S01 - 2008

Long-Term Value Modeling in the Automobile Industry

Jeff Ames, Ford Motor Company, Dearborn, MI
Cathy Hackett, Trillium Teamologies, Royal Oak, MI
Bruce Lund, Marketing Associates, Detroit, MI

ABSTRACT

Businesses often classify their customer base in terms of the customers' predicted long-term value (LTV). LTV may influence marketing strategies, particularly CRM and concern resolution. This paper describes an approach to LTV calculations in the automobile industry. The emphasis of this presentation is on one aspect of LTV which is the choice of "next new-vehicle segment". SAS code related to "next-segment" predictive modeling is outlined. The implementation of this predictive model within a SAS scoring platform is presented.

INTRODUCTION

Automotive manufacturers can identify their customers through their transactional interactions with the manufacturer. When a new-vehicle is sold by a dealer, the manufacturer obtains the name and address of the buyer of a new-vehicle as well as the VIN (vehicle identification number), sale date, and selling dealer. By searching the manufacturer's historical sales-records by name and address, the past vehicle purchase(s) by this customer can be linked to the current purchase to create a "purchase history". This matching process is usually performed "by household" rather than by consumer. Many automotive manufacturers also survey their customers to measure their satisfaction with the vehicle and dealer. service informationis reported by VIN and then linked to the customer. Demographic overlays are also available from several large consumer database compilers. These dataset sources provide the source data for the development and maintenance of a Customer Database.

Using the Customer Database, it is possible to create an estimate of the Long-Term Value (LTV) of a household to the automotive manufacturer. For the purpose of this paper, the LTV for a household is the net-present-value of future vehicle and finance profits to be earned from the household by the manufacturer.

This paper defines the LTV concept and then discusses one of the components of LTV called the "next new-vehicle segment choice” model. Both the modeling methodology and SAS programming techniques that were used to develop the "next new-vehicle segment choice” model are discussed in more detail.

THE NEXT NEW-VEHICLE SEGMENT CHOICE MODEL DEFINITION

Each automotive manufacturer organizes its portfolio of cars and trucks into “vehicle segments” based on size, price, and functional attributes. The goal is to placea vehicle into a segment directly corresponds to vehicles from other manufacturers. This segmentation is typically defined by the Product Development Office and could involve up to 20 segments. For example, in the case of Ford Motor Company, the Ford Fusion and the Mercury Milan are assigned to the “CD Car” segment – a segment which includes several competitor nameplates.

For each household, the **next new-vehicle segment choice model** provides probabilities (summing to one) for buying one of the manufacturer’s vehicles within a segment, given that the household does buy another new-vehicle from the manufacturer within the next year.

For each household, the next new-vehicle segment choice model probabilities are combined with the manufacturer's variable-profit estimates for the segments to obtain a predicted "profit of the next purchase" – to be denoted by PROFIT in the formula below:

$$\text{PROFIT} = \sum_j \text{Probability (Segment “j“ | Buys from Manufacturer)} \; * \; \text{(Profit of Segment “j”)}$$

A segment classification is used for this model rather than a name-plate approach (e.g. “CD Car” vs. “Fusion” or “Milan”) due to the relative permanence of segments versus the fluctuations in name-plates.

THE LTV CONCEPT OVERVIEW

In addition to the estimate for the "profit from the next purchase", other models in the LTV calculation provide estimates for:

1) The time in months of possible future new-vehicle purchases from the manufacturer T_1, T_2, T_3 …
2) The probabilities that these future purchases occur at these times.

A monthly discount factor "r" is added as an externally provided parameter to discount future profits to the present.

As a simplifying assumption, it is assumed that PROFIT applies to each future purchase date T_1, T_2, T_3 …

These estimates and assumptions are incorporated in the LTV formula shown below:

$$\text{LTV} = \sum_{T_i} \{\text{Prob of Purchase}\} * \{\text{Profit}\} \ / \ (1 + r)^{T_i} \text{ where } i = 1, 2, 3, \text{etc.}$$

The discussion now focuses on the development and implementation of the next new-vehicle segment choice model (hereafter, the "**segment model**")

METHODOLOGY: THE NEXT NEW-VEHICLE SEGMENT CHOICE MODEL ("Segment Model")

There is no claim to originality regarding the methodology of the modeling technique of "pair-wise logistic regression" presented in this paper in the following pages. The theoretical properties of this technique were described by Begg and Gray (1984).

