

PAKDD-2007 Competition Submission

By John L. Ries and N. Scott Cardell
Salford Systems
14 April, 2007

Acknowledgments

We acknowledge the contributions of fellow team members Dan Steinberg (President of Salford Systems and our team leader), Mikhail Golovnya, and Don Cozine to this analysis. We also thank Nathaniel Noriel, who provided helpful and timely answers to our data questions, and Jerome Friedman, the inventor of the MART algorithm, which is employed by the TreeNet® software we used throughout our analysis. Finally, we thank the organizers of the competition and the providing company for the use of this interesting data set.

Data Preparation

The modeling data, as received, included numerous variables that were not usable in their raw forms. In particular, the credit bureau variables had embedded codes signifying whether or not a credit bureau record was found (code=98) or whether a search was made (code=99). Since these codes were the same for all numeric credit bureau variables, we decided to create a new character variable, BUREAUSEARCH, defined as follows:

```
if B_ENQ_L6M_GR1=98 then let bureausearch$="No match found"  
  else if B_ENQ_L6M_GR1=99 then let bureausearch$="Did not go"  
  else let bureausearch$="Match found"
```

New versions of the original credit bureau variables were created as follows:

For numeric variables:

- Code 98 (no match found) recoded to 0
- Code 99 (did not go) recoded to missing

For character variables:

- All values other than "Y" or "N" recoded to missing

A list of these variables follows:

Old Variable	New Variable
B_ENQ_L6M_GR1	B_ENQ_L6M_CRED
B_ENQ_L6M_GR2	B_ENQ_L6M_LOAN
B_ENQ_L6M_GR3	B_ENQ_L6M_MORT
B_ENQ_L12M_GR1	B_ENQ_L12M_CRED
B_ENQ_L12M_GR2	B_ENQ_L12M_LOAN
B_ENQ_L12M_GR3	B_ENQ_L12M_MORT
B_DEF_PAID_L12M	B_DEF_PAID_L12M_v2

Old Variable	New Variable
B_DEF_PAID_IND	B_DEF_PAID_IND_v2
B_DEF_UNPD_IND\$	B_DEF_UNPD_IND_v2
B_ENQ_L1M	B_ENQ_L1M_v2
B_ENQ_L3M	B_ENQ_L3M_v2
B_ENQ_L6M	B_ENQ_L6M_v2

Since the sum of B_ENQ_L6M_GR1, B_ENQ_L6M_GR2, and B_ENQ_L6M_GR3 was sometimes less than B_ENQ_L6M, we assumed that there were some enquiries that did not fall into the three categories specified in the data dictionary. We therefore created a variable, B_ENQ_L6M_OTH, representing the difference.

The self-reported account and credit card indicators included a number of values other than “Y”, or “N”. We therefore elected to create new versions of these variables as well, where miscellaneous values were recoded to missing. A list of these variables follows:

Old Variable	New Variable
SAV_ACCT_IND	SAV_ACCT_IND_v2
AMEX_CARD	AMEX_CARD_v2
DINERS_CARD	DINERS_CARD_v2
VISA_CARD	VISA_CARD_v2
MASTERCARD	MASTERCARD_v2
RETAIL_CARDS	RETAIL_CARDS_v2

We discovered 16 records in the modeling data set where CURR_EMPL_MTHS=1000; we took this as a code for “employment length unknown” and therefore created new versions of CURR_EMPL_MTHS and PREV_EMPL_MTHS (named CURR_EMPL_MTHS_v2, and PREV_EMPL_MTHS_v2). We also created a new variable EMPYRS, defined as $(CURR_EMPL_MTHS_v2 + PREV_EMPL_MTHS_v2) / 12$ (number of years employed in the most recent two positions), and RESYRS, defined as $(CURR_RES_MTHS + PREV_RES_MTHS) / 12$ (number of years resided at two most recent addresses).

There were also a few records where AGE_AT_APPLICATION was less than RESYRS by a small amount (1-2 years). In response, we created a new variable, APPAGE2, which was equal to AGE_AT_APPLICATION if greater than RESYRS; or RESYRS rounded down to the nearest unit otherwise.

After running a series of experiments aimed at determining optimal sample size, we elected to use all of the data in our modeling. We also determined that there were no significant data patterns distinguishing the prediction data set from the

modeling data set.

We note that while our final list of predictor fields included only a small subset of these variables, we tried all of them in the early stages of our analysis.

Modeling

Our models were developed with TreeNet® which estimates a nonparametric score that is on a logistic scale. Interestingly, on the logistic scale, there was little difference in the fit between additive models and freely interacted models. In a preliminary step we selected 11 important predictors to work with. These were:

Variable Name	Description
b_enq_l3m_v2	# Bureau Enquiries in the last 3 months (codes removed)
b_enq_l12m_loan	# Bureau Enquiries in last 12 months for Loans (codes removed)
disp_income_code	Indicates monthly Disposable Income
age_at_application	Age at application time
marital_status	Marital status
curr_empl_mths_v2	Indicates total number of months at Current Employment.
b_enq_l6m_mort	# Bureau Enquiries in last 6 months for Mortgages (codes removed)
rent_buy_code	Residential Status code
a_district_applicant	District of Residence of the Applicant (using residential address)
resyrs	Total # years at last two residences (current and previous)
curr_res_mths	Indicates total number of months at Current Residence.

These eleven variables were selected from the original list of predictors on the basis of a long process of removing variables one at a time to determine how their absence affected the model (positively, negatively, or neither). This small list outperformed all the other ones attempted in numerous tests.

