**Assignment #6**

Daniel S Prusinski

**Introduction:**

A weakness in regression analysis is the tendency to build models that over-fit the data. Cross validation is a technique that splits the data and allows one to test the regression model on data that has not been associated with building the model. In this assignment, cross-validation will be utilized to assess the best multiple variable logistic regression model. Techniques such as backward selection, assessing goodness-of-fit, lift charts, and the KS test statistic all aid in selecting the best model.

**In-Sample Results:**

Throughout this assignment, two models will be compared. The first model is chosen based on management’s decision and will be called Model 1. The second model, Model 2, is based on a statistical technique that analyzes all the variables in the model and chooses the best variables based on a p-value set at the user’s desire, which is .05 for this exercise. The data being used to formulate the models are comprised of 70% of the total data. The output from running this procedure can be seen below.

| **Summary of Backward Elimination** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Step** | **Effect Removed** | **DF** | **Number In** | **Wald Chi-Square** | **Pr > ChiSq** |
| **1** | **A8** | 1 | 16 | 0.0011 | 0.9733 |
| **2** | **A6\_q** | 1 | 15 | 0.0427 | 0.8364 |
| **3** | **A7\_h** | 1 | 14 | 0.1323 | 0.7161 |
| **4** | **A3** | 1 | 13 | 0.3335 | 0.5636 |
| **5** | **A10\_t** | 1 | 12 | 0.3828 | 0.5361 |
| **6** | **A7\_v** | 1 | 11 | 0.4865 | 0.4855 |
| **7** | **A6\_k** | 1 | 10 | 0.6272 | 0.4284 |
| **8** | **A1\_a** | 1 | 9 | 0.8127 | 0.3673 |
| **9** | **A7\_bb** | 1 | 8 | 1.0080 | 0.3154 |
| **10** | **A6\_w** | 1 | 7 | 1.9597 | 0.1616 |
| **11** | **A12\_t** | 1 | 6 | 1.6777 | 0.1952 |
| **12** | **A2** | 1 | 5 | 1.9496 | 0.1626 |

| **Analysis of Maximum Likelihood Estimates**  **Variables that are staying the model.** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -4.0846 | 0.5365 | 57.9692 | <.0001 |
| **A11** | 1 | 0.2291 | 0.0680 | 11.3368 | 0.0008 |
| **A15** | 1 | 0.000597 | 0.000224 | 7.1232 | 0.0076 |
| **A4\_u** | 1 | 0.8924 | 0.4054 | 4.8450 | 0.0277 |
| **A7\_ff** | 1 | -2.1065 | 0.9038 | 5.4325 | 0.0198 |
| **A9\_t** | 1 | 4.0210 | 0.4378 | 84.3549 | <.0001 |

From the results above, it can be seen that five variables had a p-value less than .05, and as a result will be the variables in the second model. From the past EDA in assignment five, I chose A11 as my optimal model and seeing that A11 and A9\_t are both in this model my conclusion is that this model will be very strong.

The next step in assessing the two models is to compare the inferential statistics between the two models. Below is the “Model Fit Statistic”.

| **Model Fit Statistics Model 1** | | |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 620.703 | 340.739 |
| **SC** | 624.812 | 357.176 |
| **-2 Log L** | 618.703 | 332.739 |

| **Model Fit Statistics Model 2** | | |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 620.703 | 289.616 |
| **SC** | 624.812 | 314.271 |
| **-2 Log L** | 618.703 | 277.616 |

This output compares how well the models fit the data. A high -2 Log L value equates to a worse fit. It is assumed that Model 2 will fit better given the fact it has higher covariates, but the AIC and SC penalize a model for having more covariates. Interestingly, Model 2 has a lower value for each criterion than Model 1. The Global Null hypothesis tests that all the explanatory variables have coefficients equal to zero. It can be seen that both variables have at least one coefficient that does not equal zero. Both models also have a significant p-value. Model 2 has much higher scores, and its coefficients are likely to be more form fitting on the data. This could lead to over-fitting which will be analyzed later.

