Using PROC LOGISTIC, SAS MACROS and ODS Output to evaluate the consistency of independent variables during the development of logistic regression models. An example from the retail banking industry

Alex Vidras, David Tysinger

Merkle Inc.

ABSTRACT

Predictive models are used extensively in customer relationship management analytics and data mining to increase the effectiveness of marketing campaigns. Logistic regression remains at the forefront in analytics as the most popular technique used to predict customer behavior. Particularly with direct mail marketing, logistic regression models are built using previous campaigns that span several months in length, posing a major challenge to statisticians to devise a way to not only capture seasonality across these campaigns but to also evaluate the stability of these models. Millions of dollars are spent annually on marketing activities that utilize logistic regression models. Therefore the predictive ability and robustness of logistic models is essential for executing a successful direct mail campaign. This paper shows how PROC LOGISTIC, ODS Output and SAS MACROS can be used to proactively identify structures in the input data that may affect the stability of logistic regression models and allow for well-informed preemptive adjustments when necessary. Thus we are introducing a standardized process that industry analysts can use to formally evaluate the impact and statistical significance for predictors within logistic regression models across multiple campaigns and forecasting cycles.

KEY WORDS

Logistic regression, predictive model stability, independent variables screening, response models, database marketing

INTRODUCTION

Logistic regression models built using SAS procedures like PROC LOGISTIC or PROC GENMOD are frequently deployed in marketing analytics to assess the probability that:

a) A customer or prospect will purchase a product or service
b) A customer will leave the company
c) A customer/prospect will respond to a direct mail, email or other marketing stimulus
d) Other binary outcomes (e.g. cross selling, coupon redemption etc.)

Given the significant level of marketing budget spent annually on marketing activities driven by logistic regression model results, understanding the predictive ability of these models is paramount. The initial cost of developing and subsequently deploying a predictive model is substantial, although costs are declining with the use of automated tools like Enterprise Miner. The stability of a logistic regression model is largely dependent on the variables that make up the final model. During development, it is important that independent variables get screened for consistent historical performance, which will maximize the chance of stable performance in the future and thereby the longer-term power and stability of the predictive model.

BACKGROUND

Predictive models use historic results to predict the future and this can be challenging in today's fast changing market. Using input variables that are robust over time is essential for the success of predictive models and this is the focus of this paper.

Logistic regression models are often used in direct marketing campaigns to predict response. The characteristics driving successful response can change over time due to mass advertising, or other market forces. If a modeler can identify fields that are susceptible to market changes and either treat them (capping,

recoding etc.) or exclude them from the model, this will maximize the ability of the model to accurately predict future responders.

Logistic regression models tend to utilize data from multiple points in time. This is to ensure that adequate sample size is used for the model and to control for seasonality. The issue introduced by using multiple campaigns or time points for the development of predictive models, is that independent variables can have inconsistent behavior in terms of their relationship with the dependent variable, at different points in time. These points in time can be past campaigns, months or sales data from different quarters.

