Assignment #6: Multiple Logistic Regression Model (40 points)

Data Directory: Data can be accessed on the SAS OnDemand server using this libname statement.

```
libname mydata        '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/' access=readonly;
```

Data Set: mydata.credit_approval

Data Description: See the data dictionary for a full description of the data.

Assignment Instructions:

For this assignment we will fit a multiple logistic regression model for a binary response variable to the credit_approval data set using PROC LOGISTIC, and assess its predictive accuracy. We will compare the predictive performance of our multiple logistic regression model to the predictive performance of a pre-specified model. **You should use the same data formatting and dummy variables that you developed in Assignment #5.**

Split the Sample: In order to assess the predictive accuracy of this classification model we will employ a statistical methodology called *cross-validation* by splitting the sample data into a 70/30 split, which we will respectively refer to as the *training* and *testing* samples. Split the data by generating a uniform random variable using the statement u=uniform(123); in a SAS data step. If (u<0.7) then assign the observation to the training data set, else assign the observation to the testing data set. We will estimate the model on the training data set and assess its predictive accuracy on the testing data set.

```
data temp;
      set temp;
      * Flag the observations as training/testing;
      * Since we set the seed value to 123, we will get the same set of
      random numbers every time and we will all get the same set of random
      numbers;
      u=uniform(123);
      if (u<0.7) then train=1; else train=0;

      * Create a response indicator based on the training/testing split;
      if (train=1) then Y_train=Y; else Y_train=.;

      /* DEFINE ALL OF YOUR DUMMY VARIABLES HERE */

      * Delete the observations with missing values;
      if (A1='?') or (A4='?') or (A5='?') or (A6='?') or (A7='?')
      or (A2=.) or (A3=.) or (A8=.) or (A11=.) or (A14=.) or (A15=.)
      then delete;
run;
```

In order for all of us to get the same answer we must follow the outline of the above data step.

<u>Fit the Model:</u> Find the optimal model using backward variable selection. You will need to include an output statement in PROC LOGISTIC in order to output the model scores, i.e. the probability that Y=1.

```
proc logistic data=temp descending;
model Y_train = A2 A3 A8 A11 A14 A15
      A1_b A4_u A5_g
      A6_aa A6_c A6_cc A6_ff A6_i A6_k A6_m A6_q A6_w A6_x
      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;
```

We will refer to the model selected through this backward variable selection procedure as Model #1. Your report should include the backward selection summary table, the parameter estimates, the goodness-of-fit statistics, and a discussion of these results .

In addition to the optimal model that you will define, your manager wants you to fit this particular model. We will refer to this model as Model #2.

```
proc logistic data=temp descending;
model Y_train = A9_t A2 A3 ;
output out=model_data2 pred=yhat;
run;
```

We want to compare the predictive merits of the two models and suggest that one model be used instead of the other model. We will begin this analysis by discussing the in-sample (training sample) model fit including the parameter estimates and the goodness-of-fit statistics automatically produced by SAS.

<u>Assessing the Predictive Accuracy:</u> While the goodness-of-fit statistics provide an insight into the in-sample predictive accuracy of the fitted model. We must always assess the out-of-sample predictive accuracy of our model in order to guard against *overfitting*, hence the need for some type of *cross-validation*. We will assess the predictive accuracy of our model by creating a *lift chart* (also known as a *cumulative gains chart*) and computing the respective Kolmogorov-Smirnov test statistic as a model comparison measure. Here is how we compute the lift chart for Model #2.

```
proc logistic data=temp descending;
model Y_train = A9_t A2 A3 ;
output out=model_data2 pred=yhat;
run;

* The descending option assigns the highest model scores to the lowest
score_decile;
proc rank data=model_data2 out=training_scores descending groups=10;
var yhat;
ranks score_decile;
where train=1;
run;
```

```
* To create the lift chart run this exact code;
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 interpol=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;
```

You can find the scaling factor 201 by using a PROC FREQ statement. 201 is the scaling factor for the in-sample lift chart. The out-of-sample lift chart will have a different scaling factor.

```
proc freq data=temp;
tables train*Y;
run;
```

You will produce the lift chart (table) and a plot of the lift chart for both models and for both the training and testing data sets and display all four tables and graphs in your report. Note that the model lift is defined by the difference between the model classification rate and random assignment. The model lift is also the Kolmogorov-Smirnov test statistic. For more information on lift charts see:

```
(1) http://msdn.microsoft.com/en-us/library/ms175428.aspx
(2) Chapter 8 & 18 of Statistical Machine-Learning Data Mining by Ratner
```

Assignment Document:

All assignment reports should conform to the standards and style of the report template provided to you. Results should be presented and discussed in an organized manner with the discussion in close proximity of the results. The report should not contain unnecessary results or information.

The 'Results' section of this report should contain two subsections: 'In-Sample Results' and 'Out-of-Sample Results'. The 'In-Sample Results' section will contain the two fitted logistic regression models with their model parameter values, the output of a model selection procedure, a lift chart table and graph for each model for the training data set, and a discussion of the models and their goodness-of-fit statistics. The 'Out-of-Sample Results' section should contain a lift chart table and graph for each model for the testing data set with a discussion of their predictive accuracy and a recommendation for one model over the other model. The document should be submitted in pdf format.