CoIL Challenge 2000 Tasks and Results: Predicting and Explaining Caravan Policy Ownership

Peter van der Putten [1,2], Michel de Ruiter [1], Maarten van Someren [3]
[1] Sentient Machine Research, Amsterdam
{pvdputten,mdruiter}@smr.nl
[2] Leiden Institute of Advanced Computer Science
Leiden University
[3] Sociaal Wetenschappelijke Informatica
Universiteit van Amsterdam
maarten@swi.psy.uva.nl

ABSTRACT: We present the problem tasks of the CoIL Challenge as they were explained to the participants. Furthermore, a general overview is given of the Challenge results.

1 INTRODUCTION

Direct mailings to a company's potential customers - "junk mail" to many - can be a very effective way for to market a product or service. However, as we all know, much of this junk mail is really of no interest to the majority of 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. Therefore, following a successful CoIL competition in 1999, CoIL runs a new competition challenge in 2000:

Can you predict who would be interested in buying a caravan insurance policy and give an explanation why?

We encourage any type of solutions to these problems, particularly those involving any of the CoIL technologies (Fuzzy Logic, Evolutionary Computing, Machine Learning, and Neural Networks) or any combinations of these. We are also interested in other solutions using other technologies, since CoIL is interested in being able to demonstrate how CoIL technologies relate to other techniques. The winners will be invited to present a short paper on their approach at the CoIL'2000 Symposium on Computational Intelligence and Learning (19-23 June 2000), in Chios, Greece.

2 THE TASKS

The competition consists of two tasks:

1. Predict which customers are potentially interested in a caravan insurance policy.
2. Describe the actual or potential customers; and possibly explain why these customers buy a caravan policy.

For both tasks only one winner will be chosen. Information about customers consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes. The data was 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. See the appendix for more information.

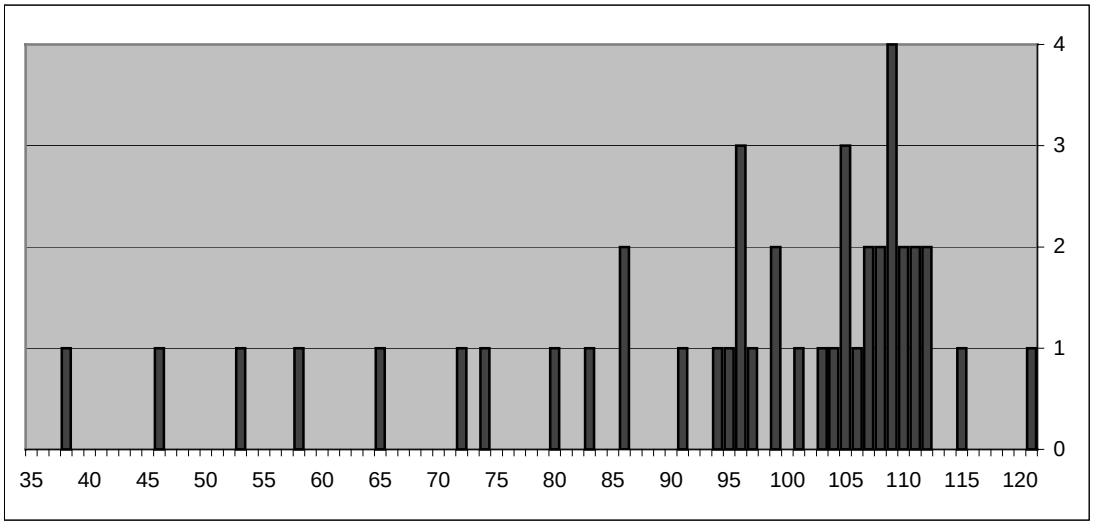

Figure 1: Frequency distribution of prediction scores.
The x-axis displays number of real policy owners in the selection that was sent in.

2.1 PREDICTION TASK

In the prediction task, the underlying problem is to find the subset of customers with a probability of having a caravan insurance policy above some boundary probability. The known policyholders can then be removed and the rest receives a mailing. The boundary depends on the costs and benefits such as of the costs of mailing and benefit of selling insurance policies. To approximate and simplify this problem, we want the participants to find the set of 800 customers in the test set of 4000 customers that contains the most caravan policy owners. For each solution submitted, the number of actual policyholders will be counted and this gives the score of a solution.

