Assignment #5

Introduction:

In this exercise, we will perform an Exploratory Data Analysis on **credit_approval** data set to identify the best variable to model in a single variable logistic regression model. Subsequently, we will fit this single variable model, and also compare it against the model selected by an automated variable selection procedure. We will assess the goodness of fit and model adequacy of these models. Finally, we will look at the ROC curve of the optimal model chosen and another with two specific variables and compare the two.

Results:

Part 1: Exploratory Data Analysis

The credit_approval data set identifies 16 variables, A1-A16, with variable A16 used to identify whether the credit request is approved or not with a "+" or "-" sign. As the first step in this exercise, we create a response variable Y with values '1' and '0' corresponding to values of '+' and '-'. This will helps us in the logistic regression exercise where the response variable takes on values of 1 and 0. The table below shows a partial snapshot of the dataset with response variable Y added.

Obs	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	Y
1	b	30.830	0.000	u	g	w	v	1.250	t	t	1.000	f	g	202.000	0.000	+	1
2	a	58.670	4.460	u	g	q	h	3.040	t	t	6.000	f	g	43.000	560.000	+	1
3	a	24.500	0.500	u	g	q	h	1.500	t	f	0.000	f	g	280.000	824.000	+	1
4	b	27.830	1.540	u	g	w	v	3.750	t	t	5.000	t	g	100.000	3.000	+	1
5	b	20.170	5.625	u	g	w	v	1.710	t	f	0.000	f	s	120.000	0.000	+	1
6	b	32.080	4.000	u	g	m	v	2.500	t	f	0.000	t	g	360.000	0.000	+	1
7	b	33.170	1.040	u	g	r	h	6.500	t	f	0.000	t	g	164.000	31285.00	+	1
8	a	22.920	11.585	u	g	cc	v	0.040	t	f	0.000	f	g	80.000	1349.000	+	1
9	b	54.420	0.500	y	p	k	h	3.960	t	f	0.000	f	g	180.000	314.000	+	1
10	b	42.500	4.915	y	p	w	v	3.165	t	f	0.000	t	g	52.000	1442.000	+	1

The next step in the EDA is to discretize the continuous variables, which will handle any non-linear effects, and also serve to simplify the EDA process. In the credit_approval data set, the following variables are continuous: A2, A3, A8, A11, A14, A15.

By generating quantilies of the continuous variables, we get a sense of the distribution of the data and we can create categories such that each category for each of the variables has an "unbalanced" distribution of responses (more "0s" or more "1s"). This is important because if a given category has an equal number of each type of response, this implies this category does not have any explanatory power (as good as a coin toss).

Based on this criterion, the following categories were created for each of the continuous variables (labeled variablename_discrete):

Table of Y by A2_discrete					
Y	A2_discrete				
Frequency Percent Row Pct Col Pct	1	2	3	4	Total
0	28 4.29 7.84 80.00	31 4.75 8.68 64.58	252 38.59 70.59 56.88	46 7.04 12.89 36.22	357 54.67
1	7 1.07 2.36 20.00	17 2.60 5.74 35.42	191 29.25 64.53 43.12	81 12.40 27.36 63.78	296 45.33
Total	35 5.36	48 7.35	443 67.84	127 19.45	653 100.00

Table of Y by A3_discrete						
Y	A3_discrete					
Frequency Percent Row Pct Col Pct	1	2	3	4	5	Total
0	96 14.70 26.89 62.34	118 18.07 33.05 67.43	43 6.58 12.04 56.58	63 9.65 17.65 41.18	37 5.67 10.36 38.95	357 54.67
1	58 8.88 19.59 37.66	57 8.73 19.26 32.57	33 5.05 11.15 43.42	90 13.78 30.41 58.82	58 8.88 19.59 61.05	296 45.33
Total	154 23.58	175 26.80	76 11.64	153 23.43	95 14.55	653 100.00

Observe the Column percent in each category.

