Assignment #5

Introduction:

In this assignment we will fit a logistic regression model for a binary response variable. In preparation for fitting the model we will conduct Exploratory Data Analysis (EDA) on the credit_approval data set. Our first step in the EDA process will be to use the proc freq and proc means statements in SAS to see the distribution of the categorical variables and continuous variables.

We will then use the output from the proc means to make educated guesses about how to discretize the continuous variables. After discretizing the continuous variables we will generate frequency tables for these variables to determine whether appropriate cutoff points were used. If necessary, we will adjust the cutoff points and discretize with new values for the bins.

Our final EDA step will be to apply proc means to all of the categorical variables and the discretized version of the continuous variables to get an idea of how well each variable predicts the outcome of the response variable.

Based on the EDA that we perform, we will choose the best single variable to use in fitting a logistic regression model. We will then apply an automated variable selection technique to let SAS determine the best single variable logistic regression model.

Finally, we will generate a ROC curve based on the optimal single variable regression model as chosen by the automated variable selection method. We will compare this ROC curve to a ROC curve generated from a given two variable logistic regression model to assess which model performs better within this data set.

Results:

Part 1: The Exploratory Data Analysis

(1) The frequency tables, like the one shown below, will be used in step 3 of part 1 in order to help us determine how to code the categorical variables into dummy variables.

A13	Frequency	Percent	Cumulative Frequency	Cumulative Percent
g	625	90.58	625	90.58
p	8	1.16	633	91.74
s	57	8.26	690	100.00

(2) The table shown below gives us various percentile values for the continuous variables, classified by value of the response variable. These figures help us to determine the values to use for discretizing the continuous predictor variables.

Y	N Obs	Variable	5th Pctl	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl	95th Pctl
0	383	A2	17.0800000	18.5800000	22.0000000	27.3300000	34.8300000	44.2500000	51.9200000
		A3	0.1650000	0.3750000	0.8350000	2.2100000	5.0000000	11.0000000	12.6250000
		A8	0	0	0.1250000	0.4150000	1.5000000	3.5000000	5.0850000
		A11	0	0	0	0	0	2.0000000	3.0000000
		A14	0	0	100.0000000	167.5000000	272.0000000	372.0000000	460.0000000
		A15	0	0	0	1.0000000	67.0000000	400.0000000	1000.00
1	307	A2	18.8300000	20.3300000	23.1700000	30.5000000	41.3300000	52.8300000	58.4200000
		A3	0.1650000	0.3750000	1.5000000	4.4600000	9.5400000	12.7500000	15.0000000
		A8	0	0.0400000	0.7500000	2.0000000	5.0000000	8.5000000	13.0000000
		A11	0	0	0	3.0000000	7.0000000	11.0000000	14.0000000
		A14	0	0	0	120.0000000	280.0000000	399.0000000	470.0000000
		A15	0	0	0	221.0000000	1210.00	4159.00	8000.00

After discretizing each of the continuous variables, we generate frequency tables like the one shown below in order to confirm that there is imbalanced distribution in each category for the two values of Y.

Table of Y by A15_discrete						
Y	A15_discrete					
Frequency Percent Row Pct Col Pct	1	2	3	4	5	Total
0	209 30.29 54.57 64.51	82 11.88 21.41 79.61	28 4.06 7.31 66.67	57 8.26 14.88 33.14	7 1.01 1.83 14.29	383 55.51
1	115 16.67 37.46 35.49	21 3.04 6.84 20.39	14 2.03 4.56 33.33	115 16.67 37.46 66.86	42 6.09 13.68 85.71	307 44.49
Total	324 46.96	103 14.93	42 6.09	172 24.93	49 7.10	690 100.00

(3) Using the information from the frequency tables that we generated in step one, we code each of the categorical variables into dummy variables. For each categorical variable we will need the number of dummy variables to be one less than the number of possible categories. In general, we will use the category with the fewest number of observations as the base category. This means that there will not be a dummy variable explicitly coded for this category, but rather the absence of a 1 in the related dummy variables will indicate that an observation belongs to the base category. For consistency, categorical

variables that can only take values of ‘t’ or ‘f’ (assumed to indicate true or false) will be converted into a dummy variable for the t category.

