CoIL Challenge 2000
Submitted Solution

Alexander K. Seewald

Austrian Research Institute for Artificial Intelligence,
Schottengasse 3, A-1010 Vienna
alex@seewald.at, alexsee@ai.univie.ac.at

Abstract. This paper describes my solution to the CoIL[1] Challenge 2000. The challenge was to predict who would buy a Caravan insurance and why. There were two subtasks: to predict caravan insurance ownership and to describe caravan owners according to this prediction model. My model was trained using a MetaCost[3] extended C4.5R8-clone and achieved a score of 109 out of a theoretical maximum of 238 while the winner achieved 121.

1 Introduction

In this paper I describe how I arrived at my solution to the CoIL Challenge, what methods and toolkits were used and some conclusions to be drawn from this confrontation with real-life data.

The WEKA[2] environment was used for all experiments. The described algorithms are all part of WEKA 3-1-7.

2 Initial considerations

At first glance it is quite problematic that there are only about 6% policy owners[3] in the training data. Any reasonable learning algorithm will therefore always predict CARAVAN=0, yielding an astounding accuracy of around 93%! To prevent this, I tried the meta-learning scheme boosting[4] with various base learners. The results were disappointing.

I also considered two types of cost sensitivity both of which strongly penalize a prediction of CARAVAN=1 for non-policy-owners[4].

[1] *CO*mputational *I*ntelligence and *L*earning Cluster (www.dsc.napier.ac.uk/coil) which aims to achieve scientific, technical and "social" integration of fuzzy logic, evolutionary computing, machine learning and neural networks

[2] Waikato Environment for Knowledge Analysis, available freely in source form at www.cs.waikato.ac.nz, see also [5].

[3] policy owners are instances with known class CARAVAN=1

[4] i.e. by giving each error that classifies a policy owner as CARAVAN=0 seventeen times more weight that vice-versa and thus compensating for the a priori unbalanced class distribution.

– to predict the class with the lowest misclassification cost (CSC[5]) based on a classifier that outputs class distributions[6]. This worked very well for NaiveBayes as base classifier.
– to make the base classifier cost-sensitive using the method described in [3]. This worked very well for j48 (a C4.5R8-clone) as base classifier.

These two variants with cost-sensitivity, namely CSC-NaiveBayes and MC-j48 were used in all subsequent experiments due to their superior performance. Some other algorithms were also evaluated in a less systematic way, the better ones were often of comparable performance.

3 Feature subset selection

Initial experiments with all the data using various machine learning algorithms yielded barely acceptable and unstable predictions. Therefore three sets of feature subset selections were considered. The first attribute index (MOSTYPE) is considered 1.

– Subset DT: 1, 6, 12, 13, 16, 17, 21, 22, 24, 29, 30, 37, 39, 42, 44, 47, 54, 59, 61, 86 which was a byproduct of running a Decision Table learner[1]. This is the final subset that was used by this learner on the training data. The classifier DecisionTable itself predicted CARAVAN=0 for all instances.
– Subset WrapperNB: 6, 8, 12, 13, 16, 21, 22, 24, 33, 38, 41, 45, 46, 47, 48, 49, 52, 55, 57, 58, 59, 60, 61, 63, 76, 79, 81, 82, 83, 86. This subset was generated by a subset evaluation wrapper[2] for CSC-NaiveBayes. CSC-NaiveBayes was used since it was the second-best algorithm but far less costly to train and evaluate than MC-J48.
– Subset Comb: 1, 5, 6, 8, 12, 13, 16, 21, 22, 24, 33, 38, 41, 42, 44, 45, 46, 47, 48, 49, 52, 54, 55, 57, 58, 59, 60, 61, 63, 76, 79, 81, 82, 83, 86, 87, 88, 89 where attr. 86 = attr. 42 concatenated with attr. 47, 87 = 65 & 68, 88 = 42 & 68, 89 = class CARAVAN. These three combined attributes[7] offered the highest lift on the training data. This subset includes most attributes from both wrapperNB and DT, some attributes were discarded and some were added arbitrarily.

Various learning algorithms evaluated on these subsets by two-fold cross validation[8] with differently randomized datasets. Interestingly, Subset wrapperNB was the best one but only slightly better than DT while Subset Comb was rather worse.

---

[5] after CostSensitiveClassifier in WEKA
[6] Distribution classifiers do not output a single class prediction but rather probabilities for every class.
[7] attribute names MINKGEM & PPERSAUT, AWAPART & APERSAUT, MINKGEM & APERSAUT
[8] This validation simulates the ratio of amount training data to test data which is also about 1:1.

