Excerpts from Data Mining Project - caravan insurance challenge
Shigeru Odani

Background

The data used for this analysis came from the 2010 Computational Intelligence and Learning Cluster (CoiL) Challenge, a data mining competition sponsored by the EU. The original data set is a real-world-business database from an insurance company, supplied by the Dutch data mining company, Sentient Machine Research. The software used for the following analysis is SAS Enterprise Miner.

Two data sets are supplied

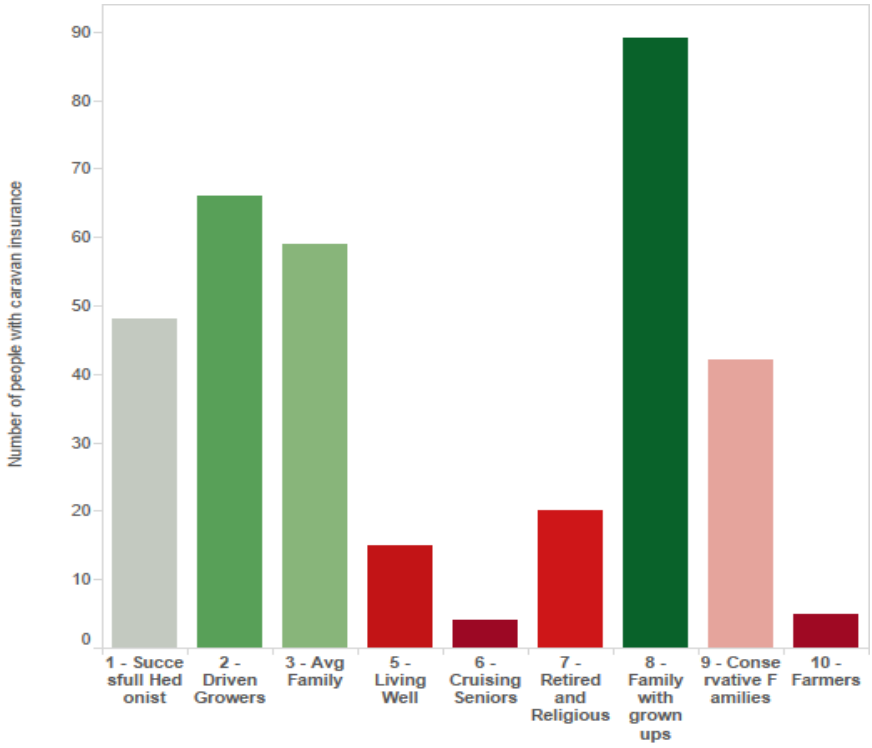
- The training data set, containing 5,922 observations, is used to build and validate the model
- The validation data set, containing 4,000 observations, is used to test the predictive power of the best performing model against other past competitors

For both data set, there are 86 attributes for each record. The attributes were grouped into 43 socio-demographic variables and 43 insurance product ownership (number of policies of various types and level of contribution to those policies) variables, one of which is the target variable, ownership of Caravan-mobile home insurance policy. All the insurance product ownership variables were coded as categorical in an ascending order (ordinal).

Socio-demographic variables

Derived using user zip code, some of these variables could potentially be good proxies to individual information such as average income, purchasing power, car ownership (found to be related to caravan insurance). However, the drawback is that customers from the same zip code have identical information pertaining to those attributes, and with data at such an aggregate level, the socio-demographic variables may not contribute significantly to the predictive power of the model. It was decided to drop all variables of this type except 2: average income and customer type because of the following reasons:

- average income derived using zip code can be a good estimate of the individual's income level, and there could be a relationship between income and Caravan ownership.
- some customer types are more likely to purchase caravan insurance as can be seen in the graph below. Alternatively, this variable could be treated as predefined clusters of customers.