(4) At this point we must employ some technique to deal with observations that have missing values. A common method, and the one we will employ here, is to remove observations that have missing values for one or more variables. After running a conditional statement in the data step to remove observations with missing values, we determine that only 37 out of 690 observations were deleted. Since we deleted just over 5% of the observations we don’t expect that this will have a meaningful impact on our model.

(5) The cross-tab tables shown below display the number of observations by category for each of the categorical and discretized variables. In addition to seeing the distribution of variables, we also see the mean value of our binomial response variable, Y, for each category. By calculating the mean value of Y for each category of the categorical and discretized variables, we can evaluate which variables are the best predictors of the response variable. If a category tends to correspond with a response variable value of 0 (no event), then the mean value of Y for that category will be close to 0 and vice versa for a category that tends to correspond with a response variable value of 1. A variable whose categories’ mean value of Y lies near the extremes (0 or 1) will likely be a strong predictor of the response variable.

Analysis Variable : Y		
A1	N Obs	Mean
a	203	0.4679803
b	450	0.4466667

Analysis Variable : Y		
A4	N Obs	Mean
l	2	1.0000000
u	499	0.4989980
y	152	0.2960526

Analysis Variable : Y		
A5	N Obs	Mean
g	499	0.4989980
gg	2	1.0000000
p	152	0.2960526

Analysis Variable : Y		
A6	N Obs	Mean
aa	52	0.3653846
c	133	0.4511278
cc	40	0.7250000
d	26	0.2692308
e	24	0.5833333
ff	50	0.1400000
i	55	0.2545455
j	10	0.3000000
k	48	0.2708333
m	38	0.4210526
q	75	0.6533333
r	3	0.6666667
w	63	0.5238095
x	36	0.8333333

Analysis Variable : Y		
A7	N Obs	Mean
bb	53	0.4528302
dd	6	0.3333333
ff	54	0.1481481
h	137	0.6350365
j	8	0.3750000
n	4	0.5000000
o	2	0.5000000
v	381	0.4278215
z	8	0.7500000

Analysis Variable : Y		
A9	N Obs	Mean
f	304	0.0592105
t	349	0.7965616

Analysis Variable : Y		
A10	N Obs	Mean
f	366	0.2540984
t	287	0.7073171

Analysis Variable : Y		
A12	N Obs	Mean
f	351	0.4301994
t	302	0.4801325

Analysis Variable : Y		
A13	N Obs	Mean
g	598	0.4682274
p	2	0.5000000
s	53	0.2830189

Analysis Variable : Y		
A2_discrete	N Obs	Mean
1	60	0.2833333
2	123	0.4065041
3	134	0.4253731
4	161	0.4223602
5	175	0.5942857

Analysis Variable : Y		
A3_discrete	N Obs	Mean
1	64	0.4375000
2	116	0.3448276
3	149	0.3154362
4	159	0.4968553
5	165	0.6181818

Analysis Variable : Y		
A8_discrete	N Obs	Mean
1	237	0.2447257
2	112	0.3303571
3	165	0.6000000
4	75	0.6666667
5	64	0.8125000

Analysis Variable : Y		
A11_discrete	N Obs	Mean
1	366	0.2540984
2	111	0.4594595
3	27	0.7037037
4	32	0.8437500
5	117	0.9059829

Analysis Variable : Y		
A14_discrete	N Obs	Mean
1	206	0.6213592
2	161	0.3229814
3	106	0.2924528
4	47	0.3404255
5	133	0.5187970

Analysis Variable : Y		
A15_discrete	N Obs	Mean
1	302	0.3642384
2	101	0.2079208
3	39	0.3589744
4	164	0.6707317
5	47	0.8723404

From these tables we see that when the variable A9 takes on a value of ‘f’ then the mean value for Y is 0.0592 which is the lowest mean for any of the categories of any variable. This is a strong indicator that a value of ‘f’ for the A9 variable would be a good predictor for a value of 0 in the response variable Y. The mean value for Y when A9 has a value of ‘t’ is 0.7966 which is among the highest means for any category of any variable. This means that a value of ‘t’ in the A9 variable would be a strong indication that the response variable Y will take on a value of 1.

