

## Predictive Models in Business Analytics – Project

Due: Wednesday December 8, 2021 at 11:00am

(Report can be handed in as late as Monday December 13 at 12:00pm (noon))

You are requested to make the following project questions. You are allowed to work in teams of at most 2 students. Please, take the following instructions into consideration when handing in this assignment:

1. Regularly use the **“save” button** to prevent losing your work when the computer or RapidMiner crashes.
2. Use the **pre-structured report** to hand in your results and your answers to the questions. You can find this form on D2L. Please stay within the boxes (they will grow while you type).
3. When figures or tables are created in RapidMiner, **include them in your written report**. You can do so by screen clipping and then pasting in Word. Make sure everything is easily readable. If needed, use multiple screenshots for legibility. Do NOT include screenshots of your RapidMiner process.
4. Include page numbers. Please use grammatically correct English.
5. Before uploading your report on D2L, covert the Word document to **PDF format**. Each group is to develop a single report write-up.
6. Create a **ZIP** file containing ONLY the PDF file of the report with the accompanying RapidMiner files (e.g., “Task1.rmp”, “Task2.rmp” and “project-answerform.pdf”). The PDF file as well as the RapidMiner files will be used for marking. Put your student numbers in the name of the ZIP file. So, for example, the name of the ZIP file becomes “123456-234567.zip”. Do NOT use any other compression software (such as RAR or ARJ).
7. Submit the ZIP file electronically using **D2L Dropbox** (under any of the group member’s names).
8. If the project is uploaded on D2L after the deadline, it is up to the instructor to lower the mark received for the project (see the Course Outline for more details).

**Good luck!!!**

# INTRODUCTION

Sending letters to potential customers may be a very effective way to market a product or a service. However, as we all know, most of this “junk mail” is not really interesting to most of the recipients and it usually ends up, unopened, in a trash bin. In this way companies are losing money, people get irritated, the environment suffers, etc. If companies could better predict the interests and needs of their potential clients, life would be much easier.

In this assignment you are supposed to help an insurance company to identify potential buyers of one of their products: a caravan insurance policy. More precisely, you will have to analyze some historical data (the file “train.txt”), in order to make some good selections of records from the test set, “test.txt”.

The file “train.txt” contains 5,822 records. Each record consists of 86 attributes. The first 85 of them describe customers with sociodemographic data and product ownership data. The sociodemographic data (attribute 1-43) is derived from zip codes, and the variables give information on the distribution of that variable in the zip code area of the customer. All customers living in areas with the same zip code have the same sociodemographic attributes. The product ownership data (attributes 44-86) is split in information regarding the contribution paid by the customer towards certain insurance policies (attributes 44-64) and the number of insurance policies the customer has for each of these insurances (attributes 65-85). The last attribute, the target, takes values 0 or 1, depending on whether the customer bought the caravan policy or not. More information about the meaning of all variables can be found at the end of this document.

The file “test.txt”, with 4,000 records, also contains 86 variables; however, the target variable is dummy: it is set to 0, so *you* don't really know which clients from the test set bought the policy. However, *we* do know the true values of this target variable - we will keep them hidden and use them to evaluate your submissions.

The project consists of two tasks. In Task 1, you will have to make a selection of 1,000 records from the test set to catch as many buyers of the policy as possible. In Task 2, you will have to make another selection of records, in such a way that the corresponding profit is maximized. To calculate the profit, we will assume that sending a single letter cost \$1 and that the company makes \$15 profit on every buyer of the policy.

## Task 1: Selection

In this task, you need to find a set of 1,000 customers in the test set that contains as many potential buyers of a caravan policy as possible. In other words, use your most accurate model that you developed on the training set to select 1,000 most likely policy buyers. Instead of submitting 1,000 selected records, you should submit only a list of 1,000 integers that represent positions of the selected records in the test data. Thus, if your selection contains the first, third, ..., and the last record in the test data set, you should submit a list of integers that looks like this:

1  
3  
...  
4000

In other words, your submission should be a file with 1,000 indices - integers between 1 and 4,000. Put each index on a separate line (without comma's!). Do not put any other text in the file, so your file should have exactly 1,000 lines.

## Task 2: Profit

Analyze the performance of your best model (which can/should be a different model than for task 1, as the objective is different) and decide which records from the test set should be selected in order to maximize the profit. As said before, each selected record costs \$1, while each selected record that is a policy buyer provides a profit of \$15. This time, the number of selected records has to be determined by you (if you submit an empty list your profit will be \$0; if you submit all 4,000 cases your profit will be about  $\$15 \cdot 240 - \$4,000 = \$-400$  - we know that there are about 240 policy buyers in the test set).

Again, only submit the indices of the selected records (as in the previous task).