4 Polishing

As I found out after the challenge, now would have been a good time to stop and submit a very good solution (a posteriori: 3rd place). Unfortunately I decided to continue optimizing...

On analyzing the data I found many inconsistent instances (i.e. same values for all attributes but differing values) and removed the minority class of them[9] for training since they might confuse the algorithm. About 20% of caravan policy owners were removed this way. This increased validation and training set performance slightly.

I also noticed that some evaluation set instances are identical to known instances from the training set and thus can be classified simple by remembering the whole training set[10] and assigning the class from the identical training instance. Thus I could classify 32 instances as CARAVAN=1 and 697 as CARAVAN=0. This known evaluation subset was used for the final comparison of learning algorithms although it was not truly "unseen data". Unfortunately, only seven of these 32 instances are actually caravan owners in the evaluation set. This points towards noise in attributes and/or class values. It may also be the case that there are simply not enough attributes to differentiate non-policy-owners from police-owners.

After choosing the final candidate algorithm by this known evaluation subset, I removed predictions that were "known" to be CARAVAN=0 and added predictions that were "known" to be CARAVAN=1, effectively restoring the perfect rote learner performance. Sadly, this further reduced my score[11].

5 Description Task

Fig. 1 describes the potential and actual caravan insurance customers. Since actual and potential customers are indistinguishable without knowledge which ones have caravan insurance, my model refers to both actual and potential customers, expecting to predict about seven times more potential customers than actual ones. The attributes are named exactly as in TICDATADESCR.TXT. Only predicted customers are shown as distinct, non-overlapping groups defined by appropriate attribute values and ranges, all subsets of the data that are not mentioned are presumed to have CARAVAN=0.

E.g. the first group would be all customers with PPERSAUT<=5 (Contribution car policies <= 62%), AFIETS<=0 (i.e. no bicycle policies), MGODRK>1 (more than 10% roman catholics in sociodemographic area) and APLEZIER>0 (at least one boat policy). It turns out that 20% of these customers are caravan policy owners. Unfortunately there are only five people with these properties in 5822 instances of training data. This is signified by '(5/20%)' behind CARAVAN=1. There are many such small groups in the decision tree which may be

[9] In case there was no clear minority class, all identical instances were removed.
[10] rote learning
[11] Successful prediction models are presumably less noisy than data.

explained because C4.5 tries to fit the target concept CARAVAN=1 too closely. A more loose fitting may give more insight in the data but less prediction accuracy and cannot be so easily verified. It should be considered if many of the people with CARAVAN=0 which have been assigned to CARAVAN=1 may be interested in a caravan insurance due to their similarity to existing caravan owners. The concept potential customer instead of actual customer may thus be considerably easier to learn.

To gain insight into the target concept I will restrict myself to groups of at least fifty persons.

The greatest group of this kind is in the lower third of the tree: PPERSAUT>5, PBRAND>2, PTRACTOR<=0, MBERBOER<=2, ALEVEN<=0 and PBRAND<=4 (i.e. 2<PBRAND<=4) yields 843 potential customers of which 16.4% are policy owners which is 2.7 times more than in the original dataset.

Others groups are PPERSAUT>5, PBRAND>2, PTRACTOR<=0, MBERBOER<=2, ALEVEN<=0, PBRAND>4, MOPLHOOG<=2 (52/8%) and PPERSAUT>5, PBRAND>2, PTRACTOR<=0, MBERBOER<=2, ALEVEN>0, MBERMIDD<=4 (111/14%). It is quite striking how all these large groups appear near to each other, and sharing at least four conditions (down to MBERBOER<=2).

It makes intuitive sense to presume that high contribution to car policies (PPERSAUT>5) and fire policies (PBRAND>2) correlates positively to caravan insurance ownership. That people with a higher contribution to tractor policies (high PTRACTOR) are less likely to own an insurance is less obvious but still plausible. In an area with lots of farmers (high MBERBOER) it also seems less likely to own a caravan insurance - presumably because less people have a caravan there. The number of life insurances (ALEVEN) seems to have a slightly negative impact. For a higher contribution to fire policies (PBRAND>4) the percentage of customers drops significantly, especially for areas with less than 23% of high level education (MOPLHOOG<=2, 52/8%).

Therefore, a description of a typical policy customer based on this model would be:

- high contribution to car policies (>62%, PPERSAUT>5) and fire policies (between 23% and 49%, 2<PBRAND<=4) [half as likely to be owner if contribution to fire policies is >=50% (PBRAND>4) especially when in area with high education < 23% (MOPLHOOG<=2)]
- no contribution to tractor policies (0%, PTRACTOR=0)
- lives in an area with at most 23% farmers (MBERBOER<=2)
- has a caravan (obviously, since otherwise s/he would not need insurance..)