From our EDA we conclude that A9 is likely to be the best single predictor of Y.

Part 2: Model Building

Since we concluded, through EDA, that A9 would likely be the best single variable predictor for the response variable Y, we start by building a model utilizing the A9_t dummy variable that takes on a value of 1 when A9 is ‘t’ and 0 when A9 is ‘f.’

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.7656	0.2430	129.5239	<.0001
A9_t	1	4.1306	0.2770	222.3474	<.0001

$$\hat{g}(A9\_t) = -2.7656 + (4.1306 * A9\_t)$$

The table shown above gives the parameter estimates for the logistic model. Given these parameters, we estimate the equation as shown to the right.

The coefficient for A9_t is 4.1306 which means that every one unit increase in A9_t raises our estimate of the log-odds ratio by 4.1306. Since a value 1 in A9_t corresponds to an A9 value of 't' we can say that when A9=t, the log-odds ratio is 4.1306 higher than when A9=f. We calculate the estimated change in the odds ratio to be $e^{4.1306} = 62.2152$. When A9=t the odds of Y being equal to one are estimated to be over 62 times higher than when A9=f.

Using the selection=score option with the proc logistic procedure, we let SAS determine the best single variable logistic model. The summary table for this automated variable selection process is shown below.

Regression Models Selected by Score Criterion		
Number of Variables	Score Chi-Square	Variables Included in Model
1	356.4519	A9_t
1	133.3312	A10_t
1	107.6653	A11
1	72.2924	A8
1	28.0037	A3
1	23.1084	A7_h
1	22.1186	A7_ff
1	21.4453	A6_ff
1	21.2165	A2
1	19.7656	A4_y
1	19.4908	A15
1	17.8360	A4_u
1	13.6820	A6_q
1	12.6935	A6_cc
1	9.5729	A6_i
1	6.9598	A6_k
1	6.7484	A13_s
1	6.3903	A13_g
1	4.7420	A14
1	3.7018	A6_d

Regression Models Selected by Score Criterion		
Number of Variables	Score Chi-Square	Variables Included in Model
1	2.8772	A7_z
1	2.3946	A7_v
1	1.7618	A6_aa
1	1.7002	A6_e
1	1.6332	A12_t
1	1.3991	A6_w
1	0.9630	A6_j
1	0.3516	A7_dd
1	0.2564	A1_b
1	0.2003	A7_j
1	0.1692	A6_m
1	0.0354	A7_n
1	0.0032	A6_c
1	0.0000	A7_bb

The table above ranks each individual variable by the Chi-square Score of the logistic regression model that results from using that variable as the single predictor. This output reflects the selection of A9_t as the best single predictor variable. This agrees with the conclusion that we reached through our EDA. The model that was fit previously still applies.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	901.544	493.254
SC	906.025	502.218
-2 Log L	899.544	489.254

The model fit statistics are shown at the left. For all three of these measures; AIC, SC, and -2 Log L; smaller values indicate a better fit. The real value of these values is in comparing the relative fit of multiple models. Since we are only looking at the values for a single model, these numbers don't tell us much except to say that the model with a predictor variable is a better fit than an intercept only model.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	75.2	Somers' D	0.740
Percent Discordant	1.2	Gamma	0.968
Percent Tied	23.6	Tau-a	0.367
Pairs	105672	c	0.870

In the table to the left we see the Somer's D, Gamma, Tau-a, and c values for our model. These figures all measure the in-sample predictive power of the model. They do so by comparing the percent of concordant pairs versus the percent of discordant pairs. Values for these measures range from 0 to 1 with values closer to 1 indicating better predictive power.