Methodology:

The eligible population for modeling consisted of households in the Customer Database that had acquired a new-vehicle from the manufacturer in the past 10 years and who purchased a new-vehicle from the manufacturer during an 11th "observation year". The target variable value was the segment of (the first) new-vehicle purchased during the observation year. There were 16 segments.

The population of households eligible for modeling was randomly split into two equal samplesone held out for performance evaluation. Sample sizes were adequate for this application.

Factors which influence the vehicle-segment value of the target variable would certainly include the vehicle-segment of prior vehicle purchases. Many households in the Customer Database, however, had only one prior new-vehicle purchase from the manufacturer. To allow for flexibility in creating predictor variables, the analysis dataset was divided into (i) households with 1 prior purchase, (ii) households with 2+ prior purchases.

Exploratory data analysis identified 121 variables for consideration as predictorsbut only about half of these could be used for subpopulation (i). The predictor variables were valuated as of the day before the start of the observation period.

Pair-wise logistic regression was selected as the modeling technique (in preference to multinomial logistic regression). Pair-wise logistic regression involved, in this case, 15 binary logistic regressions for buyers from each of 15 vehicle-segments versus a single-fixed "reference vehicle-segment". The choice of "reference vehicle-segment" is important to the success of the pair-wise technique and the optimal characteristics of the reference segment include:

1. Stability over time - the vehicle-composition of the reference segment should remain constant into the future.
2. Differentiation - the reference segment should be distinguishable from other segments both physically and in perception.
3. The size of the reference segment (total vehicle count) should at least be equal to the average segment size.

The buyers from the reference vehicle-segment were re-used for each of the 15 pair-wise logistic regressions.

The pair-wise analysis was performed separately for subpopulations: (i) households with 1 prior, (ii) households with 2+ priors. SAS PROC LOGISTIC was used with the STEPWISE selection option. The modeling was automated via a SAS "next-segment" macro which looped through the 15 regressions, with each regression being presented with the same list of candidate predictors. The coefficients from the regressions were captured as SAS datasets using the OUTEST option. Coefficients of un-selected variables are included in the OUTEST dataset with a value of missing.

The 15 regressions (separately for subpopulations (i) and (ii)) are combined according to these formulas:

- Using the "xbeta's" (= sum of coefficients * predictors) from the 15 pair-wise regressions, the probability for the reference segment is determined by:

$$P(\text{Reference}) = 1 / (1 + \sum_{i=2}^{16} e^{\text{xbeta}_i}) \ \ldots \text{(A1)}$$

- The probabilities for the other 15 vehicle-segments are determined by:

$$P(\text{Segment } i) = e^{\text{xbeta}_i} * P(\text{Reference}) \quad \text{where i=2 to 16 (15 segments)} \ldots \text{(A2)}$$

Explanation: P(Segment i) / P(Reference) estimates the number of occurrences of “i” over the number of occurrence of “reference” … this is the empirical odds-ratio. But the odds ratio of “i” over “reference” is estimated by the e^{xbeta_i} of the binary logistic regression of “i” versus “reference” This leads to (A2).

So, necessary conditions between the vehicle-segment probabilities and the binary logistic regression “xbeta’s” are:

$P(\text{Segment } i) = e^{\text{xbeta}_i} * P(\text{Reference})$ for i = 2 to 16
$P(\text{Segment } 2) + \ldots + P(\text{Segment } 16) + P(\text{Reference}) = 1$

Regarding the quantities P(Segment i) and P(Reference) as 16 unknowns in 16 linear equations given above, this set of linear equations is non-singular. By direct substitution the unique solution is shown to be given by (A1) and (A2).