These customers tend to care about safety issues (car / fire insurance), as long as it does not cost them too much. However they care more for their car, where they can afford more than 62% contribution while for fire insurance they only want to afford at most 49%. Clearly, once they have a caravan, caravan insurance will be interesting to them. They live in areas with few farmers where a caravan as home, and maybe even as the only home, is accepted and where people are more mobile and less "down-to-earth", less grounded in their surroundings. About one fifth of training data is of this type.

6 Prediction Task

These are the indices of polished predictions for CARAVAN=1 that were submitted.

2 3 14 18 21 36 39 43 44 45 51 52 53 57 78 82 85 89 92 112 114 116 123 129 141 144 145 147 151 153 162 165 171 177 180 183 208 210 211 213 216 219 224 227 230 232 243 244 255 271 276 277 279 280 285 286 287 292 297 306 308 311 312 313 314 320 322 324 331 339 341 342 347 352 354 357 362 365 369 374 382 388 390 399 401 408 411 422 425 426 442 443 444 445 454 455 460 466 470 489 506 514 519 521 529 540 542 567 572 573 575 576 578 579 588 596 601 615 629 634 639 641 649 652 655 658 667 670 683 687 692 696 707 717 723 729 732 736 737 747 751 754 758 761 762 771 775 780 787 789 797 810 818 829 834 836 839 843 848 863 877 879 888 892 895 904 913 915 918 920 922 932 935 942 945 946 948 964 966 969 971 972 982 991 993 998 1004 1007 1011 1018 1022 1029 1032 1034 1039 1051 1054 1070 1076 1083 1086 1088 1096 1097 1107 1118 1120 1141 1143 1145 1146 1153 1160 1163 1170 1175 1183 1186 1188 1191 1195 1200 1207 1214 1215 1224 1234 1236 1238 1247 1250 1253 1254 1255 1258 1271 1272 1277 1283 1287 1288 1292 1303 1304 1305 1312 1324 1331 1333 1334 1335 1348 1351 1352 1354 1364 1365 1368 1384 1385 1392 1401 1402 1406 1411 1413 1416 1417 1418 1430 1436 1443 1447 1448 1462 1465 1467 1469 1482 1490 1492 1499 1501 1506 1509 1514 1517 1519 1525 1527 1528 1535 1538 1546 1550 1561 1562 1563 1564 1566 1576 1579 1583 1585 1586 1590 1594 1607 1610 1612 1613 1619 1624 1627 1635 1637 1641 1642 1649 1657 1660 1661 1665 1671 1676 1684 1686 1690 1691 1694 1695 1697 1699 1702 1703 1706 1711 1712 1713 1717 1723 1725 1726 1735 1736 1739 1740 1764 1767 1772 1775 1779 1786 1787 1793 1796 1798 1805 1819 1820 1824 1826 1833 1835 1857 1859 1860 1865 1868 1874 1875 1876 1878 1885 1891 1896 1897 1898 1902 1913 1916 1918 1928 1929 1930 1931 1933 1935 1939 1952 1967 1968 1970 1971 1975 1979 1980 1991 1992 1996 2001 2003 2009 2017 2019 2023 2032 2033 2037 2040 2042 2043 2047 2048 2059 2073 2074 2076 2078 2082 2087 2089 2092 2099 2104 2105 2112 2113 2119 2122 2123 2129 2133 2134 2136 2141 2142 2146 2147 2148 2149 2150 2155 2156 2162 2165 2166 2171 2172 2200 2206 2208 2215 2217 2220 2225 2236 2239 2240 2241 2244 2253 2256 2267 2269 2274 2275 2279 2296 2303 2305 2310 2311 2316 2326 2335 2344 2347 2352 2353 2357 2371 2372 2373 2380 2387 2389 2390 2391 2395 2404 2405 2417 2425 2436 2443 2449 2451 2453 2454 2456 2462 2463 2466 2469 2471 2477 2479 2486 2489 2493 2497 2501 2503 2511 2516 2522 2534 2541 2545 2561 2576 2588 2597 2598 2601 2613 2622 2626 2629 2633 2635 2638 2642 2652 2654 2659 2660 2661 2663 2665 2668 2677 2679 2686 2691 2697 2698 2704 2711 2713 2714 2718 2734 2738 2749 2762 2768 2772 2775 2786 2789 2793 2798 2799 2803 2804 2809 2828 2830 2834 2836 2839 2846 2854 2855 2857 2858 2863 2865 2866 2868 2870 2872 2875 2878 2891 2892 2894 2896 2907 2930 2935 2942 2945 2956 2962 2969 2974 2982 2986 3001 3005 3008 3011 3013 3016 3017 3019 3026 3029 3034 3041