## Expectations and Deliverables

The two tasks have the form of a competition: you should submit your selections by e-mail to [REDACTED]. The maximal number of submissions is three (for each task) and the best of these three submissions counts as final submission.

There are three deliverables for this project:

- The RapidMiner files associated with the best model for Task 1 and Task 2
- A short PowerPoint presentation (around 4-5 minutes) that highlights your approach that you will need to present on Wednesday December 8 during the lecture time.
- A report that provides a detailed description how you have addressed the two individual tasks, what approach/process you have taken, which learning algorithms you have selected to analyze (including the parameter values), how the learning algorithms performed on your training and validation dataset, and which learning algorithm (including parameter values) you have selected for your final model.

## Description of the data

NR	VARIABLE	DESCRIPTION
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 ... (see L3)
8	MGODOV	Other religion (see L3)
9	MGODGE	No religion (see L3)
10	MRELGE	Married (see L3)
11	MRELSA	Living together (see L3)
12	MRELOV	Other relation (see L3)
13	MFALLEEN	Singles (see L3)
14	MFGEKIND	Household without children (see L3)
15	MFWEKIND	Household with children (see L3)
16	MOPLHOOG	High level education (see L3)
17	MOPLMIDD	Medium level education (see L3)
18	MOPLLAAG	Lower level education (see L3)
19	MBERHOOG	High status (see L3)
20	MBERZELF	Entrepreneur (see L3)
21	MBERBOER	Farmer (see L3)
22	MBERMIDD	Middle management (see L3)
23	MBERARBG	Skilled labourers (see L3)
24	MBERARBO	Unskilled labourers (see L3)
25	MSKA	Social class A (see L3)
26	MSKB1	Social class B1 (see L3)
27	MSKB2	Social class B2 (see L3)
28	MSKC	Social class C (see L3)
29	MSKD	Social class D (see L3)
30	MHHUUR	Rented house (see L3)
31	MHKOOP	Home owners (see L3)
32	MAUT1	1 car (see L3)
33	MAUT2	2 cars (see L3)
34	MAUT0	No car (see L3)
35	MZFONDS	National Health Service (see L3)
36	MZPART	Private health insurance (see L3)
37	MINKM30	Income < 30.000 (see L3)
38	MINK3045	Income 30-45.000 (see L3)
39	MINK4575	Income 45-75.000 (see L3)
40	MINK7512	Income 75-122.000 (see L3)
41	MINK123M	Income >123.000 (see L3)
42	MINKGEM	Average income (see L3)
43	MKOOKLA	Purchasing power class (see L3)
44	PWAPART	Contribution private third party insurance (see L4)
45	PWABEDR	Contribution third party insurance (firms) ... (see L4)
46	PWALAND	Contribution third party insurance (agriculture) (see L4)
47	PPERSAUT	Contribution car policies (see L4)
48	PBESAUT	Contribution delivery van policies (see L4)
49	PMOTSCO	Contribution motorcycle/scooter policies (see L4)
50	PVRAAUT	Contribution lorry policies (see L4)
51	PAANHANG	Contribution trailer policies (see L4)

52	PTRACTOR	Contribution tractor policies (see L4)
53	PWERKT	Contribution agricultural machines policies (see L4)
54	PBROM	Contribution moped policies (see L4)
55	PLEVEN	Contribution life insurances (see L4)
56	PPERSONG	Contribution private accident insurance policies (see L4)
57	PGEZONG	Contribution family accidents insurance policies (see L4)
58	PWAOREG	Contribution disability insurance policies (see L4)
59	PBRAND	Contribution fire policies (see L4)
60	PZEILPL	Contribution surfboard policies (see L4)
61	PPLEZIER	Contribution boat policies (see L4)
62	PFIETS	Contribution bicycle policies (see L4)
63	PINBOED	Contribution property insurance policies (see L4)
64	PBYSTAND	Contribution social security insurance policies (see L4)
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	ATTRACTOR	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 fire 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

L0:

VALUE	LABEL
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	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	Ethnically 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 religious families
40	Large family farms
41	Mixed rurals

L1:

VALUE	LABEL
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:

VALUE	LABEL
1	Successful hedonists
2	Driven growers
3	Average family
4	Career loners
5	Living well
6	Cruising Seniors
7	Retired and religious
8	Family with grown ups
9	Conservative families
10	Farmers

L3:

VALUE	LABEL
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:

VALUE	LABEL
0	\$0
1	\$1 - 49
2	\$50 - 99
3	\$100 - 199
4	\$200 - 499
5	\$500 - 999
6	\$1,000 - 4,999
7	\$5,000 - 9,999
8	\$10,000 - 19,999
9	>= \$20,000