.2 Assignment 1

The purpose of assignment 1 is to give a clear insight to why customers have a caravan insurance policy and how these customers are different from other customers. Descriptions can be based on regression equations, decision trees, linguistic descriptions, graphical representations or any other form. The descriptions and accompanying interpretation must be comprehensible, useful and actionable for a marketing professional with no prior knowledge of data mining.

Compare some different techniques and/or settings of parameters to see how well they perform on this problem. For this comparison you may assume some basic knowledge about data mining with the reader. In **Appendix B** details are provided how to use scikit-learn For feature selection.

2.3 Assignment 2

The purpose of assignment 2 is to find a set of 800 customers from the test set that contains the most caravan policy owners. Use your most accurate model to select the 800 most likely policy owners.

Report. Submit a report with your answers to the assignments (2-6 pages).

Appendix A Detailed data description

A.1 Relevant files

The relevant files are given in plain text format.

A.2 Data dictionary

Attribute number, Name and Description Domain.
1 MOSTYPE Customer Subtype see L0
2 MAANTHUI Number of houses 1...10
3 MGEMOMV Avg size household 1...6
4 MGEMLEEF Avg age see L1
5 MOSHOOFD Customer main type see L2
6 MGODRK Roman catholic see L3
7 MGODPR Protestant ...
8 MGODOV Other religion
9 MGODGE No religion
10 MRELGE Married
11 MRELSA Living together
12 MRELOV Other relation
13 MFALLEEN Singles
14 MFGEKIND Household without children
15 MFWEKIND Household with children
16 MOPLHOOG High level education
17 MOPLMIDD Medium level education
18 MOPLLAAG Lower level education
19 MBERHOOG High status
20 MBERZELF Entrepreneur
21 MBERBOER Farmer
22 MBERMIDD Middle management
23 MBERARBG Skilled labourers
24 MBERARBO Unskilled labourers
25 MSKA Social class A
26 MSKB1 Social class B1
27 MSKB2 Social class B2
28 MSKC Social class C
29 MSKD Social class D
30 MHHUUR Rented house
31 MHKOOP Home owners
32 MAUT1 1 car
33 MAUT2 2 cars
34 MAUT0 No car
35 MZFONDS National Health Service
36 MZPART Private health insurance
37 MINKM30 Income < 30.000
38 MINK3045 Income 30-45.000
39 MINK4575 Income 45-75.000
40 MINK7512 Income 75-122.000
41 MINK123M Income >123.000
42 MINKGEM Average income
43 MKOOPKLA Purchasing power class
44 PWAPART Contribution private third party insurance see L4
45 PWABEDR Contribution third party insurance (firms) ...
46 PWALAND Contribution third party insurance (agriculture)
47 PPERSAUT Contribution car policies
48 PBESAUT Contribution delivery van policies
49 PMOTSCO Contribution motorcycle/scooter policies
50 PVRAAUT Contribution lorry policies
51 PAANHANG Contribution trailer policies
52 PTRACTOR Contribution tractor policies
53 PWERKT Contribution agricultural machines policies
54 PBROM Contribution moped policies

55 PLEVEN Contribution life insurances
56 PPERSONG Contribution private accident insurance policies
57 PGEZONG Contribution family accidents insurance policies
58 PWAOREG Contribution disability insurance policies
59 PBRAND Contribution fire policies
60 PZEILPL Contribution surfboard policies
61 PPLEZIER Contribution boat policies
62 PFIETS Contribution bicycle policies
63 PINBOED Contribution property insurance policies
64 PBYSTAND Contribution social security insurance policies
65 AWAPART Number of private third party insurance 1 - 12
66 AWABEDR Number of third party insurance (firms) ...
67 AWALAND Number of third party insurance (agriculture)
68 APERSAUT Number of car policies
69 ABESAUT Number of delivery van policies
70 AMOTSCO Number of motorcycle/scooter policies
71 AVRAAUT Number of lorry policies
72 AAANHANG Number of trailer policies
73 ATRACTOR Number of tractor policies
74 AWERKT Number of agricultural machines policies
75 ABROM Number of moped policies
76 ALEVEN Number of life insurances
77 APERSONG Number of private accident insurance policies
78 AGEZONG Number of family accidents insurance policies
79 AWAOREG Number of disability insurance policies
80 ABRAND Number of re policies
81 AZEILPL Number of surfboard policies
82 APLEZIER Number of boat policies
83 AFIETS Number of bicycle policies
84 AINBOED Number of property insurance policies
85 ABYSTAND Number of social security insurance policies
86 CARAVAN Number of mobile home policies 0 - 1

A.3 Data domains

L0:

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

L1:

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:

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:

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:

0 f 0

1 f 1 – 49

2 f 50 – 99

3 f 100 – 199

4 f 200 – 499

5 f 500 – 999

6 f 1000 – 4999

7 f 5000 – 9999

8 f 10.000 - 19.999

9 f 20.000 - ?

Appendix B: Python modules

For the feature selection part of the assignments you can use scikit-learn implementations provided on:

https://scikit-learn.org/stable/modules/feature_selection.html

Note that some of the feature selection methods are not compatible with the classification models you might use.

In addition, that feature selection is a part of training classification models! To plug feature selection methods in the training process you use:

```
from sklearn.pipeline import make_pipeline
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from sklearn.feature_selection import RFE

from sklearn.linear_model import LogisticRegression

classifier = LogisticRegression()
pipeClassifier = make_pipeline(SelectKBest(chi2, k=4), classifier)
#k is the number of variables selected

pipeClassifier.fit(X, Y)
```