Fig. 1. Prediction Model (Decision Tree). The discussed subtree is marked with '!'

```
 PPERSAUT <= 5
 | AFIETS <= 0
 | | MGODRK > 1
 | | | APLEZIER > 0: CARAVAN=1 (5/20%)
 | AFIETS > 0
 | | PBRAND > 2
 | | | MGODOV <= 2
 | | | | MOPLHOOG <= 2
 | | | | | MINK3045 > 6: CARAVAN=1 (2/50%)
 | | | | MOPLHOOG > 2: CARAVAN=1 (2/50%)
 | | | MGODOV > 2: CARAVAN=1 (3/33%)
 PPERSAUT > 5
 | PBRAND <= 2
 | | MOPLHOOG <= 3
 | | | MGODRK > 0
 | | | | MINK3045 <= 5
 | | | | | MRELOV <= 4
 | | | | | | PBRAND <= 0
 | | | | | | | MGODOV <= 0
 | | | | | | | | MFALLEEN <= 1
 | | | | | | | | | MGODRK <= 1
 | | | | | | | | | | MBERMIDD <= 3
 | | | | | | | | | | | MINK3045 > 1: CARAVAN=1 (11/36%)
 | | | | | | | MGODOV > 0
 | | | | | | | | MBERBOER <= 1
 | | | | | | | | | MBERARBO > 2
 | | | | | | | | | | AFIETS > 0: CARAVAN=1 (4/25%)
 | | | | | | | | MBERBOER > 1
 | | | | | | | | | MGODOV > 1: CARAVAN=1 (4/25%)
 | | | | | | PBRAND > 0
 | | | | | | | MINK123M <= 0
 | | | | | | | | MGODRK <= 1
 | | | | | | | | | MBERMIDD <= 0: CARAVAN=1 (3/33%)
 | | | | | | | | | MBERMIDD > 0
 | | | | | | | | | | MGODOV > 1
 | | | | | | | | | | | MINK3045 > 2
 | | | | | | | | | | | | MINK3045 <= 3: CARAVAN=1 (3/33%)
 | | | | | | | | | | | | MINK3045 > 3
 | | | | | | | | | | | | | MRELOV <= 2: CARAVAN=1 (2/0%)
 | | | | | | | MINK123M > 0
 | | | | | | | | MBERARBO <= 2: CARAVAN=1 (3/33%)
 | | MOPLHOOG > 3
 | | | MAUT2 <= 0
 | | | | PBRAND > 1
 | | | | | MGODOV <= 3: CARAVAN=1: (2/0%)
 | | | MAUT2 > 0
 | | | | MBERBOER <= 2
 | | | | | PBRAND <= 0
 | | | | | | MFALLEEN <= 3
 | | | | | | | MGODOV <= 1
 | | | | | | | | MOPLHOOG <= 6
 | | | | | | | | | MAUT2 <= 1: CARAVAN=1 (8/38%)
 | | | | | | | | | MAUT2 > 1
 | | | | | | | | | | MFALLEEN <= 1
 | | | | | | | | | | | MINK123M <= 0
 | | | | | | | | | | | | MGODOV <= 0
 | | | | | | | | | | | | | MBERMIDD > 4: CARAVAN=1 (2/0%)
 | | | | | | | | | | | | MGODOV > 0: CARAVAN=1 (4/25%)
 | | | | | | | | | | MFALLEEN > 1: CARAVAN=1 (4/50%)
 | | | | | | | | MOPLHOOG > 6: CARAVAN=1 (3/33%)
 | | | | | | | MGODOV > 1: CARAVAN=1 (17/24%)
 | | | | | PBRAND > 0: CARAVAN=1 (5/20%)

!| PBRAND > 2
!| | PTRACTOR <= 0
!| | | MBERBOER <= 2
!| | | | ALEVEN <= 0
!| | | | | PBRAND <= 4: CARAVAN=1 (843/16%)
!| | | | | PBRAND > 4
!| | | | | | MOPLHOOG <= 2: CARAVAN=1 (52/8%)
!| | | | | | MOPLHOOG > 2
!| | | | | | | MAUT2 > 1: CARAVAN=1 (12/9%)
!| | | | ALEVEN > 0
!| | | | | MBERMIDD <= 4: CARAVAN=1 (111/14%)
!| | | | | MBERMIDD > 4
!| | | | | | MGODRK <= 1
!| | | | | | | MFALLEEN <= 2: CARAVAN=1 (12/17%)
!| | | | | | | MFALLEEN > 2
!| | | | | | | | PLEVEN <= 3: CARAVAN=1 (2/0%)

 | | | MBERBOER > 2
 | | | | PPERSAUT <= 6
 | | | | | PLEVEN <= 3
 | | | | | | MOPLHOOG <= 4
 | | | | | | | MBERARBO <= 3: CARAVAN=1 (40/13%)
 | | | | | | | MBERARBO > 3
 | | | | | | | | MGODRK > 0: CARAVAN=1 (4/0%)
 | | PTRACTOR > 0
 | | | PBRAND <= 5
 | | | | MAUT2 <= 3
 | | | | | MOPLHOOG <= 1: CARAVAN=1 (27/7%)
 | | | | | MOPLHOOG > 1
 | | | | | | MBERBOER > 1: CARAVAN=1 (4/25%)
 | | | PBRAND > 5
 | | | | PGEZONG <= 0
 | | | | | MRELOV > 1
 | | | | | | MBERARBO <= 0: CARAVAN=1 (3/0%)
 | | | | | | MBERARBO > 0
 | | | | | | | MGODOV > 0
 | | | | | | | | MBERBOER <= 0: CARAVAN=1 (5/20%)
 | | | | | | | | MBERBOER > 0
 | | | | | | | | | MBERARBO <= 1: CARAVAN=1 (2/0%)
 | | | | PGEZONG > 0: CARAVAN=1 (2/0%)
```

