COIL CHALLENGE 2000 Submission

Tom McKone and Curt Stenger
Analytic Resources
919 N. Michigan Avenue, Suite 750, Chicago, IL 60611, USA
phone: 312-787-9200
fax: 312-787-6374
email: tom.mckone@analyticres.com

1. Prediction

The model that best fit the training data in predicting caravan policy owners (CARAVAN>0) was a Boosted Decision Tree model created in SGI's MineSet using an adaptation of the AdaBoost boosting algorithm (Bauer and Kohavi, 1998) built into MineSet. Attached is the list of the top 800 (20%) index numbers of the test data set.

The prediction task was to classify potential CARAVAN policy holders without penalty for misclassifying CARAVAN policy non-holders, so long as the 6% of policy holders fell within the top 20% of the prediction. The model was run with an associated loss matrix that made the absolute loss associated with CARAVAN holders and non-holders roughly equal. In other words, a misclassification of a policy holder was penalized at about 15 times the penalty of a misclassification of a non-policy holder. This loss matrix lowers the overall classification rate of the model, but does substantially increase the rate for the target class of policy holders. Our experience has held that classification prediction works best when the classifiers have equal presence. Following this view, caravan policy holders were underrepresented. The loss matrix adjusted for this.

In the model used, we let all the proportion ranges and contribution ranges remain as integers instead of categories. Within these variables, order does matter and we felt it best to allow the algorithm the freedom to aggregate, or not, as best fit the model. We did make the two customer type variables, MOSTYPE and MOSHOOFD, into categories. To deal with the collinearity and sparseness issues of policy contribution and number of policies, we removed all the individual "number of policy " variables and created three aggregate number of policy variables; ATOTAL (all policies excluding

CARAVAN), APROPTOT (all property type policies except CARAVAN) and AEVENTOT (all life and accident policies). We felt this was a better solution than creating contribution per policy variables because of the different spans associated with each contribution level.

As Figure 1 and Table 1 indicate, the boosted decision tree does a very nice job in identifying policy holders while still keeping the classification rate of non-holders at a decent level.

Figure 1: Lift Curve for CARAVAN policy holder

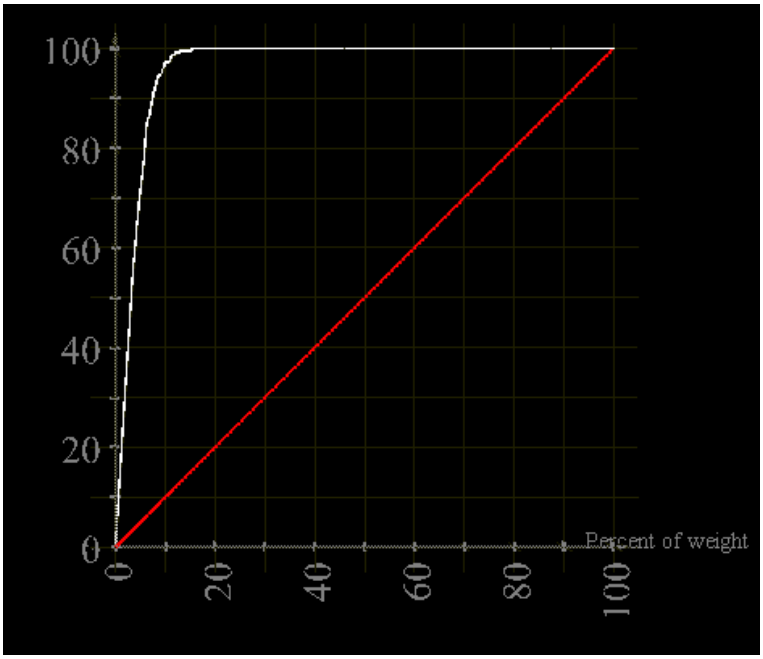

Table 1: Classification Rates

Class	Total	% have policy in top 20%	% have policy in top 6%
Have policy	1.00	1.00	0.81
Don't have policy	0.83	-	-
Total	0.84	-	-

2. Description

Even though our selected model prediction is based on a vote of a number of different tree-based models, the description can be derived from our initial decision tree. If fact, the single decision tree model had such a similar structure across several bootstrap samples that we feel confident the assertions we make hold true for the boosted model as well.

Target Group Descriptions

The model developed a number of potential CARAVAN customer groups. The majority of these groups are currently more likely to have insurance policies other than the CARAVAN policy. One interesting group does not carry insurance, other than the CARAVAN policy, and should make an excellent prospecting target.