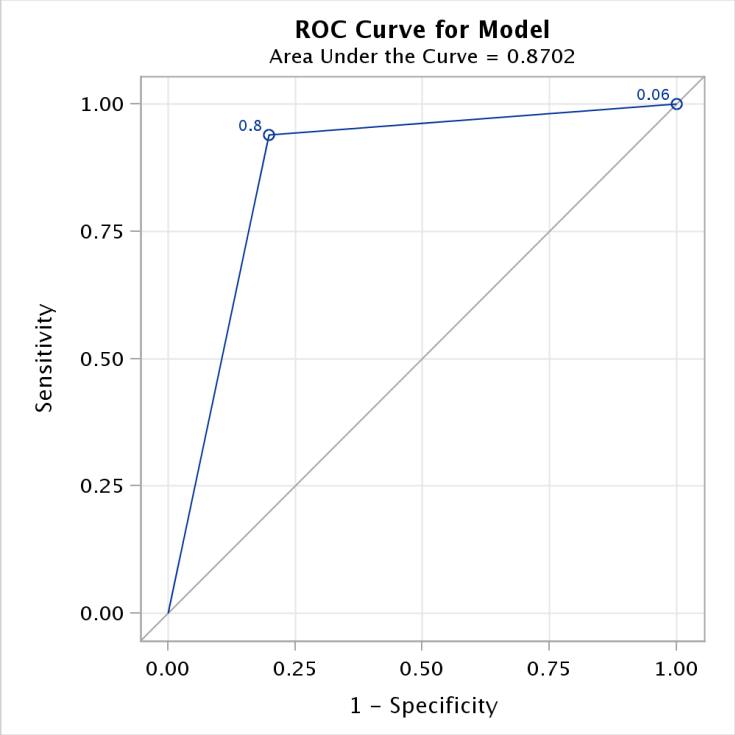
In order to determine the number of concordant and discordant pairs, we start by creating all possible pairs of observations. Then, taking only the pairs where one observation has a value of 1 for the response variable and the other observation has a value of 0 for the response variable, we calculate

whether or not the model estimated a higher predicted value for the observation that has a value of 1 for the response variable. If the model correctly predicted which of the two observations would have a value of 1 for the response variable then it is a concordant pair. If the model incorrectly predicted one observation over the other then it is a discordant pair. If the model favored neither pair then the pair is a tie.

Out of the four measures of predictive accuracy on the right hand side of the table titled "Association of Predicted Probabilities and Observed Responses," the c statistic may be the most popular. Part of the appeal of the c statistic is due to the fact that it corresponds to the area under the ROC curve which is discussed further in the next section.

Part 3: Model Assessment Using the ROC Curve

Using the single variable logistic model that we fit in the previous section we generate a ROC curve which is shown below.



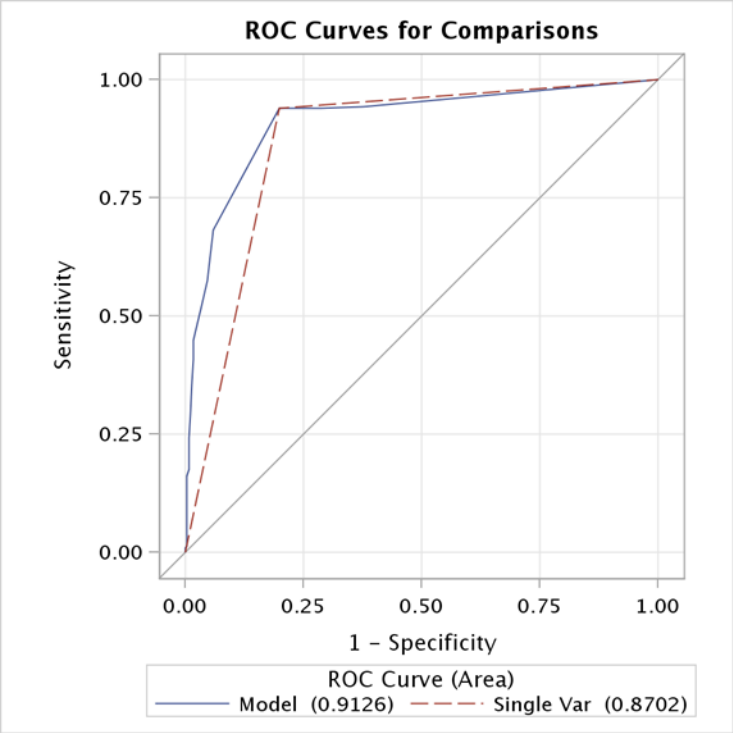
The ROC curve illustrates the relationship between the sensitivity and specificity of our model. The sensitivity is the proportion of events that are correctly predicted. The specificity is the number of non-events that are correctly predicted. An ideal model has both high sensitivity and high specificity. The higher the sensitivity and specificity are, the further the ROC curve will be from the diagonal line on the graph. The area under the curve is measured from the diagonal line to the edges of the curve with a maximum area of 1. The area under the ROC curve of 0.8702 corresponds to the value of the c statistic that was described in the previous section.

