Ensemble PROBIT Models to Predict Cross Selling of Home Loans for Credit Card Customers

Hualin Wang, AllianceData, USA
Yan Yu, AllianceData, USA
Kaixia Zhang, AllianceData, USA

ABSTRACT

Ensemble a set of PROBIT models to predict the likelihood of buying a home loan from the current company which has the credit card base customers. The buying rate is very low and data is very limited. This approach offers a stable and robust way to solve this extremely difficult and yet very common business problem.

Keywords: PROBIT Models, Ensemble Modeling, cross-selling

INTRODUCTION

The 11th Pacific-Asia Knowledge Discovery and Data Mining Conference (PAKDD, 2007) hosted another data mining competition, co-organized by the Singapore Institute of Statistics. More detailed information on the competition can be found at http://lamda.nju.edu.cn/conf/pakdd07/dmc07/.

Outline of Approach

Because the target is 0 or 1 valued, a natural candidate for modeling the cross-selling propensity is the class of PROBIT regression models:

\[ p = \text{probability}(Y=0) = C + (1-C)*F(X\beta) \]

where \( p \) is the probability for the response to be 0; \( C \) the natural response rate; \( X \) a set of explanatory variables; \( \beta \) a vector of parameter estimates; and \( F \) a link function, usually a cumulative distribution function, such as the normal, logistic function or extreme value.

When the link function is the logistic function, this is usually called logistic regression; and when the link function is the cumulative distribution function of the normal distribution, it’s called PROBIT regression. For more infor-
Some believe that PROBIT tends to outperform logistic regression if there is an underlying function with thresholds that correspond to outcome categories (0 and 1 in this problem). Our belief for the current question is that the underlying function is the tradeoffs made by consumers in order to make their decision on whether to buy home loans from the company. Factors that influence their decision include whether they own a home, how much mortgages they have, financial charges by candidate companies, and so forth. Because there is no clear and obvious answer to which approach to take, we decided to conduct some comparative studies. Our analyses show that PROBIT tends to have higher and more stable c-statistics for simply random subsamples, as well as bootstrapping samples.

Other analysis helps to determine the range of weights for the PROBIT model. The weights that we proposed are not related to sampling per se, but to assigning different weights for different outcomes in the likelihood function while estimating the parameters. Practically, we observe that models have “optimal” target rate for best performance. For example, logistic regression performs well when the target rates are around 18%. Put another way, the logistic regression does only marginally better than ordinary least linear regression if the target rates are around 50%. The weights can virtually change the target rates in the sample. With some trial and errors and comparisons of results, we find that using weights from 3 to 12 produces ideal results. The default value of natural response for all models is to be 0.

With these factors, the final model is built by following these 2 steps:

1. Pick any integer for weight between 3 and 12, build an ensemble of 10 PROBIT models using 10 bootstrapped samples and average the 10 obtained probabilities. This is the model for the selected weight. At the end of this process, there are 10 ensemble models corresponding to the 10 different weights (from 3 through 12). For more information about bootstrap, see “An Introduction to the Bootstrap” by Efron and Tibshirani (1998). For more information about ensemble model, see “Solving Regression Problems Using Competitive Ensemble Models” by Frayman, Rolfe, and Webb (2002).

2. For each observation, remove the largest as well as the smallest probabilities and compute the mean probability of the remaining eight probabilities. This average probability based on a scoring mechanism similar to a diving scoring system is the final predicted value.

Exploratory Data Analysis
There are 40 possible raw explanatory variables, but B_DEF_UNPD_L.12M is equal to 0 for all records. Because there are not many raw attributes, one of the particular challenges for this problem is to create more predictive variables out of the 39 raw ones.

For categorical variables, check their frequency distributions and the target rate which is the percent of records with TARGET_FLAG = 1. Based on these results, we do some regrouping or binning. For example, the Frequency Distribution & Target Rate for BUY_RENT_CODE is shown in Figure 1 above.

The “X” valued category is extremely small (only 137 records out of 40,700). It should be recoded as one of the other values. In this case, we recode it as “M.”

For numerical variables, we do capping, ranking, Box-Cox transformations, and other nonlinear transformations. For example, the ranks for CURR_EMPL_MTHS are shown in Table 1.

Based on this result, we create a new variable N_CURR_EMPL_MTHS as follows:

```
IF CURR_EMPL_MTHS<=5 THEN N_CURR_EMPL_MTHS =0;
ELSE IF CURR_EMPL_MTHS<=12 THEN N_CURR_EMPL_MTHS =1;
and so on.
```