THE ISSUE – An example from the retail banking industry

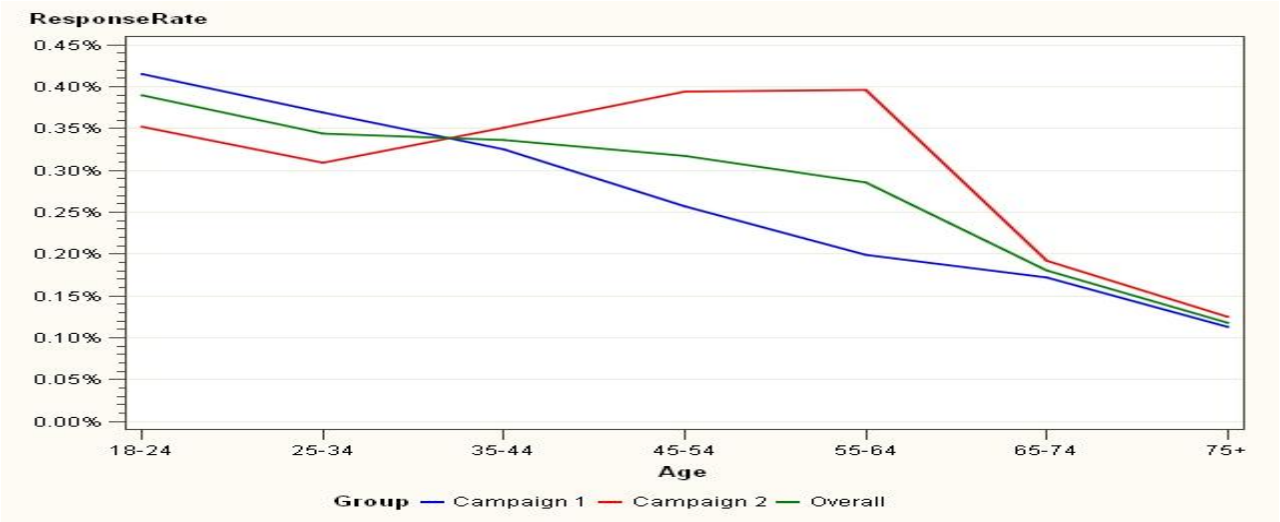
For this example we will use a direct mail campaign that targets banking prospects with an offer to open a checking account. To account for seasonality and to ensure adequate sample size, the bank chose to use two historical campaigns for the model development. 1,500 demographic, behavioral, lifestyle and geo–demographic independent variables were appended to the prospect universe. Table 1 shows the response rates for the two input campaigns, split by age, which is one of the 1,500 available independent variables:

Table 1: Response Rates by Age Group & Campaign

Age	Campaign 1			Campaign 2			Overall (Combined)		
	Mailed	Percent Mailed	Response Rate	Mailed	Percent Mailed	Response Rate	Mailed	Responders	Response Rate
18-24	13,721	2%	0.415%	9,260	1%	0.352%	22,981	90	0.390%
25-34	113,525	13%	0.369%	81,980	13%	0.309%	195,505	672	0.344%
35-44	211,009	24%	0.325%	158,910	25%	**0.351%**	369,919	1,244	0.336%
45-54	191,112	22%	0.257%	149,740	23%	**0.394%**	340,852	1,081	0.317%
55-64	146,860	17%	0.199%	115,060	18%	**0.396%**	261,920	748	0.286%
65-74	99,068	11%	0.172%	71,560	11%	0.192%	170,628	308	0.180%
75+	87,044	10%	0.113%	59,770	9%	0.125%	146,814	173	0.118%
Total	**862,340**	**100%**	**0.257%**	**646,280**	**100%**	**0.33%**	**1,508,620**	**4,315**	**0.286%**

As you can see from Table and Graph 1, the distribution of age between the two campaigns is very similar. However the response rate by age has a monotonic decrease in campaign 1, while in campaign 2 the response rate increases for age groups 35-44, 45-54 and 55-64.

Graph 1: Response Rates by Age Group & Campaign



In checking account acquisition campaigns, younger groups tend to have higher response rates, mainly because they are more active in the demand deposit accounts (DDA) market, compared to older groups that have more stable banking relationships. Based on this assumption, campaign 1 follows the expected response pattern. With regards to campaign 2, banking executives decided to shift a large percentage of

mass media dollars to the older demographics, just prior to campaign execution, in order to increase penetration in the investment accounts market. This strategy affected the effort to sell checking accounts, directly and in-directly through cross selling at the point of sale, which elevated response for the 35-64 age groups.

The issue with the above scenario is that after combining campaigns 1 and 2 to build a response model using PROC LOGISTIC, in the combined dataset as shown by the column “Overall (Combined) Response Rate” of Table 1, the inverse relationship between age and response is dampened, compared to the relationship in campaign 1. This will affect the logistic regression coefficients and therefore the final model. Based on Table 1, both the strength and the direction of the relationship between age and response, are affected by mass media advertising which decreases the confidence in using age to predict response both short term and long term.