Table of Y by A8_discrete					
Y	A8_discrete				
Frequency Percent Row Pct Col Pct	1	2	3	4	Total
0	214 32.77 59.94 75.35	134 20.52 37.54 41.61	9 1.38 2.52 24.32	0 0.00 0.00 0.00	357 54.67
1	70 10.72 23.65 24.65	188 28.79 63.51 58.39	28 4.29 9.46 75.68	10 1.53 3.38 100.00	296 45.33
Total	284 43.49	322 49.31	37 5.67	10 1.53	653 100.00

Table of Y by A11_discrete					
Y	A11_discrete				
Frequency Percent Row Pct Col Pct	1	2	3	4	Total
0	333 51.00 93.28 69.81	17 2.60 4.76 17.53	5 0.77 1.40 10.64	2 0.31 0.56 6.25	357 54.67
1	144 22.05 48.65 30.19	80 12.25 27.03 82.47	42 6.43 14.19 89.36	30 4.59 10.14 93.75	296 45.33
Total	477 73.05	97 14.85	47 7.20	32 4.90	653 100.00

Table of Y by A14_discrete					
Y	A14_discrete				
Frequency Percent Row Pct Col Pct	1	2	3	4	Total
0	78 11.94 21.85 37.86	74 11.33 20.73 66.07	56 8.58 15.69 72.73	149 22.82 41.74 57.75	357 54.67
1	128 19.60 43.24 62.14	38 5.82 12.84 33.93	21 3.22 7.09 27.27	109 16.69 36.82 42.25	296 45.33
Total	206 31.55	112 17.15	77 11.79	258 39.51	653 100.00

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

The column percentage in each category for each variable shows how many of the observations in each category have a response of 1 versus 0, indicating the degree of “unbalance” in each category, and potentially pointing to predictive power for the particular category.

Now that all of the variables are discretized, all the 15 variables are now prepared for entering into the model through the use of dummy (or design) variables. Of course each of the variables has a different number of categories, and thus would need a different number of dummy variables. Using the methodology called as “reference cell” coding, we will explicitly code only k-1 dummy variables for a categorical variable with k categories – thus treating one category (typically the one with the smallest number of observations) as the base category.

At this stage, we want to clean up our dataset and remove any observations that have missing values in any of the variables. There are 37 such observations, leaving the data set with 653 observations that are not missing in any value.

Now, we are ready to begin examining the variables to see which of them is likely to have the most predictive value in use in single variable logistic regression model.

Observing each of the variables for the following characteristics:

a) Substantial number of observations in each category (so that results are not skewed by just a few observations)

b) Substantially higher number of “1s” or “0s” in each category – indicating predictive power in each category of the variable

c) Preferring fewer categories to a large number of categories

Here is a matrix of tables for each of the 15 variables:

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	35	0.2000000
2	48	0.3541667
3	443	0.4311512
4	127	0.6377953

Analysis Variable : Y		
A3_discrete	N Obs	Mean
1	154	0.3766234
2	175	0.3257143
3	76	0.4342105
4	153	0.5882353
5	95	0.6105263

Analysis Variable : Y		
A8_discrete	N Obs	Mean
1	284	0.2464789
2	322	0.5838509
3	37	0.7567568
4	10	1.0000000

Analysis Variable : Y		
A11_discrete	N Obs	Mean
1	477	0.3018868
2	97	0.8247423
3	47	0.8936170
4	32	0.9375000

Analysis Variable : Y		
A14_discrete	N Obs	Mean
1	206	0.6213592
2	112	0.3392857
3	77	0.2727273
4	258	0.4224806

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

There are a few categorical variables that are good candidates based on the criteria mentioned above: A9, A10, A11_discrete, and A15_discrete, with A9 embodying fewer categories with substantial number

of observations in each and a high degree of “unbalanced” number of the two responses in each of the categories.

Based on the above analysis, A9 should be chosen as the single variable for the single variable logistic regression model.

Part 2: Model Building

The EDA points to A9 as the single variable to build a single variable logistic regression model with. Entering A9 through dummy variables into the model, we have:

A9_t = 1 (for category t), and A9_t = 0 (for category f).

The logit for this model is thus

g(X) = b0 + b1*X where X is A9_t for this model.

and the probability function ∏ (x) = e^g(X)/ 1+e^g(X)

Using SAS proc logistic, the following parameter estimates are generated:

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

Thus the logit g(X) = -2.7656 + 4.1306*A9_t

The Wald Chi-Square statistics for both parameters are statistically significant (p < .0001).

The estimated fitted values are thus obtained from:

∏ (x) = e^(-2.7656 + 4.1306*A9_t) / 1 + e^(-2.7656 + 4.1306*A9_t)

Now let us employ the automatic selection method that compares the Score test generated for each of the single variable logistic regression models, going through all the fifteen variables in turn.

The results of this are shown here:

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_f
1	107.6653	A11
1	72.2924	A8
1	28.0037	A3
1	23.1084	A7_h
1	22.2053	A6_x
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
1	2.8772	A7_z
1	2.3946	A7_v
1	1.7618	A6_aa
1	1.7002	A6_e
1	1.6332	A12_f
1	1.3991	A6_w
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 highest score is received for the single variable logistic regression model with A9_t as the variable of choice. This thus is identical to the choice that an Exploratory Data Analysis of the data revealed.