The graph of the ROC curve shown above displays two cut-points, which are labeled with values of 0.8 and 0.06. These values are the probability that serves as a cutoff point for concluding whether the model is predicting a positive outcome (response variable value of 1) or not. For example, at the 0.8 cut-point the model would have to estimate a probability of 0.8 or higher in order for us to predict that Y=1 for an observation.

The 0.8 cut-point reflects that at this probability level the sensitivity is somewhat higher than the specificity. At the 0.06 probability level the sensitivity is close to, or equal to, 1 and the specificity is close to, or equal to, 0. If we chose 0.06 as our cut-point, we would correctly predict nearly all of the events

(Y=1), but we would fail to predict nearly all of the non-events (Y=0). The choice of a proper cut-point depends on how the model will be used. As a default method, we can choose a cut-point where sensitivity and specificity are nearly equal to each other. However, there are situations where specificity is more important that sensitivity and vice versa.

The graph below compares our best single variable logistic model to a model containing two predictor variables, A9_t and A11.



On this graph the single variable model is shown as a dashed red line; the ROC curve for our two variable model is shown as a solid blue line. The ROC curve for the two variable model has more area under the curve so it is considered a better predictor overall. On this graph the 0.8 cut-point that was shown on our previous graph appears to be an important inflection point. For higher levels of specificity (further to the left along the x-axis), the sensitivity level is higher for the two variable model. For lower levels of specificity (further to the right along the x-axis), the single variable model has higher sensitivity than the two variable model.

Conclusion

As a result of our EDA we concluded that A9_t would be the best single predictor variable for building a logistic regression model to predict values of the response variable. The automated variable selection process that we used confirmed this finding. By looking at some measures of prediction accuracy we concluded that a single variable logistic regression model built using A9_t as the single predictor is a very good model for predicting values of the response variable. Comparing the ROC curve of our single variable model and the ROC curve for a two variable model, we showed that the two variable model is a better predictor when for higher levels of specificity, while the single variable model is better for lower levels of specificity.

Code

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

data temp;
      set mydata.credit_approval;
      if A16='+' then Y = 1;
            else if A16='-' then Y = 0;
            else Y = '.';

*Step 1:1;
/*
run;

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;
*/

*Step 2;
      if (A2 < 19) then A2_discrete=1;
            else if (A2 < 23) then A2_discrete=2;
            else if (A2 < 28) then A2_discrete=3;
            else if (A2 < 37) then A2_discrete=4;
            else A2_discrete=5;

      if (A3 < .38) then A3_discrete=1;
            else if (A3 < 1.25) then A3_discrete=2;
            else if (A3 < 3) then A3_discrete=3;
            else if (A3 < 7.2) then A3_discrete=4;
            else A3_discrete=5;

      if (A8 < .5) then A8_discrete=1;
            else if (A8 < 1.2) then A8_discrete=2;
            else if (A8 < 3.25) then A8_discrete=3;
            else if (A8 < 6) then A8_discrete=4;
            else A8_discrete=5;

      if (A11 < 1) then A11_discrete=1;
            else if (A11 < 3) then A11_discrete=2;
            else if (A11 < 4) then A11_discrete=3;
            else if (A11 < 6) then A11_discrete=4;
            else A11_discrete=5;

      if (A14 < 100) then A14_discrete=1;
            else if (A14 < 175) then A14_discrete=2;
            else if (A14 < 250) then A14_discrete=3;
            else if (A14 < 300) then A14_discrete=4;
            else A14_discrete=5;
```

```
      if (A15 < 1.5) then A15_discrete=1;
            else if (A15 < 75) then A15_discrete=2;
            else if (A15 < 200) then A15_discrete=3;
            else if (A15 < 3000) then A15_discrete=4;
            else A15_discrete=5;


ods graphics on;
proc freq data=temp;
      tables Y*A2_discrete;
run; quit;

proc freq data=temp;
      tables Y*A3_discrete;
run; quit;

proc freq data=temp;
      tables Y*A8_discrete;
run; quit;

proc freq data=temp;
      tables Y*A11_discrete;
run; quit;

proc freq data=temp;
      tables Y*A14_discrete;
run; quit;