Maintenance and Updates of the Segment Model:

- Assuming the 121 predictor variables are adequate for the future, the segment model can be easily refit by running the “next-segment” macro with the OUTEST datasets containing new coefficient values and, very likely, with a different set of predictors selected by the STEPWISE option. The impact on scoring is simply to update the coefficient datasets (the new datasets contain the same list of variables as the old – some selected by LOGISTIC, others un-selected).
- In the future, if a segment is eliminated due to changes in the automotive market, provided it is not the reference segment, the segment model does not need to be refit. The xbeta’s for the eliminated segment in formulas (A1) and (A2) are set to zero.
- If a new segment is to be added, then (assuming the 121 predictors are believed to be adequate), a single binary regression of the new segment versus the reference segment is performed. The new segment-model scores are computed by A1 and A2 (adding to the index). Provision was made in the Customer Database Scoring Logic to add new segments with no impact on the scoring program.

Advantages of the pair-wise regression approach versus multinomial regression

- PROC LOGISTIC has a multinomial option and could be used to fit the segment model. Advantages of the pair-wise approach versus multinomial include: (i) reduced size of the problem, (ii) predictors may appear in only a subset of the 15 binary regressions, in contrast to multinomial regression where a selected predictor will have coefficient values for all choices. However, a comparative study of pair-wise logistic versus multinomial logistic was not performed in preparation for this paper.

SCORING THE CUSTOMER DATABASE:

Twenty-one million households in the Customer Database are re-scored each week by more than 100 different models. The weekly scoring process starts with a merge of two large aggregated datasets, each keyed uniquely by household ID (HH_id). An efficient method to score the 100+ models is to compute and score the models in a single data step – the “Scoring Data Step”. The logic for each model is inserted into the data step either as a %Include statement or as a SAS macro.

In the case of the segment model the coefficients from the 30 (= 2 * 15) regressions were saved from the PROC LOGISTIC OUTEST option including both used and un-used coefficients for a total of 3,630 (= 121 * 2 * 15). These 30 datasets of coefficients are merged together to form a dataset with a single row and then read into the "Scoring Data Step" as part of the “%Macro run_scores” – shown below:

Reading in the coefficients as a dataset is certainly superior to hard coding 3,630 coefficients. These 3,630 coefficients are retained for the duration of the data step where they remain in memory but are dropped when the data step completes executing. The impact of the segment model on the time to perform the weekly re-scoring of the Customer Database was insignificant.

The segment model scoring code is outlined below.

Index definitions for the code below:
i = the different models (1 or 2). Model 1 is for population 1 of households with only one historical new-vehicle purchase from the manufacturer. Model 2 is for households with 2+ such purchases.
j = segments (2 to 16) Segment "1" is excluded since "1" refers to the base segment)
k = predictor variables or coefficients (1 to 121) There are 121 coefficients for each segment 2 to 16.

The code below has two PARTS: (1) Creation of a dataset with one observation containing the 3,630 coefficients. (2) Applying the coefficients to the household predictor variables to create segment model scores.

PART 1

```
/* In the Proc Logistic Outest option the names of the predictors and the coefficients
are the same.  These names appear in the &var list */

%let vars = intercept /* and another 120 variables */;
%let d1 = _link_ _type_ _status_ _name_ _lnlike_;

/* if the coefficient is unused (= .), then assign 0 */
%macro z_coeff(i,j);
data z_coeff_&i._&j;
/* Set the coefficient datasets from Proc Logistic Outest option */
   set coeff_&i._&j (drop = &d1);
   drop k &vars;
   array coeff_&i._&j{121} &vars ;
   array array_of_coeff_&i._&j{121} c_&i._&j._1 - c_&i._&j._121;
   /* if the coefficient is unused (= .), then assign 0 */
   do k = 1 to 121;
      if coeff_&i._&j{k} = . then coeff_&i._&j{k} = 0;
      array_of_coeff_&i._&j.{k} = coeff_&i._&j{k};
      end;
run;
%mend;

%macro initialize_coeff;
     %do i = 1 %to 2;
        %do j = 2 %to 16;
           %z_coeff(&i,&j);
           %end;
        %end;
%mend;
%initialize_coeff;

/* z_coeff is a dataset with one observation consisting of 3630 coefficients named c_i_j_k
   where i = 1 to 2, j = 2 to 16, and k =1 to 121 with missing values re-set to zero */
%macro merge_coeff;
data z_coeff;
     merge
     %do i = 1 %to 2;
        %do j = 2 %to 16;
           z_coeff_&i._&j
           %end;
        %end;
        ;
%mend;
%merge_coeff;
run;
```