Thus the logit and probability function show up as:

g(X) = -2.7656 + 4.1306*A9_t and

∏ (X) = e^(-2.7656 + 4.1306*A9_t) / 1 + e^(-2.7656 + 4.1306*A9_t)

The best way to interpret this model is by looking at the Odds Ratio (OR), which for a logistic regression with a dichotomous independent variable (as in this case where A9_t =1 or 0 representing the two categories) is:

OR = e^b1 = e^4.1306 = 62.215

That is, the odds of the response variable being a '1' is 62.2 times greater if the value of A9_t = 1.

Thus, in other words, the coefficient of the dummy variable is simply ln(OR).

Incidentally, the same interpretation extends to a polychotomous independent variable (which enters through multiple dummy variables into the model), where the OR is calculated between the particular category and the control category (which is typically represented by all '0' in the dummy variables). Each of the coefficients of the dummy variables then represent the appropriate ln(OR).

If a dummy variable is dropped, ideally all the dummy variables associated with all the categories of that variable should be dropped. Doing so would ensure that the entire variable is dropped from the model. This would then become a reduced model. If however, only one of the dummy variables that represents a particular category of a variable is dropped – this amounts to recoding of the variable and the interpretations of the results must change accordingly (for example it may no longer represent reference cell coding and all the convenient results it leads us to).

Now let us test this model for adequacy and goodness of fit.

Goodness of fit statistics for this model is generated as shown here:

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 statistics clearly show that the covariate here has predictive power as the value of the deviance (-2LogL) goes down with the covariate (the single variable in this case) in the model. Thus deviance is declining, and the fit is improving.

Also, the values of AIC and BIC (or SC), which add a measure of penalty for each predictor added, goes down in value with the covariate in the model as well.

To verify that the model is adequate:

1) We test that the Global test of significance for the model holds. The results are shown as below:

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	410.2891	1	<.0001
Score	356.4519	1	<.0001
Wald	222.3474	1	<.0001

This test clearly shows that the Likelihood Ratio, Score and Wald tests all indicate significance ($p<0.0001$), indicating that the variable added to the model has statistically significant explanatory power.

2) The use of a classification table and its resulting measures of association begin giving a sense of how well the individual predictions made by the model are doing. One approach is to compare predicted value from the model for a '0' versus predicted value for a '1' and set the higher probability outcome as the predicted value. Then if the predicted value matches the observed value – these become concordant pairs, differences become discordant pairs. For cases where the predicted model values are identical for a '0' or a '1' are marked as ties. All possible pairs are considered when conducting this analysis.

   Useful measures of association include Tau-a, Gamma and Somer's D, where higher values indicate a strong association between observed and predicted values. All of these measures rely on the notion of concordant and discordant pairs of data.

Given C concordant pairs, D discordant, T ties and N pairs, these are calculated as:

Tau-a = C – D / N

Gamma = C-D / C + D

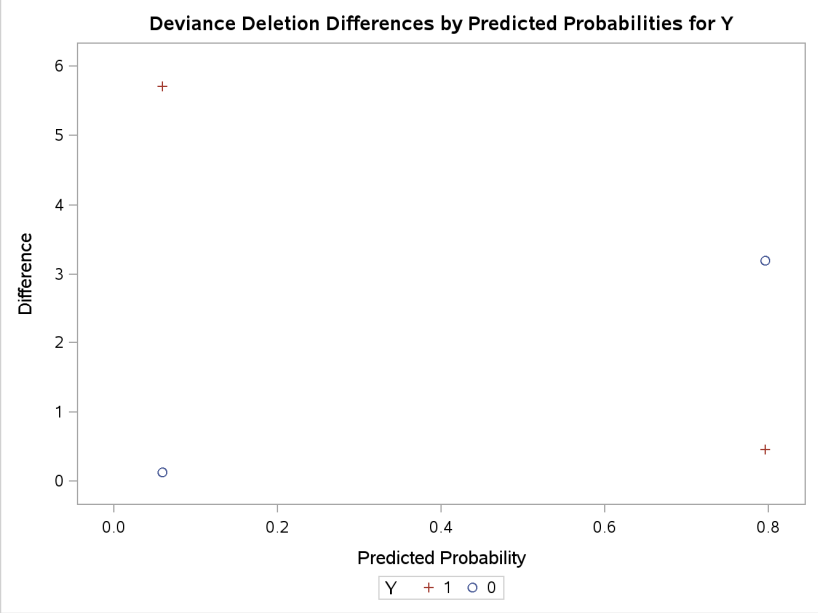
Somer's D = C –D / C + D + T

These tests for this model yield:

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

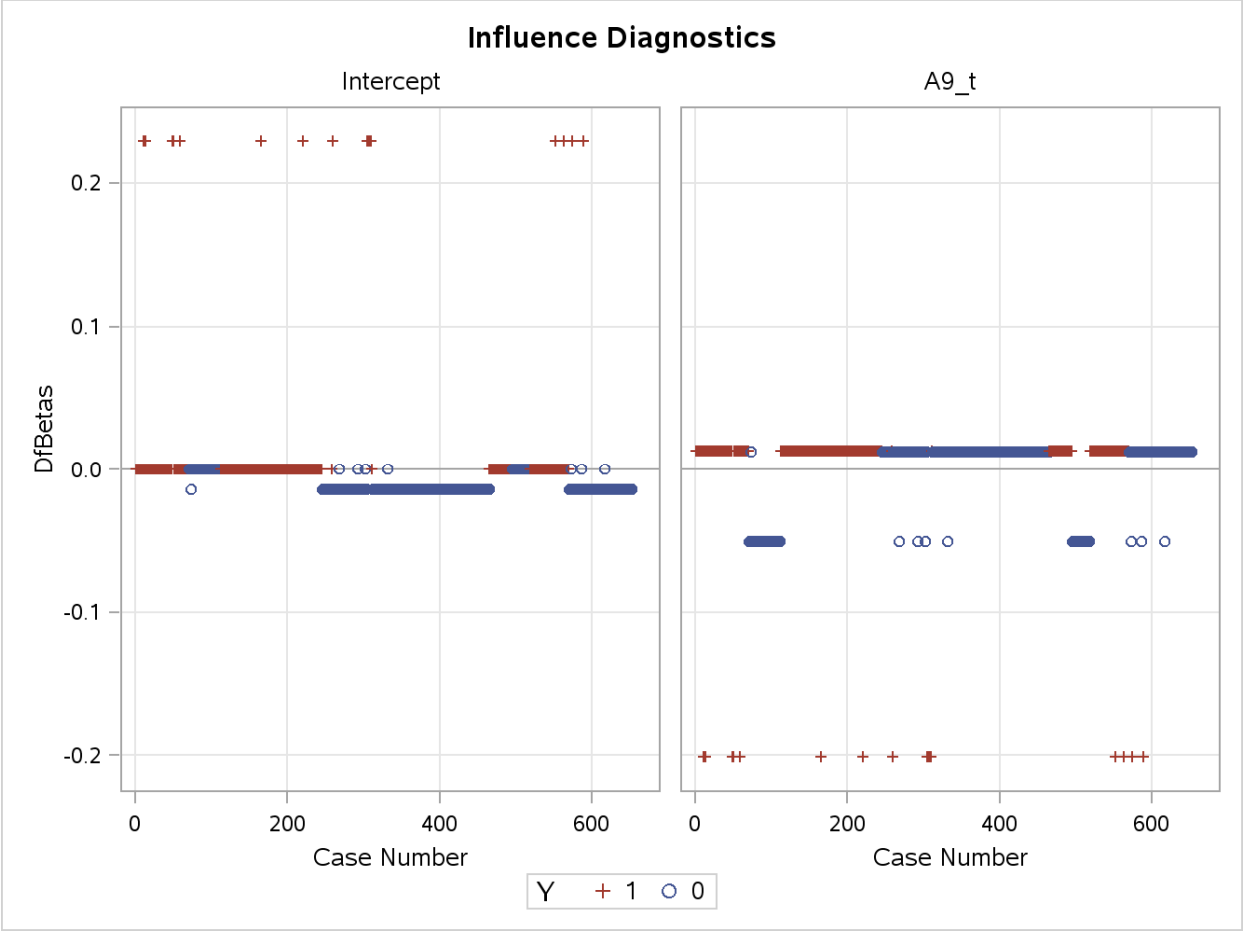
Thus among all possible pairs of combinations, 75.2 are concordant ("correctly classified"), 1.2 discordant and 23.6 are tied. The measure Gamma is indicating a very high level of association between the predicted and observed values, while Somer's D adjusts for the fairly high number of ties in this model, and gives a lower measure. Tau-a value tends to most closely mirror the value of a generalized R-square that is derived from the likelihood ratio.