2.2 DESCRIPTION TASK

The purpose of the description task 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, neural network weights, linguistic descriptions, evolutionary programs, 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 computational learning technology. Since the value of a description is inherently subjective, submitted descriptions will be evaluated by the jury and an expert in insurance marketing.

3 RESULTS

The response to the CoIL Challenge has been very high, when compared to previous CoIL and ERUDIT Challenges and the U.S. counterpart data mining competition, the KDDCUP (24 participants in 1999). Overall 147 participants registered, of which 43 have sent in a solution (29%). We tried to stimulate cooperation and consequently improve the response rate by introducing a CoIL Challenge mailing list, which was used extensively during and after the Challenge period (March 17 - May 8 2000; over 50 messages). The majority of the participants used and some even combined approaches from more than one area in computational intelligence and statistics. Submissions came from 18 different countries[1] from Europe (58%), North and South America (30%), Asia (7%) and Australia (5%), both from industry (31%) and academics (59%; remainder unknown).

[1] All stats are measured by the person sending in solution. Actually, there was a significant number of joint submissions combining people from different countries, academics and industry professionals.

3.1 PREDICTION TASK

The frequency distribution of scores for the prediction task are displayed in figure 1. The maximum number of policy owners that could be found was 238, the winning model selected 121 policy owners. Random selection results in 42 policy owners. Our standard benchmark tests result in 94 (k-nearest neighbor), 102 (naïve bayes), 105 (neural networks) and 118 (linear!) policy owners.
The scores for academic participants versus industry participants were similar; 'Europe' scored on an average 98 policy owners versus 89 policy owners for American submissions; however these findings are hardly significant given the standard deviations (19 vs 21).
A wide variety of methodological approaches were used including using derived attributes, recoding attributes, feature selection, -construction, and -reduction, boosting, bootstrapping and cost-sensitive classification. Algorithms used included standard statistics, neural networks, evolutionary algorithms, genetic programming, fuzzy classifiers, decision and -regression trees, support vector machines, ILP and others. Relating this wide variety of modelling approach to model performance will be interesting, but it requires more in depth study and is therefore out of the scope of this general paper.
The winning solution was provided by Charles Elkan from the University of California (121 policy owners). He used a naïve bayes approach. Apart from 'Purchasing Power Class', all sociodemographic variables derived from zip codes were discarded, because they did not add predictive power to the model. Furthermore, 2 derived variables were introduced that summarized the usage of car and fire policy products. Runners up were Petri Kontkanen from the University of Helsinki, Finland (115 policy owners), Andrew Greenyer from The Database Group, United Kingdom and Arnold Koudijs, Cap Gemini, The Netherlands (both 112). The extended abstracts or short papers describing their approach are all in this report.

3.2 DESCRIPTION TASK

To repeat, the goal in the description task was to explain why people own a caravan policy, given the data, modelling methods and subjective, domain-based interpretation. The descriptions and accompanying interpretation had to be comprehensible, useful and actionable for a marketing professional with no prior knowledge of computational learning technology. Submitted descriptions were evaluated by a marketing expert, Mr Stephan van Heusden from MSP Associates in Amsterdam (www.mspa.nl/uk/uk_index_htm). Descriptive solutions were *not* checked for statistical validity.
Mr van Heusden remarked: "Allmost all entries lack a good description in words: participants seem to forget that most marketeers find it difficult to read statistics (and understand them!)". Mr. van Heusden also stressed the importance of actionability. "A good entry combines a description of the results with a 'tool' to use the results."
There were no significant differences in submissions from the Americas versus Europe. As could be expected (but should be regretted), participants from industry had a better score than academic participants (4.3 versus 3.5 out of a maximum of 6), although these differences were not significant given the standard deviations (1.6 resp. 2.1).
Similar to the prediction solutions, a wide variety of approaches was chosen, although there was a tendency to rule based solutions.
About the winner Mr van Heusen remarked a.o.: "This participant clearly explained which steps preceded his conclusions. He may have discovered all conclusions the others also found, but additionally he really tries to interpret them." Mr. van Heusden also appreciated the association rule results, the discussion of the complex nonlinear relation between purchasing power and policy ownership and the explanations why people were *not* interested in a caravan policy.
Techniques used by the winner for description where Evolutionairy Local Search Algorithms (ELSA), chi square tests and association rules. The winning solution was provided by YongSeog Kim and W. Nick Street, from the Management Science Dept., University of Iowa. Runner up was Tomislav Smuc from the Rudjer Boskovic Institute, Zagreb, Croatia. The solutions of Chris Leckie, Marc Krogel , Phil Brierley and Uzay Kaymak were also especially appreciated.

