Dr Philip Brierley
Tiberius Data Mining Pty Ltd
http://www.tiberius.biz/

Problem Outline

Our real-world dataset for this year's competition was donated to us by a consumer finance company with the aim of possibly finding better solutions for a cross-selling business problem.

The company currently has a customer base of credit card customers as well as a customer base of home loan (mortgage) customers. Both of these products have been on the market for many years, although for some reason the overlap between these two customer bases is currently very small. The company would like to make use of this opportunity to cross-sell home loans to its credit card customers, but the small size of the overlap presents a challenge when trying to develop a effective scoring model to predict potential cross-sell take-ups.

A modeling dataset of 40,700 customers with 40 modeling variables (as of the point of application for the company's credit card), plus a target variable, will be provided to the participants. This is a sample of customers who opened a new credit card with the company within a specific 2-year period and who did not have an existing home loan with the company. The target categorical variable "Target_Flag" will have a value of 1 if the customer then opened a home loan with the company within 12 months after opening the credit card (700 random samples), and will have a value of 0 if otherwise (40,000 random samples).

A prediction dataset (8,000 sampled cases) will also be provided to the participants with similar variables but withholding the target variable.

The data mining task is a to produce a score for each customer in the prediction dataset, indicating a credit card customer's propensity to take up a home loan with the company (the higher the score, the higher the propensity).

Data Preparation

The data pre-processing required depends on the software or code that is building the model and how it can automatically handle different data types.

The main data manipulation carried out prior to modelling was in regard to the special values of 98 and 99 for the bureau variables. These cannot be used ‘as is’ if the variables are to be treated as continuous values as opposed to nominal categorical. As most of the variables represent counts it makes sense to have the ability to use them as continuous variables.

These special values were replaced with null values in the data. Two extra binary fields were then created for each original field with a ‘1’ representing the special value (either 98 or 99) and a ‘0’ for all other values.

Modelling Technique Used

The modelling technique used was a blend of the results from two models built using different algorithms. The first algorithm was a neural network and the second a form of general additive model. Both of these algorithms are present in the Tiberius data mining suite and were developed by the author.

The two models were developed with 70% of the data, with the remaining 30% used as a validation set to ensure over learning was not taking place. Each model selected its own random 70%.

The scores from the two models were then used as the inputs to a neural network, with the known target as the output. This final model gave the weightings of how the two models should be combined. Only 1 hidden neuron was used in the neural network, so essentially it was logistic regression, which could also have been used.

The results of the blending are given below in terms of the AUC, the metric that the models are being evaluated by. The higher the values then the better the model.

	Train (70%)	**Validation (30%)**	**All**
Neural Network	0.741	0.7375	0.727
GAM	0.713	0.7365	0.721
Combined	-	-	0.7375

Note, the original values were the Gini coefficients and were transformed to the AUC using the formual AUC = (0.5 * Gini) + 0.5.

It is interesting to note that calculating the AUC on the combined training and validation set for the neural network model gives a worse result than the two individual sets.

The combined model AUC is a significant improvement on each of the individual models.

More models were generated and the blending process repeated. The combined models AUC continued to fall as more models were added, but not massively. As the submission of the 2 models had already been made to the competition, we could not thus submit the models where the size of the committee was larger.

Finally the data distribution of the modelling and scoring sets were compared, along with the final score distributions of the model on the two sets. The comparisons showed the distributions to be very similar, so it is expected that the final ROC on the scoring set will be in the range 0.72 – 0.74.

Discussion

The algorithms used contained automatic variable removal methods, so that the final models only contained a handful of the 40 variables available. Examining these variables can give an understanding of what the important characteristics are for cross sell opportunities.

A reasonable model (training ROC = 0.701) can be achieved using just 4 variables

Number of Bureau enquiries in the last 6 months for mortgages
Number of bureau enquiries in the last 12 months for loans
Age at application
Number of months at current residence

Don't target these…

Mortgage enquires last 6 months = 0
and
Months at current address > 40
and
Age at application < 23 or > 44

22.3% of the population but only 7.8% of the mortgages

Target these…

Mortgage enquires last 6 months > 1
and
Months at current address < 36

Only 2.6% of the population but 12.4% of the mortgages.