3) Let us now check for high-influence points or points that are poorly fit in the model. For this, we will use the Deviance Deletion plot (this is essentially a check for change in deviance as each covariate pattern is deleted).



The graph shows two curves, one connecting points with dependent variable '1' (the + signs) and sloping down to the right, and one rising up to the right, connecting points with dependent variable '0' (the o signs). This graph shows up as expected and no points show up as a poor fit.

4) Another useful check is to look for any outliers in the data through the change in estimated coefficient when a particular observation is deleted. Large changes in this value can indicate presence of outliers in the data. The plot is below (DFBETAS indicate the standardized change in parameter estimates)

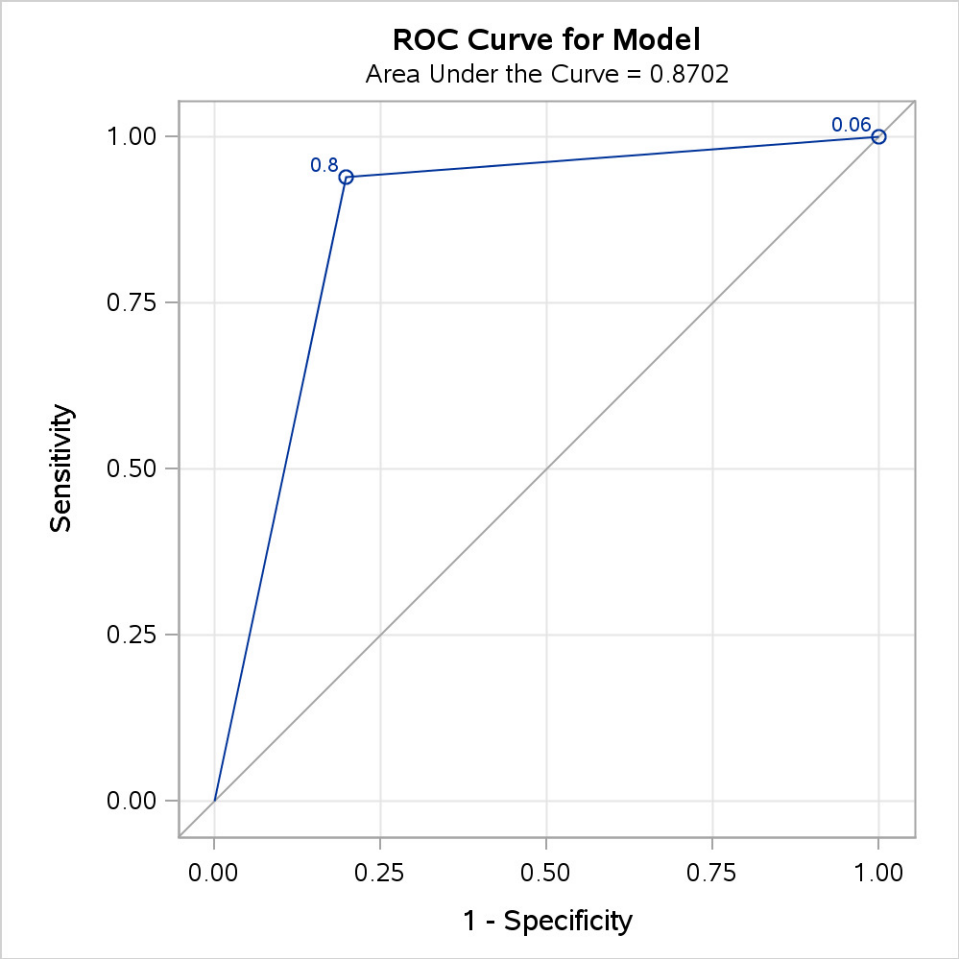


The plot does not indicate any significant change (> 1), thus there are no outliers in the dataset.

The model thus shows itself as adequate, and as having statistically significant explanatory powers over a model without the selected variable.

Part 3: Model Assessment using the ROC Curve

An ROC curve for the optimal model is shown below:



There are two cut-points shown in the curve, 0.8 and 0.06. Cut-points are probability values where predicted probabilities above that value are treated as a '1' in the response, and values below that are treated as a '0'. Obviously, a very low cut-point would begin marking all predicted probabilities as a '1' in the response, and a very high cut-point would mark all response values as a '0'.

These are better interpreted by the use of the terms Sensitivity and Specificity, where Sensitivity is the proportion of "true positives" – i.e. the proportion of predicted '1's that were observed to be '1' among all observed '1's. Specificity does the same for "true negatives" – i.e. the proportion of predicted '0's that were observed to be '0' among all observed '0's.