| **Model 1 Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 285.9640 | 3 | <.0001 |
| **Score** | 246.5494 | 3 | <.0001 |
| **Wald** | 151.7473 | 3 | <.0001 |

| **Model 2 Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 341.0870 | 5 | <.0001 |
| **Score** | 267.3283 | 5 | <.0001 |
| **Wald** | 131.5481 | 5 | <.0001 |

The maximum likelihood analysis pared with the odds ratio estimates reveal statistically significant individual coefficients and their prospective magnitudes. In Model 1, A2 and A3 lack the statistical significance at the .05 threshold, and I would alert this point to management. Model 2 only has one variable that is not statistically significant, but it is rather close to the .05 threshold. In order to better understand the magnitude of the coefficients, interpreting the odds ratio is helpful. The odds ratio of a coefficient communicates that the predicted odds for that coefficient are the Point Estimate times the odds compared to that specific non-coefficient. For example, A9\_t has 53 times the odds of non A9\_t values of being 1. The magnitude for A9\_t in model 1 is huge compared to Model 2. Also, A15 almost has a non-existent coefficient.

| **Model 1 Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -3.6287 | 0.5051 | 51.6051 | <.0001 |
| **A9\_t** | 1 | 3.9836 | 0.3302 | 145.5842 | <.0001 |
| **A2** | 1 | 0.0227 | 0.0127 | 3.1641 | 0.0753 |
| **A3** | 1 | 0.0527 | 0.0314 | 2.8241 | 0.0929 |

| **Model 2 Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -3.5542 | 0.4514 | 61.9895 | <.0001 |
| **A11** | 1 | 0.2229 | 0.0607 | 13.5025 | 0.0002 |
| **A15** | 1 | 0.000555 | 0.000207 | 7.1670 | 0.0074 |
| **A4\_u** | 1 | 0.6854 | 0.3706 | 3.4200 | 0.0644 |
| **A7\_ff** | 1 | -2.1243 | 0.8686 | 5.9816 | 0.0145 |
| **A9\_t** | 1 | 3.6107 | 0.3630 | 98.9523 | <.0001 |

| **Model 1 Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **A9\_t** | 53.712 | 28.122 | 102.590 |
| **A2** | 1.023 | 0.998 | 1.049 |
| **A3** | 1.054 | 0.991 | 1.121 |

| **Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **A11** | 1.250 | 1.110 | 1.407 |
| **A15** | 1.001 | 1.000 | 1.001 |
| **A4\_u** | 1.985 | 0.960 | 4.103 |
| **A7\_ff** | 0.120 | 0.022 | 0.656 |
| **A9\_t** | 36.992 | 18.161 | 75.349 |

The goodness-of-fit statistics include the percent concordant, percent discordant, Somer’s D, Gamma, and Tau-a. The output for these statistics are listed below.

| **Model 1 Association of Predicted Probabilities and Observed Responses** | | | |
| --- | --- | --- | --- |
| **Percent Concordant** | 89.1 | **Somers' D** | 0.787 |
| **Percent Discordant** | 10.5 | **Gamma** | 0.790 |
| **Percent Tied** | 0.4 | **Tau-a** | 0.390 |
| **Pairs** | 50049 | **c** | 0.893 |

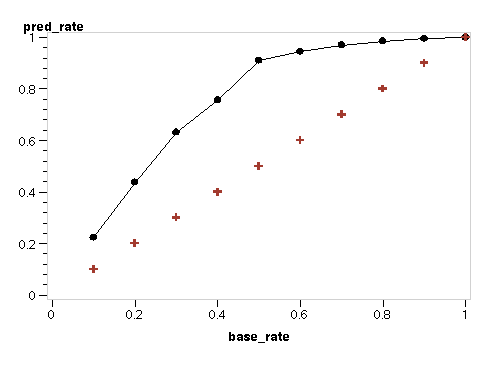
| **Model 2 Association of Predicted Probabilities and Observed Responses** | | | |
| --- | --- | --- | --- |
| **Percent Concordant** | 92.6 | **Somers' D** | 0.871 |
| **Percent Discordant** | 5.5 | **Gamma** | 0.888 |
| **Percent Tied** | 1.9 | **Tau-a** | 0.432 |
| **Pairs** | 50049 | **c** | 0.936 |