Insurance product ownership/usage variables

The variables in this group include the number of insurance policies and the contribution toward certain type of insurance by each individual. Some types of insurance can provide predictive power such as:

- car insurance: customers with car insurance are more likely to own a caravan and have caravan insurance
- property insurance: risk-averse customers are more likely to insure their caravan

Choice of algorithms and rationale

For a binary target variable, the following supervised learning algorithms were applied:

- Logistic Regression
- Decision Tree (CART)
- kNN
- Neural Networks
- SVM

There are advantages and disadvantages for each model used. Decision Tree deals best with missing value and categorical variables while the output is usually easy to interpret and provides insights to the data. Some models (Logistic Regression, Decision Tree) provide more

transparency of predictors' importance in explaining the target variable than others (Neural Networks, kNN). In addition, because the predictor variables are predominantly categorial, kNN and Neural Networks modeling process can be computationally intensive if the data set is large.

Analysis

a) What's the best way to model the problem?

2 judgement calls were made early in the modeling process

- Because of the rare case of the target variable (~6%), I decided to use ROC Index for the validation set as the selection criteria.
- Wanted to cut down the number of variables from 86 to 40-50. For this reason, with the exception of Customer type and Avg Income, I removed 41 Zip-code level variables.

Modeling started with 45 variables (44 from the dataset and 1 cluster id) and it was found that if you reduce the number of variables to just the significant ones, you get much better results, and that this could be achieved using Stepwise Regression. Further, ordinal variables can be coded as interval, and in some cases this can lead to better results as well as greater computational efficiency (this helped when working with the larger set of variables).

The chart below shows the 5 best performing model / variable combinations. The ones in green have a reduced set of variables, coding the ordinal variables as ordinal. The ones in white code the ordinal variables as interval.

Model	Valid: ROC Index
Neural Network6	0.797
Standard reg training6	0.79
Standard reg training4	0.787
Stepwise reg training4	0.785
Stepwise reg training6	0.785

b) Which predictors are important?

Some variables are more intuitive than other in terms of their predictive power, for example, people with higher average income are more likely to have caravan insurance and people who have boat policies are more likely to also have caravan insurance, possibly because of an overlap in lifestyles. On the other hand, predictors such as contribution to car policies are not as easy to intepret especially given the mixed signs of the coefficients.

The table below shows the significant variables in the model.

Parameter	Codes	Coefficient Estimate
Number of boat policies	0	-1.74
Number of boat policies	1	0.50
Number of agriculture insurance	0	0.67
Avg Income		0.10
Customer Category	1	1.43
Customer Category	2	1.86
Customer Category	3	1.27
Customer Category	4	-8.14
Customer Category	5	1.14
Customer Category	6	-0.31
Customer Category	7	0.98
Customer Category	8	1.29
Customer Category	9	1.27
Contribution fire policies	0	1.47
Contribution fire policies	1	1.11
Contribution fire policies	2	0.24
Contribution fire policies	3	1.99
Contribution fire policies	4	2.26
Contribution fire policies	5	1.57
Contribution fire policies	6	0.46
Contribution bicycle policies	0	-0.44
Contribution car policies	0	4.12
Contribution car policies	4	-2.74
Contribution car policies	5	4.13
Contribution car policies	6	5.50
Contribution car policies	7	-5.23
Contribution disability policies	0	3.10
Contribution disability policies	4	-5.60
Contribution disability policies	5	-2.77

How the model can be applied to solve business challenges

The model is seen as a solution to a marketing problem: user-level targeting, low list costs, high marketing costs, low conversion rates. Often in marketing, target customer lists can be acquired fairly inexpensively. However, many lists contain large numbers of unqualified individuals, and these unqualified individuals can consume a good amount of marketing budget. As such, marketers often want a way of marketing to only qualified individuals. This model is a way of achieving this.

A marketer can take the target customer list (list of people to potentially market to) and calculate an expected profit for each one and then target only people with a positive expected profit using the formula P(CARAVAN=1) * Customer Lifetime Value - Marketing cost.

The graph below shows the profits a marketer can achieve at any given marketing budget allocation comparing this model with a random draw of leads. With an average customer lifetime value of $3,000 and average marketing cost of $100/lead, it can be seen that a profit of above $250,000 can be achieved by spending as low as $80,000 on marketing to the most qualified (highest probability of buying caravan insurance) 800 leads in the database, while it would cost $320,000 on marketing to 3,200 randomly chosen leads.