Table 2 has the output of PROC LOGISTIC when fitting a simple PROC LOGISTIC model using the combined modeling dataset and age as the only independent variable. Under this scenario, the parameter estimate of the independent variable age is -0.1391, meaning that the log of the odds of responding to the direct mail campaign, decreases by 0.1391 when age increases by one year. The parameter estimate of age when fitting a simple logistic regression for only campaign1 is -0.1684. Fitting a model just for campaign2 the estimate increases to -0.1112. From the value of these parameter estimates, it is clear that:

a) the relationship between age and response differs between campaigns 1 and 2
b) when combining campaigns 1 and 2, the relationship is dampened

Based on this parameter estimates analysis, the inverse relationship between age and response rate may not hold in the future, since response rate by age group is not stable over time and is affected by increases of mass media advertising supporting investment products. As a result, including age in the predictive model without any treatment, can greatly affect the accuracy of the model when applied as a targeting tool for future direct mail campaigns.

Table 2: PROC LOGISTIC Output

The SAS System

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Label
Intercept	1	-1.6340	0.0265	3805.1640	<.0001		Intercept: DV=1
Age	1	-0.1391	0.00628	490.8242	<.0001	-0.1161	

Why Population Stability Index (PSI) is not appropriate for this issue?

The Population Stability Index (PSI) can help in monitoring data that are used as inputs to predictive models. Modeling data shift for a variety of reasons and this can directly affect the accuracy of predictive models. Various factors can cause data shifts including changes in data sourcing or data definitions, ETL errors etc.

In database marketing, PSI can be utilized to identify changes in the distribution of independent variables and provide an early warning system. However PSI only evaluates changes in the distribution of each independent variable and cannot be used to identify changes in the nature of the relationship between the independent variables and the dependent variable.

As shown in Table 1, the variable age has almost identical distribution between campaigns 1 and 2. As a result, the use of PSI will not assist in identifying the change in the relationship between age and response, which was caused by the mass media advertising campaign, of the retail bank, just before campaign 2.

THE SOLUTION

As shown in the retail banking example above, the variable age appears to be problematic due to the fact that the inverse relationship with response does not hold over time. Depending on its importance or future use, the variable can be treated through recoding or can be excluded from the model development. Treatment of problematic variables extends beyond the scope of this paper. The main purpose of this paper is to introduce an automated process, using PROC LOGISTIC, SAS MACROS and ODS Output to formally evaluate thousands of potential independent variables and identify the ones that show inconsistent behavior in terms of the nature of their relationship with the dependent variable, in logistic regression models.

Whenever multiple historical campaigns or data from different time periods (monthly, quarterly data etc.) are used in a logistic regression model, the modeler can manually examine each independent variable through univariate plots using PROC GPLOT and PROC SGPANEL. The challenge with examining univariate plots of each input variable for patterns lies in the fact that thousands of exploratory variables comprise initial marketing analytics modeling datasets, which exceeds the ability of the analyst to examine patterns within any reasonable time period. This introduces the need for a less subjective and automated way of identifying and flagging independent variables with inconsistent performance across different datasets, marketing campaigns, historical data etc.

To automate the knowledge discovery, we recommend fitting a simple logistic regression model for each independent variable and use the parameter estimate that is part of the annotated output of PROC LOGISTIC to compare independent variables across campaigns, datasets etc. The parameter estimate in logistic regression is a measure of the linear relationship between the independent variable and the log of the odds of the Dependent Variable (DV). If the parameter estimates are significantly different across campaigns, the bi-variate relationship between the independent variable and the dependent variable is not consistent for these campaigns, and therefore the independent variable may need special treatment. The process below identifies these variables by formally testing for significant differences between the parameter estimates.