4 DISCUSSION

The challenge has shown that the computational intelligence community has many tools for addressing data mining problems such as the target selection problem, where the interesting patterns are masked by huge amounts of irrelevant information. The challenge results deserve a more detailed analysis, but in the scope of this report we will limit ourselves to voicing some general comments which were also supported by the marketing expert and challenge participants (special thanks to Uzay Kaymak , Janos Abonyi and Hans Roubos).

Firstly, the spread in the prediction scores is rather large. Apparently, you must have the method and know how to apply it: expertise with the method is required. This stresses the importance for more research in the direction of automating the application of intelligent techniques for datamining. To turn datamining into a tool for end users, this expertise must be made explicit, formalized and automated where possible. This should include steps that go beyond the core algorithm, such as data preparation (e.g. feature selection) and evaluation (application specific evaluation measures, boosting, combining models).
Secondly, it is disappointing to conclude most entries for the description task lack a good description in words. If the valuable but complex patterns which are detected by advanced intelligent techniques are not explained properly, end users like marketers will still prefer the simple but crude and limited solutions.
Thirdly, an approach which was suggested to be the most prudent in the after challenge discussions was to "Try the simplest first and be self-confident." On real world prediction problems like the one in the challenge, one should try a wide variety of approaches, starting with the most simple ones, because they seem to work best. This indicates that simple computational learning algorithms can play an important role when used as alternatives to standard statistical techniques and thus improving the choice range of algorithms. One the other hand, simple statistics should always be included.
To conclude the discussion we would like to remark that it would be very interesting to set up competitions like these as a meta learning experiment from the start, in addition to experiments where all modelling is performed in a standardized, controlled environment.

5 CONCLUSION

The CoIL Challenge was a success for multiple reasons. In a short time period, a large number of people have joined forces working on the same real world problem, exchanging and discussing ideas and trying out competitive approaches. This has lead to a wide variety of solutions, a selection of which is collected in this report. The top solutions indicate a substantial return on investment if these models were used by the insurance company. We hope that this report forms a starting point for further analysis of the challenge results and a source of inspiration for future competition organizers.

6 APPENDIX: DETAILED DATA DESCRIPTION

THE INSURANCE COMPANY (TIC) - COIL CHALLENGE 2000

DISCLAIMER

This dataset is owned and supplied by the Dutch datamining company Sentient Machine Research, and is based on real world business data. This dataset and accompanying information can only be used for the COIL CHALLENGE 2000. For any other use, please contact Peter van der Putten, pvdputten@smr.nl. An earlier version of this dataset has been used in an small scale machine learning competition in the Benelux, however the data has been changed for the purpose of the COIL CHALLENGE 2000.

RELEVANT FILES

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

TICEVAL2000L.txt:
Dataset for predictions (4000 customer records). It has the same format as TICDATA2000.txt, only the target is missing. Contestants 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.

DATADICTIONARY

Nr	Name	Description
1	MOSTYPE	Customer Subtype
2	MAANTHUI	Number of houses
3	MGEMOMV	Avg size household
4	MGEMLEEF	Avg age
5	MOSHOOFD	Customer main type
6	MGODRK	Roman catholic
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
45	PWABEDR	Contribution third party insurance (firms)
46	PWAAND	Contribution third party insurane (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
66	AWABEDR	Number of third party insurance (firms) ...
67	AWALAND	Number of third party insurane (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 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 policie