Assignment #5: Binary Response Exploratory Data Analysis and a Single Variable Logistic Regression Model (30 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 perform an Exploratory Data Analysis (EDA) for a binary response variable and fit a single variable logistic regression model to the credit_approval data set using PROC LOGISTIC.

Part 1: The Exploratory Data Analysis: First, define the response variable Y=1 if A16='+' and Y=0 if A16='-'. The response variable will need to be defined in a SAS Data Step where you define a SAS data set named 'temp' (see the IF-THEN/ELSE statement pp. 88-89 in *The Little SAS Book*). Second, prepare the data to perform an EDA by examining the distributions of attribute variables for each class. **All of the data coding (the dummy variables and the discretization of the continuous variables) outlined below should be performed in a single SAS data step.**

(1) The first step of the EDA is to look at the conditional distributions of the predictor variables using PROC MEANS with a CLASS statement for the continuous predictor variables and PROC FREQ for the categorical predictor variables. This step serves as a data check so that you can view the possible values taken by categorical variables and the quantiles of the continuous variables.

```
proc freq data=temp;
tables A1 A4 A5 A6 A7 A9 A10 A12 A13 A16;
run;

proc means data=temp p5 p10 p25 p50 p75 p90 p95;
class Y;
var A2 A3 A8 A11 A14 A15;
run;
```

(2) Use the summary statistics from PROC MEANS to discretize the continuous predictor variables using an IF-THEN/ELSE-IF/ELSE ladder for each attribute.

Analysis Variable : A15								
Y	N Obs	5th Pctl	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl	95th Pctl
0	357	0	0	0	1.0000000	67.0000000	400.0000000	1000.00
1	296	0	0	0	210.5000000	1223.00	4159.00	8000.00

For example using the conditional distributions for A15 we could construct the following discrete categories.

```
if (A15 < 1.5) then A15_discrete=1;
else if (A15 < 50) then A15_discrete=2;
else if (A15 < 100) then A15_discrete=3;
else if (A15 < 200) then A15_discrete=4;
else if (A15 < 4000) then A15_discrete=5;
else A15_discrete=6;
```

Notice how the observations with (Y=0) tend to have much smaller values for A15 than the observations with (Y=1). By discretizing A15 we hope construct classes that are unbalanced in Y, i.e. each class will have more (Y=0) or (Y=1). We will see this imbalance in Step 5 when we compute the mean for each category, but you can also see the discrete distribution by producing a cross-tabulated table using PROC FREQ.

```
proc freq data=temp;
tables Y*A15_discrete;
run; quit;
```

Table of Y by A15_discrete							
Y	A15_discrete						
Frequency Percent Row Pct Col Pct	1	2	3	4	5	6	Total
0	192 29.40 53.78 63.58	68 10.41 19.05 79.07	16 2.45 4.48 76.19	21 3.22 5.88 63.64	55 8.42 15.41 31.61	5 0.77 1.40 13.51	357 54.67
1	110 16.85 37.16 36.42	18 2.76 6.08 20.93	5 0.77 1.69 23.81	12 1.84 4.05 36.36	119 18.22 40.20 68.39	32 4.90 10.81 86.49	296 45.33
Total	302 46.25	86 13.17	21 3.22	33 5.05	174 26.65	37 5.67	653 100.00

In general discretization allows us to capture a non-linear relationship, but in the context of logistic regression discetizing a continuous variables allows us to approach logistic regression in the mindset of a *contingency table*, for example see Table 3.2 on p. 51 of *Applied Logistic Regression*. In a regression setting discretized variables capture the non-linear relationship by acting like *smoothers*, e.g. LOESS, by computing class means. You can read more about specifying a logistic regression model in Section 4.2 (p. 92) of *Applied Logistic Regression.*

For this assignment you can select any cut-points that you wish, but you must create at least four categories for each continuous variable. Selecting the cut-points is an iterative process! If your initial cut-points do not create an imbalanced distribution for each category, then simply revise them. If you cannot find any cut-points that do create an imbalanced distribution across the categories, then the variable will not be a good predictor variable.

