CoIL Challenge 2000: Choosing and Explaining Likely Caravan Insurance Customers

YongSeog Kim and W. Nick Street
Management Sciences Department
University of Iowa
Iowa City, IA 52242 USA
{yong-s-kim,nick-street}@uiowa.edu

May 26, 2000

1 Task 1: Customer Prediction

Our prediction method combines artificial neural networks (ANNs) for prediction with evolutionary search for choosing the predictive features. The result is a predictive model that uses only a subset of the original features, thus simplifying the model and reducing the risk of overfitting while maintaining accuracy.

The feature subset search uses the Evolutionary Local Search Algorithm (ELSA). ELSA performs a "local" search in the space of feature subsets by evaluating genetic individuals based on both their quality (hit rate, which should be maximized, and complexity, which should be minimized) and on the number of individuals in the neighborhood of the individual in objective space. ELSA's bias toward diversity makes it ideal for multiobjective optimization, giving the decision maker a clear picture of Pareto-optimal solutions from which to choose. Previous research has demonstrated the effectiveness of ELSA for feature selection in both supervised [2, 3] and unsupervised [1] learning.

We employ a wrapper model by evaluating feature subsets selected by ELSA in the following manner. Each individual contains a subset of predictive features. A training set is constructed by randomly dividing the training set into two thirds of the points for training and one third for testing. The input features selected by the ELSA individual are used to train an ANN that predicts "buy" or "not buy." The ANN is tested on the test set, and the individual is evaluated both on the hit rate and the complexity (number of features) of the solution. The neural networks used standard sigmoid units trained with backpropagation. Each ANN had 5 hidden units.

Promising solutions found by ELSA were tested more extensively with five ten-fold cross-validation runs, where "promising" means high hit rate with not too many features. Our chosen solution uses only 21 features and has an expected hit rate of 16.72% in the top 20% of the predictions. Though we found another solution with slightly better hit rate of 16.92% using 43 features, we believe that reducing the feature subset is more valuable in this case. We note that the first feature (customer subtype) was omitted, and the fifth feature (customer main type) was recorded as ten binary variables. Other features were considered continuous and scaled to be in the same range. We further discuss the selected features in the next section, and include the 800 top predictions from the test set separately.

We also tested a number of other classification methods, including C4.5 decision trees, decision tree ensembles, and a "sieve" method that create a decision list by discarding attribute-value pairs that were likely to belong to customers who would not buy the insurance. The expected hit rates of these methods were found to be inferior to the ELSA/ANN combination.

Post-contest Analysis: Due to a simple indexing error on the test set results, our predictive model was not competetive in the contest. Following the release of the true results, we tested three different models found by ELSA to be good predictors. These models differ in the number of input features used. The following table summarizes the performance of these models on the test data, averaged over five runs.

	# Features	# Correct $\pm$ s.d	Hit Rate
Model 1	21	107.4 $\pm$ 0.89	13.45%
Model 2	43	113.6 $\pm$ 1.51	14.20%
Model 3	77	114.2 $\pm$ 2.16	14.28%

Of these, our chosen solution (Model 1) turned out to be the least accurate; all of the results are less accurate than our cross-validated estimates. Somewhat surprisingly, the neural network models with more features did not appear to overfit the training data, indicating that small increases in accuracy can be obtained by including marginally relevant features. As expected, the more complex models displayed a much higher variance.

2 Task 2: Customer Description

2.1 ELSA

The ELSA-based solution described above searches for and evaluates whole feature subsets, and thus incorporates feature interactions. The subset from our submitted solution includes the following features:

1. Customer subtype: Average family
2. Customer subtype: Career Loners
3. Customer subtype: Farmers
4. Other relation
5. High level education
6. Medium level education
7. Social class A
8. 2 cars
9. No car
10. Average income

11. Purchasing power class
12. Contribution private third party insurance
13. Contribution car policies
14. Contribution agricultural insurance polices
15. Contribution private accident insurance policies
16. Contribution social security insurance polices
17. Number of third party insurance (firms)
18. Number of delivery van polices
19. Number of motorcycle/scooter policies
20. Number of boat policies
21. Number of bicycle policies

2.2 Chi-square Test

ELSA evaluates feature subsets, rather than individual features, which we believe results in discovery of relationships that might otherwise go uncovered. However, subsets like the one above are difficult to evaluate all at once. In order to rank individual features, we constructed the distribution values for each feature, given the classification (yes or no). These distributions were normalized to the size of the smaller one (the no's), and a Chi-square performed to see if the distributions were significantly different. Twenty-one of the original 85 features had different distributions at the 95% confidence level. They are listed below, in order of significance, along with the significance level.

1. Contribution of car policies (1.0000)
2. Contribution fire policies (1.0000)
3. Number of car policies (1.0000)
4. Customer main type (1.0000)
5. Average income (1.000)
6. Purchasing power class (0.9999)
7. Income $<$ 30.000 (0.9997)
8. Lower level education (0.9992)
9. Contribution private 3rd party insurance (0.9991)
10. No car (0.9987)
11. Rented house (0.9965)
12. Home owners (0.9961)

13. Social class A (0.9908)
14. 1 car (0.9899)
15. Income 45-75 (0.9866)
16. High level education (0.9861)
17. Customer subtype (0.9847)
18. Number private 3rd party insurance (0.9845)
19. Married (0.9806)
20. Other relation (0.9684)
21. Social class D (0.9568)

There were large drops in the value of the test statistic between items 1 and 2, and between 3 and 4.