Thus a low cut-point value would make the model have high Sensitivity while a high cut-point value would make the model have high Specificity.

While the choice of a cut-point depends on the nature of the model, as a general rule of thumb one seeks to maximize sensitivity while maintaining a high specificity. In the ROC curve, which is drawn between Sensitivity and 1-Specificity, we are looking for a cut-point somewhere near the upper-left part of the curve.

The first cut-point in the above curve, at 0.8, is showing the point nearest to the upper-left part of the curve – it is a point of high Sensitivity & high Specificity together.

The second cut-point at the upper-right corner, 0.06, is where the value of sensitivity has reached 1.0, and is thus an extreme point on the scale. The Specificity has declined considerably at this stage. It is instructive to note that this point also coincides with the ROC for a model with an "intercept-only" line – in other words, for a model with no predictive power.

Also, it is worth noting that the area under the ROC-curve, which gives a measure of how good the model is at classifying the data under study, is 0.8702. This value ranges between 0.5 (no predictive power) and 1 (perfect predictive power). This is also the same as the C value in the "Association of Predicted Probabilities and Observed Responses" table.

Now, fitting a model with predictor variables A9_t and A11 both, we get

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-2.9245	0.2500	136.8938	<.0001
A9_t	1	3.7048	0.2837	170.4803	<.0001
A11	1	0.2076	0.0426	23.8107	<.0001

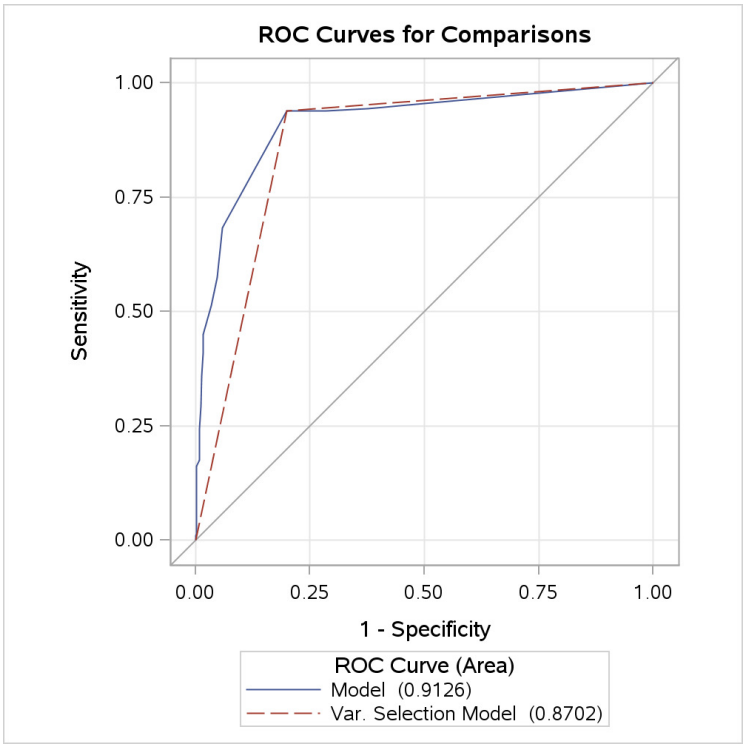
Both the intercepts are statistically significant, and the model is globally significant as shown below:

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	443.7932	2	<.0001
Score	368.6614	2	<.0001
Wald	211.9953	2	<.0001

The Association of predicted probabilities and observed responses are as below:

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	87.5	**Somers' D**	0.825
Percent Discordant	4.9	**Gamma**	0.893
Percent Tied	7.6	**Tau-a**	0.410
Pairs	105672	**C**	0.913

Now, comparing the optimal model ROC curve with this model with predictor variables A9_t and A11, we get this picture:



Observing the two models, we can see that the model with A9_t and A11 has a higher "area under the ROC curve", 0.9126 versus 0.8702 for the variable selection optimal model. This makes the model with predictor variables A9_t and A11 as the more suitable model based on the ROC criterion.

Observing the curve, we can also see that it has higher Sensitivity at high values of Specificity (low values of 1-Specificity) compared to the model chosen by variable selection.

However, as the level of Specificity needed decreases, we see that the optimal model (with only A9_t as the predictor variable), has a slightly higher level of Sensitivity for the same value of Specificity. Thus for select values of lower Specificity values (say ranging from 0.7-0.4 – corresponding to 1-Specificity of 0.3 to 0.6), the optimal model may be a better choice owing to higher corresponding Sensitivity.