PART 2

```
%macro run_scores;

data scored_household; merge indata1 indata2; by HH_id;

/* Not shown are many line of code to create the “vars” and “population” */
```

```
   drop i j k
   %do i = 1 %to 2;
      %do j = 2 %to 16;
         %do k = 1 %to 121;
            c_&i._&j._&k
            %end;
         %end;
      %end;
      ;
   array pvar{121} &vars ;
   array xbeta{2,15}
   %do i = 1 %to 2;
      %do j = 2 %to 16;
         xbeta_&i._&j
         %end;
      %end;
      ;
   array pred{15}
   %do j = 2 %to 16;
      pred_&j
      %end;
      ;
   array coeff{2,15,121}
   %do i = 1 %to 2;
      %do j = 2 %to 16;
         %do k = 1 %to 121;
            c_&i._&j._&k
            %end;
         %end;
      %end;
      ;
   array xcoeff{2,15,121}
   %do i = 1 %to 2;
      %do j = 2 %to 16;
         %do k = 1 %to 121;
            xc_&i._&j._&k
            %end;
         %end;
      %end;
      ;
   retain
   %do i = 1 %to 2;
      %do j = 2 %to 16;
         %do k = 1 %to 121;
            xc_&i._&j._&k 0
            %end;
         %end;
      %end;
      ;

   if _n_ = 1 then
   do;
/* z_coeff is a dataset with one observation consisting of 3630 coefficients named c_i_j_k
   where i = 1 to 2, j = 2 to 16, and k =1 to 121 with missing values re-set to zero */
      set z_coeff;
      do i = 1 to 2;
         do j = 2 to 16;
            do k = 1 to 121;
               xcoeff{i,j-1,k} = coeff{i,j-1,k};
               end;
            end;
         end;
      end;

/* Begin scoring - segment probabilities are saved in “pred_j” where j = 1 to 16 */

   intercept = 1;
   if population = 2 then i = 2; else i = 1;

   do j = 2 to 16;
```

```
      xbeta{i,j-1} = 0;
      do k = 1 to 121;
         xbeta{i,j-1} = xbeta{i,j-1} +  xcoeff{i,j-1,k}*pvar{k};
         end;
      end;

   if population = 1 then do;
      temp01 = 0;
      do j = 2 to 16;
         temp01 = temp01 + exp(xbeta{1,j-1});
         end;
      pred_1 = 1 / (1 + temp01);
      do j = 2 to 16;
         pred{j-1} = exp(xbeta{1,j-1}) * pred_1;
         end;
      end;
   if population = 2 then do;
      temp02 = 0;
      do j = 2 to 16;
         temp02 = temp02 + exp(xbeta{2,j-1});
         end;
      pred_1 = 1 / (1 + temp02);
      do j = 2 to 16;
         pred{j-1} = exp(xbeta{2,j-1}) * pred_1;
         end;
      end;
%mend;

%run_scores; run;
```

CONCLUSION

The technique of pair-wise logistic regressions is a practical alternative to multinomial logistic regression and has more flexibility in the selection of predictor variables. Choices (i.e. values of the target variable) can be dropped (but not the reference choice) or added without having to refit the complete model. The development of a comprehensive collection of predictor variables and a SAS macro program which loops through the pair-wise logistic regressions provide a simple process to refit the coefficients and, very likely, change the selection of predictors. The program to score the database is not impacted by drops, adds, or refits.

REFERENCES

Begg, C. B. and Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions, Biometrika 71, 1, pp. 11-18

ACKNOWLEDGMENTS

Bruce Reading of Trillium Teamologies, Inc. initially developed the Customer Database scoring process.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Jeff Ames
Ford Motor Company
Dearborn, MI
E-mail: james@ford.com

Cathy Hackett
Trillium Teamologies, Inc.
Royal Oak, MI
E-mail: chackett1@ford.com

Bruce Lund
Marketing Associates
Detroit, MI
E-mail: blund@ford.com