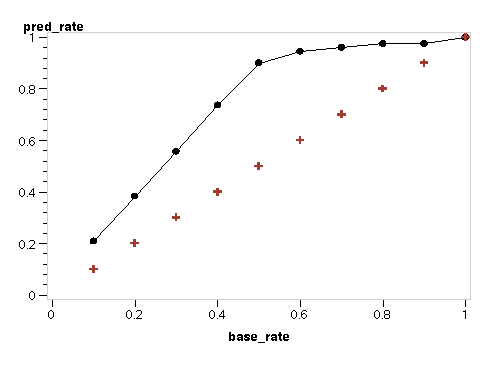
Both models have high percent concordant values. I look forward to analyzing this information on the test data. Model 2 is slightly better, and this comes as no surprise based on the prior analysis.

For Model 1, the lift model reflects what is seen in the lift chart. When targeting 50% of the population, the lift is around 40%. For Model 2, the results are very similar, except the lift is 1 percent greater.

| **Model 2** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs** | **score\_decile** | **Y\_Sum** | **Nobs** | **cum\_obs** | **model\_pred** | **pred\_rate** | **base\_rate** | **life** |
| **1** | 1 | 45 | 45 | 45 | 45 | 0.22388 | 0.1 | 0.12388 |
| **2** | 2 | 43 | 45 | 90 | 88 | 0.43781 | 0.2 | 0.23781 |
| **3** | 3 | 39 | 45 | 135 | 127 | 0.63184 | 0.3 | 0.33184 |
| **4** | 4 | 25 | 31 | 166 | 152 | 0.75622 | 0.4 | 0.35622 |
| **5** | 5 | 31 | 56 | 222 | 183 | 0.91045 | 0.5 | 0.41045 |
| **6** | 6 | 7 | 48 | 270 | 190 | 0.94527 | 0.6 | 0.34527 |
| **7** | 7 | 5 | 44 | 314 | 195 | 0.97015 | 0.7 | 0.27015 |
| **8** | 8 | 3 | 60 | 374 | 198 | 0.98507 | 0.8 | 0.18507 |
| **9** | 9 | 2 | 25 | 399 | 200 | 0.99502 | 0.9 | 0.09502 |
| **10** | 10 | 1 | 51 | 450 | 201 | 1.00000 | 1.0 | 0.00000 |

| **Model 1** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs** | **score\_decile** | **Y\_Sum** | **Nobs** | **cum\_obs** | **model\_pred** | **pred\_rate** | **base\_rate** | **life** |
| **1** | 1 | 42 | 45 | 45 | 42 | 0.20896 | 0.1 | 0.10896 |
| **2** | 2 | 35 | 45 | 90 | 77 | 0.38308 | 0.2 | 0.18308 |
| **3** | 3 | 35 | 45 | 135 | 112 | 0.55721 | 0.3 | 0.25721 |
| **4** | 4 | 36 | 45 | 180 | 148 | 0.73632 | 0.4 | 0.33632 |
| **5** | 5 | 33 | 45 | 225 | 181 | 0.90050 | 0.5 | 0.40050 |
| **6** | 6 | 9 | 45 | 270 | 190 | 0.94527 | 0.6 | 0.34527 |
| **7** | 7 | 3 | 45 | 315 | 193 | 0.96020 | 0.7 | 0.26020 |
| **8** | 8 | 3 | 45 | 360 | 196 | 0.97512 | 0.8 | 0.17512 |
| **9** | 9 | 0 | 45 | 405 | 196 | 0.97512 | 0.9 | 0.07512 |
| **10** | 10 | 5 | 45 | 450 | 201 | 1.00000 | 1.0 | 0.00000 |