Conclusions:

In this exercise, we prepared a dataset for performing a single variable logistic regression, followed by an initial Exploratory Data Analysis to identify the best candidate predictor variable. After examining the data for a categorical variable that shows significant difference in response for its different categories, we selected variable A9 which was then entered into the model as dummy variable A9_t.

The model was then fitted using logistic regression and the coefficients were found to be significant, and the model was found to be globally significant. An automatic variable selection method using the Score statistic was used to identify the best variable to fit a single variable logistic regression model. This method also selected A9, thus producing the same model as the EDA. The fitted model was then evaluated for goodness-of-fit and other measures of model adequacy and found to be adequate.

Finally, an ROC curve was generated for this model and cut-points assessed. It was then compared against another model with an additional variable, A11. This new model with both A9 and A11 showed itself to be more predictive (higher value of "area under the curve" for ROC), though for certain values of Specificity, the optimal single variable model showed higher Sensitivity on a comparative basis.

Code:

```
/* Assignment 5*/

/* Library at the SAS Servers
*/

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

/* Enabling plotting and other graphics
*/
ods graphics on;

/*****************************
 Part 1. Preparing the Data
*****************************/


/* Loading the data into 'temp' dataset
*/
data temp;
  set mydata.credit_approval;

  /* Setting response variable
  */
  if (A16="+") then Y=1;
  else if (A16="-") then Y=0;

  /* Discretizing continuous variables
  */
  if (A2 < 18) then A2_discrete=1;
  else if(A2 < 20) then A2_discrete=2;
  else if(A2 < 41) then A2_discrete=3;
  else A2_discrete=4;

  if (A3 < 1) then A3_discrete=1;
  else if(A3 < 3) then A3_discrete=2;
  else if(A3 < 4.5) then A3_discrete=3;
  else if(A3 < 11) then A3_discrete=4;
  else A3_discrete=5;


  if (A8 < .75) then A8_discrete=1;
  else if(A8 < 7.5) then A8_discrete=2;
  else if(A8 < 15) then A8_discrete=3;
  else A8_discrete=4;
```

```
if (A11 < 2.1) then A11_discrete=1;
else if(A11 < 7.1) then A11_discrete=2;
else if(A11 < 11.1) then A11_discrete=3;
else A11_discrete=4;

if (A14 < 100) then A14_discrete=1;
else if(A14 < 150) then A14_discrete=2;
else if(A14 < 200) then A14_discrete=3;
else A14_discrete=4;

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;

/**** Adding Dummy variables ***/
/* Treating 'a' as the base category */
if (A1='b') then A1_b=1; else A1_b=0;

/* Treating '1' as the base category*/
if (A2_discrete='2') then A2_discrete_2=1; else A2_discrete_2=0;
if (A2_discrete='3') then A2_discrete_3=1; else A2_discrete_3=0;
if (A2_discrete='4') then A2_discrete_4=1; else A2_discrete_4=0;

/* Treating '3' as the base category*/
if (A3_discrete='1') then A3_discrete_1=1; else A3_discrete_1=0;
if (A3_discrete='2') then A3_discrete_2=1; else A3_discrete_2=0;
if (A3_discrete='4') then A3_discrete_4=1; else A3_discrete_4=0;
if (A3_discrete='5') then A3_discrete_5=1; else A3_discrete_5=0;

/* 'l' is the samllest, set as base*/
if (A4='u') then A4_u=1; else A4_u=0;
if (A4='y') then A4_y=1; else A4_y=0;

/* 'gg' is the smallest category, set as base*/
if (A5='g') then A5_g=1; else A5_g=0;
if (A5='p') then A5_p=1; else A5_p=0;

/* 'r' is the smallest category, set as base*/
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='w') then A6_w=1; else A6_w=0;
if (A6='x') then A6_x=1; else A6_x=0;

/* 'o' is the smallest category, set as base*/
if (A7='bb') then A7_bb=1; else A7_bb=0;
```

```
if (A7='dd') then A7_dd=1; else A7_dd=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='j') then A7_j=1; else A7_j=0;
if (A7='n') then A7_n=1; else A7_n=0;
if (A7='v') then A7_v=1; else A7_v=0;
if (A7='z') then A7_z=1; else A7_z=0;