(3) Code each categorical attribute using a family of dummy variables (see *SAS Statistics By Example* pp. 153-155 and recall that a categorical variable with k categories requires (k-1) dummy variables). **If an attribute has fifteen categories, then you need to explicitly code fourteen dummy variables.** The dummy variables should be constructed in the following format.

```
if (A1='a') then A1_a=1; else A1_a=0;
```

(4) Write a single IF statement at the bottom of your data step to purge any missing values in any of the fifteen attributes. (Hint: What values do the missing categorical variables take and what values do the missing continuous variables take?)

(5) Use a PROC MEANS statement with a class statement to assess the predictive accuracy of an attribute. Here is a macro function to help you perform the task. Note that the input variable needs to be a categorical variable, hence this macro will also work for your discretized continuous variables.

```
%macro class_mean(c);
proc means data=temp mean;
*class A1 A4 A5 A6 A7 A9 A10 A12 A13;
class &c. ;
var Y;
run;
%mend class_mean;
```

Here is the output from running `%class_mean(c=A15_discrete);` Do these numbers look familiar? (Hint: Compare these numbers to the cross-tabulated table above produced by PROC FREQ.)

Analysis Variable : Y		
A15_discrete	N Obs	Mean
1	302	0.3642384
2	86	0.2093023
3	21	0.2380952
4	33	0.3636364
5	174	0.6839080
6	37	0.8648649

The results of your EDA should we well organized into your report with a discussion of the results and their interpretation. This discussion should cover the entire EDA.

Part 2: Model Building: First, select the single variable logistic regression model of your choice based on the EDA that you performed and fit the model. Second, find the best single variable logistic regression model using the *selection=score* option in PROC LOGISTIC with *start=1* and *stop=1*. **For this model you will include the continuous attributes as continuous predictor variables and the categorical attributes using your dummy variables.** Note that you will have to use the *descending* option in PROC LOGISTIC, see Chapter 11 of *SAS Statistics By Example*. In your discussion be sure to answer the following discussion questions.

(1) Did you select the optimal regression model using EDA?
(2) How do we interpret the estimated coefficient for a dummy variable?
(3) What does it mean when a dummy variable is dropped from the model?

In your report be sure to show the summary table from the model selection procedure, the parameter estimates, and the goodness-of –fit statistics. Your discussion of the results should include a thorough discussion of the following two questions. In logistic regression what constitutes an assessment of the model adequacy? How do we interpret the percent concordant mean, Somer's D, Gamma, and Tau-a statistics produced by SAS (see pp. 68-72 in *Logistic Regression Using SAS*)?

Part 3: Model Assessment Using the ROC Curve: First, produce a ROC curve with cut-points and the corresponding data set containing the coordinates on the ROC for those cut-points for the optimal model in a PROC LOGISTIC statement.

```
* Make ROC curve with cut-points;
ods graphics on;
proc logistic data=temp descending plots(only)=roc(id=prob);
model Y = A9_t / outroc=roc1;
run;
ods graphics off;

proc print data=roc1; run;quit;
```

How do we interpret these cut-points? This particular model produces two cut-points. How do we interpret each of these cut-points?

Second, we want to compare the optimal model from the variable selection procedure with an alternate model containing the predictor variables A9_t and A11using the ROC curve (see pp. 78 in *Logistic Regression Using SAS*). Base on the results of the ROC curves, which model is better? Is one model always better than the other or does the model preference change with respect to the desired value for specificity?

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.

Treat each Part of the assignment as a separate subsection in your report. Part 1 will contain the results and discussion of the EDA. Part 2 will contain the results from two fitted logistic regression models, the variable selection summary table, a discussion of the modeling results and the three listed discussion questions. Part 3 will contain the results from two more fitted models with their associated ROC curve information, and a discussion of the ROC curve results. The document should be submitted in pdf format.