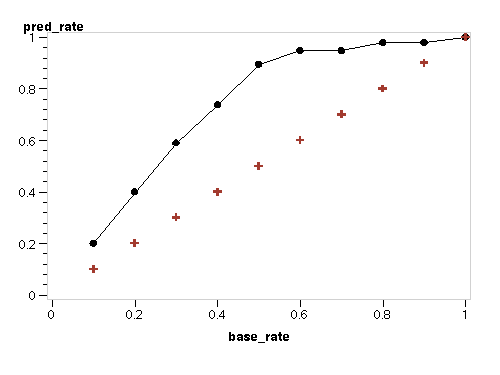


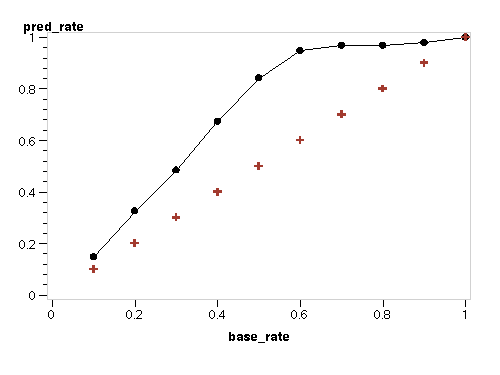


Interpreting the lift table and lift chart is key to understanding how the models perform on data that were not used to make the data. The red crosses on the graph represent a random guess, which lies at the 45 degree angle mark. For both models, the first five decile’s are strongly predictive. Model 2 has stronger prediction. In my opinion, both models peak around 50% and have an optimum added lift of 39% percent for Model 2, and 34% for Model 1. Both models perform very similarly to their “in-sample” models, which leads me to believe that neither models are over fit. The Kolmogorov-Smirnov (KS) test is the same as the lift for the models. But, the importance of the KS test lies in its statistical validation between the two models. While I can see that the distributions are different, I need to statistically verify that they are different. For model 2, I multiplied the lift times the square root of (23\*15/ (23+15) which equals 1.517 and rejects the null hypothesis that the distributions are the same. I would recommend to management to use Model 2, but I would want to know more about Model 1 and the significance of the variables. Perhaps after better understanding the variables, I would make an additional model with variables from both models.

| **Model 2** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs** | **score\_decile** | **Y\_Sum** | **Nobs** | **cum\_obs** | **model\_pred** | **pred\_rate** | **base\_rate** | **Lift** |
| **1** | 1 | 19 | 20 | 20 | 19 | 0.20000 | 0.1 | 0.10000 |
| **2** | 2 | 19 | 20 | 40 | 38 | 0.40000 | 0.2 | 0.20000 |
| **3** | 3 | 18 | 21 | 61 | 56 | 0.58947 | 0.3 | 0.28947 |
| **4** | 4 | 14 | 17 | 78 | 70 | 0.73684 | 0.4 | 0.33684 |
| **5** | 5 | 15 | 23 | 101 | 85 | 0.89474 | 0.5 | 0.39474 |
| **6** | 6 | 5 | 21 | 122 | 90 | 0.94737 | 0.6 | 0.34737 |
| **7** | 7 | 0 | 16 | 138 | 90 | 0.94737 | 0.7 | 0.24737 |
| **8** | 8 | 3 | 32 | 170 | 93 | 0.97895 | 0.8 | 0.17895 |
| **9** | 9 | 0 | 12 | 182 | 93 | 0.97895 | 0.9 | 0.07895 |
| **10** | 10 | 2 | 21 | 203 | 95 | 1.00000 | 1.0 | 0.00000 |

| **Model 1** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs** | **score\_decile** | **Y\_Sum** | **Nobs** | **cum\_obs** | **model\_pred** | **pred\_rate** | **base\_rate** | **Lift** |
| **1** | 1 | 14 | 20 | 20 | 14 | 0.14737 | 0.1 | 0.04737 |
| **2** | 2 | 17 | 20 | 40 | 31 | 0.32632 | 0.2 | 0.12632 |
| **3** | 3 | 15 | 21 | 61 | 46 | 0.48421 | 0.3 | 0.18421 |
| **4** | 4 | 18 | 20 | 81 | 64 | 0.67368 | 0.4 | 0.27368 |
| **5** | 5 | 16 | 20 | 101 | 80 | 0.84211 | 0.5 | 0.34211 |
| **6** | 6 | 10 | 21 | 122 | 90 | 0.94737 | 0.6 | 0.34737 |
| **7** | 7 | 2 | 20 | 142 | 92 | 0.96842 | 0.7 | 0.26842 |
| **8** | 8 | 0 | 21 | 163 | 92 | 0.96842 | 0.8 | 0.16842 |
| **9** | 9 | 1 | 20 | 183 | 93 | 0.97895 | 0.9 | 0.07895 |
| **10** | 10 | 2 | 20 | 203 | 95 | 1.00000 | 1.0 | 0.00000 |





**Conclusion:**

This assignment demonstrated how to split data and utilize cross-validation as a technique to hone the predictive modeling process. The code for this assignment was the most complex to date, and while interpreting the results I felt underwater. The new techniques learned in this assignment are very applicable, but I need much more practice before I remotely feel competent.

