COIL CHALLENGE 2000 ENTRY

Charles Elkan
Department of Computer Science and Engineering 0114
University of California, San Diego
La Jolla, California 92093-0114
elkan@cs.ucsd.edu

CUSTOMERS WITH CARAVAN POLICIES

The strongest single predictor of having a caravan insurance policy is having a single car insurance policy where the contribution is high (level 6), or having two car policies.

The other predictors that are most statistically significant are:
- "purchasing power class" is high (5 or higher, especially 7 or higher)
- a private third party insurance policy
- a boat policy
- a social security insurance policy
- a single fire policy with higher contribution (level 4)

Intuitively, these predictors identify customers who have a car and are wealthier than average, and who in general carry more insurance coverage than average. It is not surprising that these are the people who are most likely to have caravan insurance.

A customer with a single car policy with a low premium (level 5) is less likely than average to have caravan insurance. This finding is actionable and it was not obvious in advance. A customer of this type is presumably less wealthy or less risk-averse, so less likely to own a caravan or less likely to buy insurance for it.

Because of the small number of training examples of holders of caravan policies, there are no other predictors of having caravan insurance that are both tatistically significant and not highly correlated with the predictors mentioned above. Any analysis of who has a caravan policy that is more complex than the analysis above is likely to be the result of overfitting the training data.

In particular, none of the available demographic attributes are statistically significant independent predictors. This includes which lifestyle segment a customer belongs to. It also includes information about the population in a customer's geographic area. As is common in commercial data mining, only data about the wealth and personal behavior of individuals is useful here.

LEARNING METHOD

The method used here to obtain a predictive model, and to answer the questions above, is naive Bayesian learning. Boosting was tried also, and some derived attributes were added to the training set. After two important derived attributes were added, boosting did not give any significant increase in accuracy on a validation set taken from the training set, and neither did adding more derived attributes. Therefore, the results here do not use boosting, and only use two derived attributes.

In commercial data mining, demographic attributes, including customer segmentations by lifestyle, income, etc., often do not add any predictive power when behavioral data is available. This phenomenon is clearly visible here, so all demographic attributes were discarded, except attribute 43, "MKOOPKLA_Purchasing_power_class".

The two derived attributes added give the most detailed information possible about existing car and fire insurance policies. The range of the new "car" attribute is the cross-product of the ranges of the existing attributes 47 and 68. These attributes are "PPERSAUT_Contribution_car_policies" and "APERSAUT_Number_of_car_policies". The range of the new "fire" attribute is a similar cross-product for fire policies.

Taking a cross-product allows a different probability of having a caravan policy to be associated with each combination of a number of policies and a total premium amount (called "contribution"). Because the "number" and "contribution" attributes are already discretized, any other derived attribute constructed from these two attributes (for example an average policy premium defined as "contribution" divided by "number") must lose information compared to the cross-product.