Among current insurance holders, car policies (PPERSUAT) and fire policies (PBRAND) are the only insurance contribution factors that indicate CARAVAN coverage. The minimum contribution levels of these two types of policies set by the decision tree do isolate 53% of the CARAVAN holders, indicating their importance. Drilling down from these two policy indicators, a variety of demographic variables describing wealth, class, education, marital status and religious preference refine the CARAVAN policy targets. Table 2 lists the insurance contributions and demographic characteristics of the key current insurance holding targets. These targets represent the best chance of finding a CARAVAN holder among the total training group, not necessarily the most CARAVAN holders.

Table 2: Insurance Holding Targets

Target	% CARAVAN holders	% CARAVAN holder among rule group
Car policies > 1000 gilders + Fire policies > 200 gilders + < 50% of area with low education levels + Average income of area > 68,000 gilders + < 25% of area single + < 25% of area defined as having no religion	27.3%	28.7%
Car policies > 1000 gilders + Fire policies < 100 gilders + > 63% of area is considered high class	3.2%	22.5%
Car policies > 1000 gilders + Fire policies between 100 and 199 gilders + > 50% of area with low education levels + belonging to one of 6 customer types; LIVING WELL, RETIRED AND RELIGIOUS, FAMILY WITH GROWN UPS, CONSERVATIVE FAMILIES, SUCCESSFUL HEDONISTS or DRIVEN GROWERS	8.1%	15.1%

Among those without any insurance, except CARAVAN policies, a prospect target emerges. The demographic variable, proportion of skilled laborers in the area, keys this segment. The full rule associated with this group is:

Target	% CARAVAN holders	% CARAVAN holder among rule group
< 25% of area has an income below 30,000 gilders + < 36% of area defined as having no religion + the number of total insurance policies excluding CARAVAN = 0 + > 0% of the area are skilled laborers	5.8%	12.4%

These descriptions of the rules associated with particularly key CARAVAN targets are extremely useful in developing future target mailings. Interpreting the input variables and the splits or division points of the decision tree also provide additional insights that often transcend the specifics of the data set examined.

Interpretation – Who buys CARAVAN insurance and why?

People who have CARAVANs have CARAVAN insurance policies. The insurance has to be driven by the presence of the product. Whether CARAVANs are hitch-trailers or full-sized RVs (Recreational Vehicles), both are rather sizeable investments that are more than likely to be insured. The insurance marketer has two sets of inputs to consider, the types and levels of policies that current CARAVAN policy holders have and the demographic characteristics of them, which are probably the demographics of CARAVAN owners.

From our model, CARAVAN policy holders contribute quite a bit of money to their car policies. This could be because they need a rather substantial car or truck to pull a hitch-trailer or, more likely, because they have a high valuation for vehicles. From previous RV studies done in the U.S., we have seen that recreational vehicle owners enjoy driving and hold their vehicles in higher regard than other portions of the population. The higher car contributions in this data may indicate that this holds true in Europe as well.

The second insurance indicator found in our model, fire insurance contributions, is less clear. Fire insurance could be an indicator of moderate wealth and safety. If the property insurance input also represents homeowner type insurance, then the fire policies may be additional policies carried for the safety of the home. This would seem to represent a conservative investment, consistent with many of the demographic indicators. In fact, high insurance contributions for both fire and car may be an indication of the conservative nature of CARAVAN policy holders.

Different income inputs, proportion of area under 30,000 gilders and average income, were early indicators for CARAVAN policy holders. Both developed similar economic profiles, moderate to high income but definitely not low income. This fits the cost profile of CARAVANs themselves, which may be out of the purchase range of lower income individuals.

Even though one target group identified High Class as an important variable and another had the customer type of SUCCESSFUL HEDONISTS, most of the target groups had a class profile that was middle class and conservative. The largest target group had an educational level that was not low, but was not high either. Areas with low proportions of single people and low proportions of non-religious people were strong indicators, representing the conservative bent of CARAVAN owners. The customer types of FAMILY WITH GROWN UPS, CONSERVATIVE FAMILIES and RETIRED & RELIGIOUS support this conservatism as well. The prospect target of skilled laborers found in the model also has the income and conservative religious

demographic indicators found in the insurance carrying targets. Age may also be a factor in identifying CARAVAN owners even though it was not an indicator in this model. Since age came from the demographic overlay, it is the average age of the area, not the individual's actual age, which could be quite misleading.

CARAVAN insurance policy holders own CARAVANs. Since CARAVAN owners appear to be primarily middle or working class with conservative values and enough money to afford the product, it is unlikely someone will own a CARAVAN without insuring it. Insurance marketers could identify prospective CARAVAN owners or existing owners looking to switch insurance through lists like church and crafts union membership. They may also be able to entice switching by bundling CARAVAN and car policies, since existing CARAVAN holders also have high car policy contributions.