SAS Code:

\*Daniel Prusinski Assignment 6 Version 1\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*Lift Chart For Training A11 A15 A4\_u A7\_ff A9\_t\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\*\*\*\*Statement to access where the data is stored\*\*\*\*\*;

**libname** mydata '/courses/u\_northwestern.edu1/i\_833463/c\_3505/SAS\_Data/'**;**

ods graphics on**;**

\*\*\*\*\*This creates the response variable\*\*\*\*\*;

**data** temp**;**

set mydata.credit\_approval**;**

u=uniform**(123);**

if **(**u<**0.7)** then train=**1;** else train=**0;**

if **(**A16='+'**)** then Y =**1;**

else Y=**0;**

if **(**train=**1)** then Y\_train=Y**;** else Y\_train=**.;**

\*\*\*\*\*Categorical Variables\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

if **(**A1='a'**)** then A1\_a=**1;** else A1\_a=**0;**

if **(**A4='u'**)** then A4\_u=**1;** else A4\_u=**0;**

if **(**A5='g'**)** then A5\_g=**1;** else A5\_g=**0;**

if **(**A6='aa'**)** then A6\_aa=**1;** else A6\_aa=**0;**

if **(**A6='c'**)** then A6\_c=**1;** else A6\_c=**0;**

if **(**A6='cc'**)** then A6\_cc=**1;** else A6\_cc=**0;**

if **(**A6='d'**)** then A6\_d=**1;** else A6\_d=**0;**

if **(**A6='e'**)** then A6\_e=**1;** else A6\_e=**0;**

if **(**A6='ff'**)** then A6\_ff=**1;** else A6\_ff=**0;**

if **(**A6='i'**)** then A6\_i=**1;** else A6\_i=**0;**

if **(**A6='j'**)** then A6\_j=**1;** else A6\_j=**0;**

if **(**A6='k'**)** then A6\_k=**1;** else A6\_k=**0;**

if **(**A6='m'**)** then A6\_m=**1;** else A6\_m=**0;**

if **(**A6='q'**)** then A6\_q=**1;** else A6\_q=**0;**

if **(**A6='r'**)** then A6\_r=**1;** else A6\_r=**0;**

if **(**A6='w'**)** then A6\_w=**1;** else A6\_w=**0;**

\*\*\*\*\*I left off a few of the small variables, I want to see what this does\*\*\*\*\*;

if **(**A7='bb'**)** then A7\_bb=**1;** else A7\_bb=**0;**

if **(**A7='ff'**)** then A7\_ff=**1;** else A7\_ff=**0;**

if **(**A7='h'**)** then A7\_h=**1;** else A7\_h=**0;**

if **(**A7='v'**)** then A7\_v=**1;** else A7\_v=**0;**

if **(**A9='t'**)** then A9\_t=**1;** else A9\_t=**0;**

if **(**A10='t'**)** then A10\_t=**1;** else A10\_t=**0;**

if **(**A12='t'**)** then A12\_t=**1;** else A12\_t=**0;**

if **(**A13='g'**)** then A13\_g=**1;** else A13=\_g=**0;**

\*\*\*\*\*This purges the Data, 90 LSB\*\*\*\*\*;

if A1 = '?' then delete**;**

else if A2 = '.' then delete**;**

else if A3 = '.' then delete**;**

else if A4 = '?' then delete**;**

else if A5 = '?' then delete**;**

else if A6 = '?' then delete**;**

else if A7 = '?' then delete**;**

else if A8 = '.' then delete**;**

else if A9 = '?' then delete**;**

else if A10 = '?' then delete**;**

else if A11 = '.' then delete**;**

else if A12 = '?' then delete**;**

else if A13 = '?' then delete**;**

else if A14 = '.' then delete**;**

else if A15 = '.' then delete**;**

**run;**

**proc** **logistic** **data**=temp descending**;**

model Y\_train = A2 A3 A8 A11 A15

A1\_a A4\_u A5\_g A6\_k A6\_q A6\_w A7\_bb A7\_ff A7\_h A7\_v

A9\_t A10\_t A12\_t A13\_g / selection=backward**;**

output out=model\_data pred=yhat**;**

**run;**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

This is the beginning of building the

lift chart for A11 A15 A4\_u A7\_ff A9\_t\*\*\*\*\*\*\*\*\*;