2.3 Association Rules

We also conducted a search for simple association rules (one attribute-value pair) that would predict the purchase of a caravan policy. Minimum support was set at 10% of the positive cases (0.00598), and minimum confidence was set at 10%. Seven rules were found using these thresholds. They are listed below in no particular order, along with the confidence and support values.

1. if (middle class families = 8) then yes (0.009, 0.150)
2. if (customer type = driven growers) then yes (0.011, 0.131)
3. if (high level education = 37-49%) then yes (0.006, 0.113)
4. if (purchasing power class = 7) then yes (0.012, 0.141)
5. if (contribution car policies = 6) then yes (0.045, 0.113)
6. if (contribution fire policies = 6) then yes (0.026, 0.123)
7. if (number of car policies = 2) then yes (0.007, 0.154)

2.4 Conclusions

Clearly the strongest single predictors of caravan policy purchases are those features measuring the number of and contribution to car policy purchases. Based on the Chi-square test, Contribution Car Policies showed by far the largest difference between buyers and non-buyers. This variable was also found to be relevant by the other selection methods. The effect is roughly linear; people who spend more than 1000 guilders on car insurance are most likely to be caravan policy buyers, and the more they spend, the more likely a buyer they are. The number of car policies is also significant, as found by two of the methods; surprisingly it was not included in the ELSA/ANN model, probably because of a high degree of correlation with the corresponding contribution variable. Based on both the Chi-square test and the association rules, it appears that the most likely caravan policy buyers purchase insurance on more than one car. People with no car policies are extremely unlikely to buy caravan insurance.

The income level of the customer (or, more specifically, the customer's neighborhood) is also important, as several income and purchasing power-related variables were found to be significant. The effect here is not necessarily linear, however. For instance, the likelihood of buying a caravan policy for those with very low purchasing power is correspondingly low, and increases with increased purchasing power, but only up to a point. At very high levels of purchasing power, the likelihood of buying a caravan policy turns around and comes down somewhat. This seems to indicate that the target population contains most of the upper quartile of purchasing power, but *not* the highest levels.

Variables relating to education level were also found to be significant. We attribute this to the fact that education correlates closely with income, and do not believe that education level should be considered to have an independent effect on caravan policy purchases.

The amount spent on fire policies was also found to be important. This result is less intuitive and may warrant further investigation. In some sense, the fact that a family buys *any* kind of insurance makes them more likely to buy a caravan policy, and surely purchasers of fire insurance are likely to be in the upper-middle-class target population.

Further analysis of customer type was warranted by the significance of both customer main type and customer subtype. Looking first at customer main type, we find that the most likely buyers belong to the driven grower category, followed by successful hedonist (whatever that may be). However, the living well category was unlikely to buy the policy, again suggesting that the target population is upper-middle-class, but not extremely wealthy. One of the biggest surprises in our analysis was the fact that the categories of "cruising seniors" and "retired and religious" were unlikely to buy the caravan policy. This would seem like a group that would be likely to both buy a caravan, and to insure it. Perhaps the stereotype of the retired couple cruising around the county in their mobile home is more of an American phenomenon. More likely, we have found some inconsistency in the way that customer types are recorded by the policy sales staff. This should be further investigated. Less surprising was the fact that farmers are extremely unlikely to buy caravan policies.

Customer subtype is harder to analyze because of the small number of individuals in most categories. Consistent with our previous conclusions, the most significant difference in terms of positive correlation with caravan policy purchases is the category of "middle class families." Other categories with strong likelihood of buying policies include, in order, "affluent young families," "high income, expensive child," and "high status seniors." Categories unlikely to buy policies (at least twice as many no's as yes's, when normalized for class priors) include "young and rising," "mixed rurals," "large family farms," "young, low educated," and "low income Catholics." All of these are consistent with previous conclusions, and are based on small samples.

In conclusion, we suggest a partial customer profile based on the available input features: Established families, upper-middle to upper wealth level, at least two insured automobiles, with other insurance policies a plus. People to avoid include low-income, no cars, farmers and, surprisingly, senior citizens.

References

[1] Y. Kim, W. N. Street, and F. Menczer. Feature selection in unsupervised learning via evolutionary search. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-00)*, 2000, accepted.

[2] F. Menczer, M. Degeratu, and W. N. Street. Efficient and scalable pareto optimization by evolutionary local selection algorithms. *Evolutionary Computation*, 8(2):223–247, Summer 2000.

[3] F. Menczer, W.N. Street, and M. Degeratu. Evolving heterogeneous neural agents by local selection. In V. Honavar, M. Patel, and K. Balakrishnan, editors, *Advances in the Evolutionary Synthesis of Neural Systems*. MIT Press, to appear.