3048 3056 3062 3069 3077 3084 3088 3092 3093 3094 3099 3111 3112 3125 3132 3137 3139 3147 3148 3151 3154 3169 3172 3181 3186 3189 3192 3193 3194 3198 3199 3201 3207 3208 3220 3221 3229 3234 3240 3246 3254 3260 3261 3262 3264 3281 3282 3284 3292 3294 3297 3304 3305 3321 3323 3325 3329 3330 3334 3336 3337 3353 3354 3355 3357 3365 3370 3373 3379 3382 3397 3403 3404 3412 3414 3415 3418 3424 3442 3446 3451 3456 3460 3468 3469 3471 3473 3475 3483 3485 3492 3493 3504 3508 3510 3512 3515 3519 3526 3531 3534 3542 3558 3560 3566 3576 3584 3590 3594 3602 3603 3604 3605 3610 3616 3624 3634 3636 3650 3652 3658 3662 3664 3668 3673 3676 3678 3695 3698 3704 3717 3723 3726 3727 3736 3746 3760 3771 3773 3784 3792 3794 3795 3801 3810 3811 3813 3832 3833 3834 3837 3840 3841 3843 3852 3859 3863 3866 3872 3881 3883 3885 3892 3893 3895 3900 3903 3907 3911 3912 3913 3914 3915 3916 3919 3920 3922 3931 3937 3961 3974 3976 3984 3987 3996 3997 3998.

7 Conclusion

The dataset offered the same view over many different algorithms and subset selections: the more instances correctly classified as caravan policy holders, the more false positives there are while the ratio between true and false positives stays almost constant. This may be because the concept 'caravan policy owner' also applies to potential customers that do not yet have caravan insurance. A similarity between would-be- and already-customers is to be expected, so the learning algorithms may approximate the concept (would-be OR already)-customer.

My experiences with inconsistent instances and the consistent best-case performance of many different approaches during the competition hint that there may be too much noise in the data to get significantly better results. It may also be the case that there are not enough attributes to better differentiate non-policy-owners from policy-owners. Maybe there never will be enough - unless we can also resolve the issue of "free will" which undoubtedly also plays a role in choosing caravan insurance. From this viewpoint, finding 121 out of 248 policy owners is a fairly good result.

Personally, I enjoyed this challenge greatly and learned valuable practical lessons about data mining. I will certainly be back the next time.

References

1. Kohavi R.: The Power of Decision Tables, in Proceedings of European Conference on Machine Learning, 1995.
2. Kohavi, R., John G.: Wrappers for Feature Subset Selection, *Artificial Intelligence*, Special Issue on Relevance, Vol. 97, Nos 1-2, pp.273-324.
3. Domingos, P.: MetaCost: A general method for making classifiers cost-sensitive, Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining, pp. 155-164, 1999. Also available online at `http://www.cs.washington.edu/homes/pedrod/kdd99.ps.gz`.
4. Freund Y., Schapire R.E.: Experiments with a new boosting algorithm, Proc International Conference on Machine Learning, pages 148-156, Morgan Kaufmann, San Francisco, 1996.
5. Witten I.H., Frank E.: *Data Mining*, Morgan Kaufmann, Los Altos/Palo Alto/San Francisco, 1999.