**proc** **logistic** **data**=temp descending**;**

model Y\_train = A11 A15 A4\_u A7\_ff A9\_t**;**

output out=model\_data2 pred=yhat**;**

**run;**

**proc** **npar1way** date=temp**;**

class Y**;**

var A11, A15**;**

**run**

**proc** **rank** **data**=model\_data2

out=training\_scores descending groups=**10;**

var yhat**;**

ranks score\_decile**;**

where train=**1;**

**run;**

\*\*\*\*\*This creates the lift chart\*\*\*\*\*;

**proc** **means** **data**=training\_scores sum**;**

class score\_decile**;**

var Y**;**

output out=pm\_out sum**(**Y**)**=Y\_Sum**;**

**run;**

**proc** **print** **data**=pm\_out**;**

**run;**

**data** lift\_chart**;**

set pm\_out **(**where=**(**\_type\_=**1));**

by \_type\_**;**

Nobs=\_freq\_**;**

score\_decile = score\_decile+**1;**

if first.\_type\_ then do**;**

cum\_obs=Nobs**;**

model\_pred=Y\_Sum**;**

end**;**

else do**;**

cum\_obs=cum\_obs+Nobs**;**

model\_pred=model\_pred+Y\_Sum**;**

end**;**

retain cum\_obs model\_pred**;**

\*\*\*\*\* 201 represents the number of successes

This value will need to be changed with different samples\*\*\*\*\*;

pred\_rate=model\_pred/**201;**

base\_rate=score\_decile\***0.1;**

lift = pred\_rate-base\_rate**;**

drop \_freq\_ \_type\_**;**

**run;**

**proc** **print** **data**=lift\_chart**;**

**run;**

ods graphics on**;**

title 'In-Sample Lift Chart'**;**

symbol1 color=red interprol=join value=dot height=**1;**

symbol2 color=black interpol=join value=dot height=**1;**

**proc** **gplot** **data**=lift\_chart**;**

plot pred\_rate\*base\_rate base\_rate\*base\_rate / overlay**;**

**run;** **quit;**

ods graphics off**;**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

This is the beginning of building the

lift chart for A11 A15 A4\_u A7\_ff A9\_t\*\*\*\*\*\*\*\*\*;

**proc** **logistic** **data**=temp descending**;**

model Y\_train = A11 A15 A4\_u A7\_ff A9\_t**;**

output out=model\_data pred=yhat**;**

**run;**

**proc** **rank** **data**=model\_data

out=training\_scores descending groups=**10;**

var yhat**;**

ranks score\_decile**;**

where train=**0;**

**run;**

\*\*\*\*\*This creates the lift chart\*\*\*\*\*;

**proc** **means** **data**=training\_scores sum**;**

class score\_decile**;**

var Y**;**

output out=pm\_out sum**(**Y**)**=Y\_Sum**;**

**run;**

**proc** **print** **data**=pm\_out**;**

**run;**

**data** lift\_chart**;**

set pm\_out **(**where=**(**\_type\_=**1));**

by \_type\_**;**

Nobs=\_freq\_**;**

score\_decile = score\_decile+**1;**

if first.\_type\_ then do**;**

cum\_obs=Nobs**;**

model\_pred=Y\_Sum**;**

end**;**

else do**;**

cum\_obs=cum\_obs+Nobs**;**

model\_pred=model\_pred+Y\_Sum**;**

end**;**

retain cum\_obs model\_pred**;**

\*\*\*\*\* 201 represents the number of successes

This value will need to be changed with different samples\*\*\*\*\*;