THE %VAR_CHECK MACRO

The following is a description of each step of the macro that will help us evaluate the consistency of the relationship of each independent variable with the dependent variable (DV), between campaigns 1 and 2. The example can be generalized to more than 2 campaigns, points of time, datasets etc. by minor modifications of the code. For the example below we have two datasets, one for campaign 1 and one for campaign 2.

STEP 1 – Sample the campaign datasets to same counts/response rates

The first step is to sample the campaign 1 and campaign 2 datasets to the same overall counts and response rates. As shown in Table 1, campaign 1 involved 862,340 direct mail pieces and 2,214 responses. To ensure the validity of statistical tests used in step 5, the campaign 1 dataset will be sampled to the same number of total observations and responses, as the campaign 2 dataset. PROC SURVEYSELECT will be used to draw a random sample of both responders and non-responders from the campaign 1 dataset.

Step1 %VAR_CHECK Code

```
/* Create Dataset For Campaign 1 with Equal Number of Observations and Response Rate with Campaign 2 */

proc surveyselect data=campaign1 (where = (DV = 1) method=srs n = 2101
seed = 12345 out= dv1_campaign1;
run;

proc surveyselect data=campaign1 (where = (DV = 0) method=srs n = 646280
seed = 12345 out= dv0_campaign1;
run;

data campaign1;
   set dv1_campaign1
       dv0_campaign1;
run;
```

STEP 2 – PROC LOGISTIC and ODS Output

Next step is to run a simple PROC LOGISTIC for each campaign, using a SAS macro that will cycle through all the independent variables for both campaigns. This helps us uncover the bi-variate relationships between the independent variables and the dependent variable. Using ODS output we can create an output dataset that we can use to extract the parameter estimates and standard errors for each independent variable.

Step2 %VAR_CHECK Code

```
%macro Var_Check (var);

/* Simple Proc Logistic on Campaign 1 */

proc logistic data =campaign1  desc;
model  DV=&var/STB PARMLABEL;
/* ODS Output for Campaign 1*/
ods output parameterestimates= est_campaign1 (keep = Variable Estimate StdErr
                                              rename = (Estimate = Est_campaign1 StdErr=StdErr_campaign1));
run;

/* Simple Proc Logistic on Campaign 2 */

proc logistic data =campaign2  desc;
model  DV=&var/STB PARMLABEL;
/* ODS Output for Campaign 2*/
ods output parameterestimates= est_campaign2 (keep = Variable Estimate StdErr
                                              rename = (Estimate = Est_campaign2 StdErr=StdErr_campaign2));
run;
```

STEP 3 – Merge the two ODS output datasets and create one dataset for each independent variable using Proc SQL

The third step is to merge into one dataset, the campaign 1 and campaign 2 datasets that contain the parameter estimates and standard errors for each variable enabling a comparison of regression coefficients from campaign1 to the relative coefficients from campaign 2.

Step3 %VAR_CHECK Code

```
/* Create one dataset for each independent variable */

data &var._campaign1;
  set est_campaign1;

  if Variable = 'Intercept' then delete;

run;

data &var._campaign2;
  set est_campaign2;

  if Variable = 'Intercept' then delete;

run;

proc sql;
   create table &var._c1_2 as
   select a.Variable as Variable length = 15, a.Est_campaign1, b.Est_campaign2, a.StdErr_campaign1, b.StdErr_campaign2
   from &var._campaign1 as a,
        &var._campaign2 as b
   where a.Variable = b.Variable;
quit;
```

STEP 4 – Use a Standard Normal Test Statistic To Compare Coefficients For Each Independent Variable

We now can formally evaluate each independent variable and determine whether the logistic regression coefficients from campaign 1 are significantly different from the logistic regression coefficients of campaign 2. This tells us if the variable in question exhibits similar characteristics across both campaigns. If these coefficients are found to be significantly different, this suggests that we may need to evaluate the importance of this variable within the model or determine an appropriate method to standardize or transform.