/* Treating '4' as the base category*/
if (A8_discrete='1') then A8_discrete_1=1; else A8_discrete_1=0;
if (A8_discrete='2') then A8_discrete_2=1; else A8_discrete_2=0;
if (A8_discrete='3') then A8_discrete_3=1; else A8_discrete_3=0;


/* 'f' is the smallest category, set as base*/
if (A9='t') then A9_t=1; else A9_t=0;

/* 't' is the smallest category, set as base*/
if (A10='f') then A10_f=1; else A10_f=0;

/* Treating '4' as the base category*/
if (A11_discrete='1') then A11_discrete_1=1; else A11_discrete_1=0;
if (A11_discrete='2') then A11_discrete_2=1; else A11_discrete_2=0;
if (A11_discrete='3') then A11_discrete_3=1; else A11_discrete_3=0;

/* 't' is the smallest category, set as base*/
if (A12='f') then A12_f=1; else A12_f=0;

/* 'p' is the smallest category, set as base*/
if (A13='g') then A13_g=1; else A13_g=0;
if (A13='s') then A13_s=1; else A13_s=0;

/* Treating '3' as the base category*/
if (A14_discrete='1') then A14_discrete_1=1; else A14_discrete_1=0;
if (A14_discrete='2') then A14_discrete_2=1; else A14_discrete_2=0;
if (A14_discrete='4') then A14_discrete_4=1; else A14_discrete_4=0;

/* Treating '3' as the base category*/
if (A15_discrete='1') then A15_discrete_1=1; else A15_discrete_1=0;
if (A15_discrete='2') then A15_discrete_2=1; else A15_discrete_2=0;
if (A15_discrete='4') then A15_discrete_4=1; else A15_discrete_4=0;
if (A15_discrete='5') then A15_discrete_5=1; else A15_discrete_5=0;
if (A15_discrete='6') then A15_discrete_6=1; else A15_discrete_6=0;

/* Deleting Missing Values Observations*/
if(A2='.')
  or (A3='.')
  or (A8='.')
  or (A11='.')
  or (A14='.')
  or (A15='.')
  or (A1='?')
  or (A4='?')
  or (A5='?')
  or (A6='?')
  or (A7='?')
  or (A9='?')
```

```
    or (A10='?')
    or (A12='?')
    or (A13='?') then delete;

run;

/* Macro to examine categorical variables against response variable Y
*/
%macro class_mean(c);
proc means data=temp mean;
class &c. ;
var Y;
run;
%mend class_mean;

/* Tables to help examining each categorical variable against Y*/
%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=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);


/* Cross-tabulated table to observe the discrete distribution
*/
proc freq data=temp;
tables Y*A1;
tables Y*A2_discrete;
tables Y*A3_discrete;
tables Y*A4;
tables Y*A5;
tables Y*A6;
tables Y*A7;
tables Y*A8_discrete;
tables Y*A9;
tables Y*A10;
tables Y*A11_discrete;
tables Y*A12;
tables Y*A13;
tables Y*A14_discrete;
tables Y*A15_discrete;
run;quit;

/* Initial examination of categorical variables
*/
proc freq data=temp;
tables A1 A4 A5 A6 A7 A9 A10 A12 A13 A16;
```

```
run;

/* Initial examination of continuous variables
*/
proc means data=temp p5 p10 p25 p50 p75 p90 p95;
class Y;
var A2 A3 A8 A11 A14 A15;
run;


/*****************************
 Part 2. Model Building
*****************************/

/* Fitting the model determined by EDA, including diagnostics
*/
proc logistic data=temp plots=phat (UNPACK);
  model Y (event='1') = A9_t;
run;quit;

/*
  Selecting variables by comparing models using the 'score' option
*/

proc logistic data=temp;
  model Y(descending) = A1_b A2 A3 A4_u A4_y A5_g A5_p A6_aa A6_c A6_cc
                        A6_d A6_e A6_ff A6_i A6_k A6_m A6_q A6_w A6_x
                        A7_bb A7_dd A7_ff A7_h A7_j A7_n A7_v A7_z A8
                        A9_t A10_f A11 A12_f A13_g A13_s A14 A15/ selection =
score start=1 stop=1;
run;quit;


/*****************************
 Part 3. Generate and compare ROC curves
*****************************/


/* ROC Curve for the optimal model
*/
proc logistic data=temp descending plots(only)=roc(id=prob);
  model Y (event='1') = A9_t / outroc=roc1;
run;
proc print data=roc1; run;quit;

/* Fitting the model with two variables A9, A11
*/
proc logistic data=temp;
  model Y (event='1') = A9_t A11;
run;quit;


/* Comparing ROC curve for the two models
*/
proc logistic data=temp descending plots(only)=roc(id=prob);
  model Y (event='1') = A9_t A11;
  ROC 'Var. Selection Model' A9_t;
```

```
run;


ods graphics off;
quit;
```