The final set of scores was created by averaging the scores from two different TreeNet models. The first, consisting of 7203 six-node decision trees, was estimated with a 70% learning sample and a 30% test sample. The second, consisting of 10,000 six-node decision trees, was estimated on the whole modeling data set. Since TreeNet is a nonparametric procedure, based on

ensembles of decision trees, there are no parameters to report, but the list of relative importances is as follows:

A_DISTRICT_APPLICANT	100.0000000	100.00	*****
RESYRS	88.6246915	88.62	*****
RENT_BUY_CODE\$	87.5236923	87.52	*****
B_ENQ_L6M_MORT	84.9339535	84.93	*****
MARITAL_STATUS\$	82.1344945	82.13	*****
CURR_RES_MTHS	79.5451987	79.55	*****
AGE_AT_APPLICATION	79.2231350	79.22	*****
CURR_EMPL_MTHS_V2	73.3787106	73.38	*****
DISP_INCOME_CODE\$	71.2483373	71.25	*****
B_ENQ_L3M_V2	50.6855800	50.69	*****
B_ENQ_L12M_LOAN	45.1015367	45.10	*****

We also estimated an additive model (employing two-node trees), discussed in the following section, which provided the following list of relative importances:

A_DISTRICT_APPLICANT	100.0000000	100.00	*****
RENT_BUY_CODE\$	87.9204882	87.92	*****
RESYRS	87.6147839	87.61	*****
MARITAL_STATUS\$	82.5447646	82.54	*****
B_ENQ_L6M_MORT	81.4174899	81.42	*****
CURR_RES_MTHS	79.0658000	79.07	*****
AGE_AT_APPLICATION	76.0926410	76.09	*****
CURR_EMPL_MTHS_V2	70.1394479	70.14	*****
DISP_INCOME_CODE\$	67.9143183	67.91	*****
B_ENQ_L3M_V2	49.8343311	49.83	*****
B_ENQ_L12M_LOAN	42.7099129	42.71	*****

Discussion of the Model Results

District of residence of the applicant was the single most important predictor. Of the eight districts only three of them were very different from each other and the other five. Inflation in housing prices varies by geographic area. Those home owners with little or no equity in their house will not be able to cash out home equity by refinancing or taking a second mortgage. Even refinancing in order to reduce their interest rate may be impossible. These results suggest that geographic indicators can pick up enough of these effects to be useful predictors. They also suggest that better proxies for a homeowner's equity in their house could improve targeting models. Finally, as district appears to be acting as a proxy for other variables it could improve targeting to make a definitive determination of the variables for which it is acting as a proxy.

The RESYRS (total number of years in their current residence and their last residence), and RENT_BUY_CODE (categorical variable with levels RENT, MORTGAGE, OWNER, PARENTS, and BOARD) are virtually tied as the next two most important predictors. It is interesting that RESYRS is more important than time at the current residence, which is also in the model. RESYRS that is less than 4 years has a large positive effect. For larger times the effect slowly drops down until it reaches a small negative asymptote at about 20 years. In between, there are some small but noticeable peaks that may have to do with the typical

lengths of mortgages and the recent past history of mortgage rates. The most important policy consequence is to suggest focusing on individuals whose are in their first residence for less than 4 years or who have moved twice in the last four years. The RENT_BUY_CODE results are very much what one would expect, those with mortgages are the most likely to apply for a mortgage, this would include refinancing, second mortgages and a new mortgage if they move. Those who own with no mortgage and those who board are next. Boarding is likely to be a temporary arrangement, either while looking for a house or will attending college this makes sense. Renters are next followed by those who live with there parents. Living with one's parents may be a negative indicator of the wherewithal to purchase a home and many renters are likely to continue renting rather than purchase.

Other variables had largely expected effects. Some notable exceptions were that widowed had a much larger negative effect than separated or divorced; the difference between married and widowed is about 2 and a half times the difference between married and separated or divorced. Age shows the highest positive effect for age less than 28 and then declines to its lowest level for ages between 50 and 58, then it rise sharply to a mid range value from age 66 or so on. Time at current employment has a small positive effect for very low tenures then gradually settles to a small negative effect for tenures between 8 and 22 years, after which it increases with the rate of increase becoming quite sharp after 26 years. When combined with lower likelihood of owners without mortgages seeking a mortgage this may indicate a time at which individuals will either seek a new mortgage or choose to pay off their current mortgage and become owners.

An interesting result occurs with enquiries. Enquiries for mortgage loans in the last six months has the expected monotonic increasing effect. However, total bureau enquiries in the last 3 months has a sharp peak at 3 inquiries, with zero enquiries having the only negative effect; while for loan enquiries in the last 12 months zero enquiries has a positive effect, with 1, 2 and 3 to 5 enquiries having increasingly negative effects, for more than 5 enquiries the effect increases monotonically becoming large and positive by 10 enquiries. For income being in the lowest income group has a strong positive effect, but the other groups show very differences among themselves.

The results on time in current residence show the only noticeable difference between the interacted model and the additive model. In both case very low tenures have positive effects which decline to a negative trough from about 4 to 13 years tenure and then increase to a higher level. However, for the interacted model the level increases steadily to a new asymptote at around 30 years tenure the positive effect at the asymptote is about have the positive effect of the very shortest tenures. The only break in the last increase is a small local peak around 20 years tenure.

For the additive model the asymptote for large tenures is at a lower level and there is a noticeable peak between 18 and 24 years that is at a higher level then the asymptote. This suggest that some behavior in that time period, perhaps refinancing of current mortgages, tends to occur. Furthermore it suggests, that

a more complicated model with interactions picks up that effect without needing to use tenure in current residence.

Copyright © 2007 Salford Systems, San Diego, California, USA.