The statistical test to compare logistic regression coefficients across campaigns is defined below:

$$P = \Phi\left(\frac{\beta_1 - \beta_2}{\sqrt{\sigma_1 + \sigma_2}}\right)$$

Where:

P is the P-value associated with the statistical test

Φ is the standard normal CDF

βn is the parameter estimate of the variable in campaign n

σn is the standard error of the coefficient in campaign n

In the code below, we calculate Z scores and their associated p values that are needed for the two-tailed t-test. The t-test will be used to evaluate the following hypothesis:

> Ho: There is no statistically significant difference between the parameter estimate of the independent variable in campaign 1 and the parameter estimate of the independent variable in campaign 2.
>
> Ha: There is a statistically significant difference between the parameter estimate of the independent variable in campaign1 and the parameter estimate of the independent variable in campaign 2

Step4 %VAR_CHECK Code

```
/* Two tailed t-test on campaign1&2 coefficients */

data  &var._c1_2;
  set &var._c1_2;

  Z           = (Abs(Est_campaign2 - Est_campaign1))/Sqrt(StdErr_campaign1**2 + StdErr_campaign2**2);
  pvalue      = 2*(1 - cdf('normal', Z, 0, 1)); /* TWO - TAILED TEST */

  if  0 < Pvalue < 0.1  then Sig_Flag_at_90=1; else if Pvalue="" then Sig_Flag_at_90=.; else Sig_Flag_at_90=0;
  if  0 < Pvalue < 0.05 then Sig_Flag_at_95=1; else if Pvalue="" then Sig_Flag_at_95=.; else Sig_Flag_at_95=0;
  if  0 < Pvalue < 0.01 then Sig_Flag_at_99=1; else if Pvalue="" then Sig_Flag_at_99=.; else Sig_Flag_at_99=0;

run;
```

STEP 5 – Call the macro for all the independent variables in the dataset and output using ODS

The final step creates an output file (SAS dataset, txt file, Excel) with the results of the t-test for each of independent variables. The user also has the option to change the significance level of the test, thereby allowing for changes in the tolerance of the test to detect differences in coefficients.

The code uses the output dataset of PROC CONTENTS and the EXECUTE subroutine, to execute the %VAR_CHECK macro and output a file with the test results. The output file identifies whether the logistic regression coefficients from campaign 1 are significantly different from the coefficients of campaign 2, telling us if the variable in question exhibits similar characteristics across both campaigns.

Step5 %VAR_CHECK Code

```
%mend;


proc contents data=campaign1 out=campaign1_contents noprint; run ;

DATA _NULL_;
  SET  campaign1_contents(where = (name not in
         ('/* Space to add variables not needed in the analysis, IDs, Keys etc.')));
         Var_Check = '%Var_Check(' !! TRIM(LEFT(name))  !!  ');';
    CALL EXECUTE(Var_Check);
run;

proc sql noprint ;
 select cats(name,"_c1_2")
             into :set_datasets
             separated by ' '
             from campaign1_contents(where = (name not in
             ('/* Space to add variables not needed in the analysis, IDs, Keys etc.')));
quit;

data dt.var_check_by_cmpn;
  set &set_datasets. ;
run;

ods html file = "var_check_by_cmpn.xls" style = minimal;

proc print data = var_check_by_cmpn;
run;

ods html close;
```