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


*Step 3;
      if (A1 = 'b') then A1_b=1;
            else A1_b=0;

      if (A4 = 'u') then A4_u=1;
            else A4_u=0;
      if (A4 = 'y') then A4_y=1;
            else A4_y=0;

      if (A5 = 'g') then A5_g=1;
            else A5_g=0;
      if (A5 = 'p') then A5_p=1;
            else A5_p=0;

      if (A6 = 'c') then A6_c=1;
            else A6_c=0;
      if (A6 = 'q') then A6_q=1;
            else A6_q=0;
      if (A6 = 'w') then A6_w=1;
            else A6_w=0;
      if (A6 = 'i') then A6_i=1;
            else A6_i=0;
      if (A6 = 'ff') then A6_ff=1;
            else A6_ff=0;
```

```
      if (A6 = 'aa') then A6_aa=1;
            else A6_aa=0;
      if (A6 = 'k') then A6_k=1;
            else A6_k=0;
      if (A6 = 'cc') then A6_cc=1;
            else A6_cc=0;
      if (A6 = 'm') then A6_m=1;
            else A6_m=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 = 'j') then A6_j=1;
            else A6_j=0;

      if (A7 = 'v') then A7_v=1;
            else A7_v=0;
      if (A7 = 'h') then A7_h=1;
            else A7_h=0;
      if (A7 = 'bb') then A7_bb=1;
            else A7_bb=0;
      if (A7 = 'ff') then A7_ff=1;
            else A7_ff=0;
      if (A7 = 'j') then A7_j=1;
            else A7_j=0;
      if (A7 = 'z') then A7_z=1;
            else A7_z=0;
      if (A7 = 'dd') then A7_dd=1;
            else A7_dd=0;
      if (A7 = 'n') then A7_n=1;
            else A7_n=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;
      if (A13 = 's') then A13_s=1;
            else A13_s=0;

if A1='?' or A4='?' or A5='?' or A6='?' or A7='?' or A9='?' or A10='?' or
A12='?' or A13='?'
      or A2=' ' or A3=' ' or A8=' ' or A11=' ' or A14=' ' or A15=' '
then delete;

run;

ods graphics on;

*Step 5;
%macro class_mean(c);
```

```
      proc means data=temp mean;
      class &c. ;
      var Y;
run;
%mend class_mean;

%class_mean(c=A1);
%class_mean(c=A4);
%class_mean(c=A5);
%class_mean(c=A6);
%class_mean(c=A7);
%class_mean(c=A9);
%class_mean(c=A10);
%class_mean(c=A12);
%class_mean(c=A13);
%class_mean(c=A16);

%class_mean(c=A2_discrete);
%class_mean(c=A3_discrete);
%class_mean(c=A8_discrete);
%class_mean(c=A11_discrete);
%class_mean(c=A14_discrete);
%class_mean(c=A15_discrete);


*Part 2;
proc logistic data=temp;
      model Y (event='1') = A9_t / clodds = pl;
run;

proc logistic data=temp;
      model Y (event='1') = A2 A3 A8 A11 A14 A15
            A1_b A4_u A4_y A5_g A5_p A6_c A6_q A6_w A6_i A6_aa A6_ff
            A6_k A6_cc A6_m A6_d A6_e A6_j A7_v A7_h A7_bb A7_ff A7_j
            A7_z A7_dd A7_n A9_t A10_t A12_t A13_g A13_s / selection=score
start=1 stop=1 clodds=pl;
run; quit;

*Part 3;
proc logistic data=temp plots(only)=roc(id=prob);
      model Y (event='1') = A9_t / outroc=roc1;
run;

ods graphics off;

proc print data=roc1;
run; quit;

ods graphics on;

proc logistic data=temp;
      model Y (event='1') = A9_t A11;
      ROC 'Single Var' A9_t;
      ROCCONTRAST / ESTIMATE=ALLPAIRS;
run;

ods graphics off;
```