pred\_rate=model\_pred/**95;**

base\_rate=score\_decile\***0.1;**

life = pred\_rate-base\_rate**;**

drop \_freq\_ \_type\_**;**

**run;**

**proc** **print** **data**=lift\_chart**;**

**run;**

ods graphics on**;**

title 'In-Sample Lift Chart'**;**

symbol1 color=red interprol=join value=dot height=**1;**

symbol2 color=black interpol=join value=dot height=**1;**

**proc** **gplot** **data**=lift\_chart**;**

plot pred\_rate\*base\_rate base\_rate\*base\_rate / overlay**;**

**run;** **quit;**

ods graphics off**;**

**proc** **logistic** **data**=temp descending**;**

model Y\_train = A9\_t A2 A3**;**

output out=model\_data2 pred=yhat**;**

**run;**

**proc** **rank** **data**=model\_data2

out=training\_scores descending groups=**10;**

var yhat**;**

ranks score\_decile**;**

where train=**1;**

**run;**

\*\*\*\*\*This creates the lift chart\*\*\*\*\*;

**proc** **means** **data**=training\_scores sum**;**

class score\_decile**;**

var Y**;**

output out=pm\_out sum**(**Y**)**=Y\_Sum**;**

**run;**

**proc** **print** **data**=pm\_out**;**

**run;**

**data** lift\_chart**;**

set pm\_out **(**where=**(**\_type\_=**1));**

by \_type\_**;**

Nobs=\_freq\_**;**

score\_decile = score\_decile+**1;**

if first.\_type\_ then do**;**

cum\_obs=Nobs**;**

model\_pred=Y\_Sum**;**

end**;**

else do**;**

cum\_obs=cum\_obs+Nobs**;**

model\_pred=model\_pred+Y\_Sum**;**

end**;**

retain cum\_obs model\_pred**;**

\*\*\*\*\* 201 represents the number of successes

This value will need to be changed with different samples\*\*\*\*\*;

pred\_rate=model\_pred/**201;**

base\_rate=score\_decile\***0.1;**

life = pred\_rate-base\_rate**;**

drop \_freq\_ \_type\_**;**

**run;**

**proc** **print** **data**=lift\_chart**;**

**run;**

ods graphics on**;**

title 'In-Sample Lift Chart'**;**

symbol1 color=red interprol=join value=dot height=**1;**

symbol2 color=black interpol=join value=dot height=**1;**

**proc** **gplot** **data**=lift\_chart**;**

plot pred\_rate\*base\_rate base\_rate\*base\_rate / overlay**;**

**run;** **quit;**

ods graphics off**;**

**proc** **rank** **data**=model\_data2

out=testing\_scores descending groups=**10;**

var yhat**;**

ranks score\_decile**;**

where train=**0;**

**run;**

\*\*\*\*\*This creates the lift chart\*\*\*\*\*;

**proc** **means** **data**=testing\_scores sum**;**

class score\_decile**;**

var Y**;**

output out=pm\_out sum**(**Y**)**=Y\_Sum**;**

**run;**

**proc** **print** **data**=pm\_out**;**

**run;**

**data** lift\_chart**;**

set pm\_out **(**where=**(**\_type\_=**1));**

by \_type\_**;**

Nobs=\_freq\_**;**

score\_decile = score\_decile+**1;**

if first.\_type\_ then do**;**

cum\_obs=Nobs**;**

model\_pred=Y\_Sum**;**

end**;**

else do**;**

cum\_obs=cum\_obs+Nobs**;**

model\_pred=model\_pred+Y\_Sum**;**

end**;**

retain cum\_obs model\_pred**;**

\*\*\*\*\* 201 represents the number of successes

This value will need to be changed with different samples\*\*\*\*\*;

pred\_rate=model\_pred/**95;**

base\_rate=score\_decile\***0.1;**

life = pred\_rate-base\_rate**;**

drop \_freq\_ \_type\_**;**

**run;**

**proc** **print** **data**=lift\_chart**;**

**run;**

ods graphics on**;**

title 'Out-Of-Sample Lift Chart'**;**

symbol1 color=red interprol=join value=dot height=**1;**

symbol2 color=black interpol=join value=dot height=**1;**

**proc** **gplot** **data**=lift\_chart**;**

plot pred\_rate\*base\_rate base\_rate\*base\_rate / overlay**;**

**run;** **quit;**