Table 3: Output Dataset

Variable	Est_Cmpn1	Est_Cmpn2	StdErr_Cmpn1	StdErr_Cmpn2	Est_Dif	Est_Dif_abs	Z	Pvalue	Sig_Flag_90	Sig_Flag_95	Sig_Flag_99
Var1	-0.2768	0.011	0.1276	0.1146	103.96	103.96	1.67795	0.09336	1	0	0
Var2	-0.0381	-0.176	0.0162	0.0758	-362.29	362.29	1.77823	0.07537	1	0	0
Var3	0.00103	0.017	0.00634	0.00426	-1551.38	1551.38	2.09237	0.03641	1	1	0
Var4	-0.309	-1.5125	0.265	0.451	-389.42	389.42	2.30065	0.02141	1	1	0
Var5	0.1581	0.5448	0.1217	0.1191	-244.64	244.64	2.27089	0.02315	1	1	0
Var6	-0.4917	-0.329	0.1227	0.136	33.1	33.1	0.88843	0.37431	0	0	0
Var7	-0.0502	-0.1447	0.1153	0.1118	-188.02	188.02	0.58802	0.55652	0	0	0
Var8	0.3641	0.3891	0.0829	0.0915	-6.88	6.88	0.20278	0.83931	0	0	0
Var9	0.2075	0.0547	0.12	0.1384	73.62	73.62	0.83395	0.40431	0	0	0
Var10	0.0221	0.0315	0.0063	0.00611	-42.22	42.22	1.06381	0.28742	0	0	0

As shown on Table 3, an independent variable (e.g. Var1) can have opposite coefficient signs between multiple input campaigns, datasets, samples etc. With macro %VAR_CHECK, we introduce an automated way to detect this abnormality.

CONCLUSION

The goal of this paper is to illustrate a process that statisticians and marketing analysts can follow to aid in the modeling development phase of a database marketing campaign. The example we provided was from the financial services industry, however, the described process can easily be adapted and applied across a variety of different fields and industries where statistical models are used. With a few modifications in the code, the process can also be used during the development of linear regression models, since the t-test for statistically significant differences between coefficients, can also be applied in linear regression models.

The process described provides users a step-by-step guide to evaluate independent variables relationship with a dependent variable across multiple points in time. Through the use of PROC LOGISTIC, SAS MACROS, and ODS Output we were able to automate this process, allowing users to quickly identify inconsistent variables. The automation and logic of the %Var_Check macro reduces the development time of logistic regression models while increasing their ability to accurately predict future events of interest.

REFERENCES

SAS/STAT Software, Volume 2: the LOGISTIC Procedure, SAS Institute, Inc:

Paper 248-26 GETTING STARTED WITH PROC LOGISTIC Andrew H. Karp

Paper 140-2007 Easy Graphs with PROC FORMAT, PROC GPLOT, and ODS, Apryl DeLancey, Warner Home Video, Burbank, CA

Paper 264-26 Model Fitting in PROC GENMOD, Jean G. Orelien, Analytical Sciences, Inc

Applied Logistic Regression, Wiley: Hosmer and Lemeshow,, 1989

Paper 288-201 The Applied Use of Population Stability Index (PSI) in SAS® Enterprise Miner, Rex Pruitt, PREMIER Bankcard, LLC, Sioux Falls, SD

Paper 039-31 SAS® Macro Dynamics - From Simple Basics to Powerful Invocations Rick Andrews, Centers for Medicare and Medicaid Services, Baltimore, MD

ACKNOWLEDGMENTS

The authors would like to thank Paul D. Berger, Ph.D., of Bentley University, Pat D. Gerard, Ph.D. of Clemson University, Nan Flaaten and Dean Westervelt of Merkle Inc. for their helpful comments and suggestions.

TRADEMARK CITATIONS

Merkle Inc. is the largest privately-held customer relationship marketing agency in the U.S. For more than 20 years, Fortune 1,000 companies and leading non-profit organizations have partnered with Merkle to build and maximize the value of their customer portfolios and return on their marketing investments. With a variety of quantitative, information-based solutions, Merkle works with clients to plan, design, execute and measure fully integrated customer relationship marketing (CRM) solutions.

CONTACT INFORMATION

Alex Vidras is Analytics Manager and David Tysinger is Statistician in the Quantitative Solutions Group (QSG) at Merkle Inc.

Comments and suggestions are valued and encouraged. Contact the authors at:

Alex Vidras
E-mail: avidras@merkleinc.com

David Tysinger
E-mail: dstysing@yahoo.com