Derek Hughes
Assignment #5
Predict 410 – Sec 57

INTRODUCTION:

Here we are conducting an EDA analysis on a data set that has a dichotomous response variable. We will explore the data using EDA and try to estimate the single variable that will give us the highest probability to predict whether our response variable is positive or negative. Next, we will use the proc logistic procedure with the score selection method to determine which variable the procedure selects compared to the one we selected with our EDA. Next we will evaluate the goodness of fit metrics for a logistic regression to determine if the model does a good job determining the probability of the different response variables. Finally, we will explore the ROC curve with our model and then compare it to a model with an additional variable.

RESULTS:

Our EDA begins by looking at the categories in each categorical variable. We are looking at the frequency amounts of each category to determine which categories can be lumped into a base category. For example, in categorical variable A6, we can lump the category j and r into a base category because each has a very low frequency amount that's too low to use as a reasonable predictor. For the remaining categories (non-base categories) we create dummy variables for each category. Again, this is to provide us with more predictive and evaluative effectiveness when doing our EDA.

A1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
a	203	31.09	203	31.09
b	450	68.91	653	100.00

A4	Frequency	Percent	Cumulative Frequency	Cumulative Percent
l	2	0.31	2	0.31
u	499	76.42	501	76.72
y	152	23.28	653	100.00

A5	Frequency	Percent	Cumulative Frequency	Cumulative Percent
g	499	76.42	499	76.42
gg	2	0.31	501	76.72
p	152	23.28	653	100.00

A6	Frequency	Percent	Cumulative Frequency	Cumulative Percent
aa	52	7.96	52	7.96
c	133	20.37	185	28.33
cc	40	6.13	225	34.46
d	26	3.98	251	38.44
e	24	3.68	275	42.11
ff	50	7.66	325	49.77
i	55	8.42	380	58.19
j	10	1.53	390	59.72
k	48	7.35	438	67.08
m	38	5.82	476	72.89
q	75	11.49	551	84.38
r	3	0.46	554	84.84
w	63	9.65	617	94.49
x	36	5.51	653	100.00

A7	Frequency	Percent	Cumulative Frequency	Cumulative Percent
bb	53	8.12	53	8.12
dd	6	0.92	59	9.04
ff	54	8.27	113	17.30
h	137	20.98	250	38.28
j	8	1.23	258	39.51
n	4	0.61	262	40.12
o	2	0.31	264	40.43
v	381	58.35	645	98.77
z	8	1.23	653	100.00

A9	Frequency	Percent	Cumulative Frequency	Cumulative Percent
f	304	46.55	304	46.55
t	349	53.45	653	100.00

A10	Frequency	Percent	Cumulative Frequency	Cumulative Percent
f	366	56.05	366	56.05
t	287	43.95	653	100.00

A12	Frequency	Percent	Cumulative Frequency	Cumulative Percent
f	351	53.75	351	53.75
t	302	46.25	653	100.00

A13	Frequency	Percent	Cumulative Frequency	Cumulative Percent
g	598	91.58	598	91.58
p	2	0.31	600	91.88
s	53	8.12	653	100.00

A16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
+	296	45.33	296	45.33
-	357	54.67	653	100.00

Below we have the PROC MEANS summary statistics for the continuous variables. We use this table to determine effective “cut off” points to divide up the values of each continuous variable into discrete values that represent a range of values. This allows us to analyze the results more effectively. Generally we try to divide the values up into quantiles but the goal is to produce discrete ranges of values that can predict the binary response variable categories with a high degree of probability. This means that we want each discrete category to have a mean either close to 0 or close to 1, as a probability closer to 1 means that particular category has a stronger chance of making a correct prediction that the response variable will be 1 and vice versa if the value is close to 0. We want to avoid discrete categories with means close to .5, as that indicates it doesn’t have any more predictive power than the probability of guessing between two numbers. Through trial and error, we try different “cut off” points within each continuous variable to create discrete categories that have more predictive power (away from a mean of .5), although sometimes an effective “cut off” point isn’t possible with some variables (which indicates they should not be used in our predictive model).

Y	N Obs	Variable	5th Pctl	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl	95th Pctl
0	357	A2	17.0800000	18.4200000	21.9200000	26.9200000	34.8300000	44.2500000	51.9200000
		A3	0.1700000	0.3750000	0.8350000	2.2100000	5.0000000	11.0000000	12.7500000
		A8	0	0	0.1250000	0.4550000	1.5000000	3.3350000	5.0000000
		A11	0	0	0	0	0	2.0000000	3.0000000
		A14	0	0	100.0000000	160.0000000	260.0000000	360.0000000	454.0000000
		A15	0	0	0	1.0000000	67.0000000	400.0000000	1000.00
1	296	A2	18.8300000	20.4200000	23.2500000	31.0400000	41.4600000	52.8300000	58.4200000
		A3	0.2050000	0.4600000	1.5000000	4.4800000	9.5825000	13.0000000	16.0000000
		A8	0.0400000	0.0850000	0.7500000	2.0000000	5.0000000	8.5000000	13.8750000
		A11	0	0	0	3.0000000	7.0000000	12.0000000	14.0000000
		A14	0	0	0	120.0000000	280.0000000	400.0000000	480.0000000
		A15	0	0	0	210.5000000	1223.00	4159.00	8000.00

The following tables show the means of the categories we created of each continuous variable using the cut off points we selected for each variable. Here I created at least four categories within each continuous variable. Hence, we changed each continuous variable into a discrete variable with four categories that represent the continuous range of values for that particular continuous variable.

Analysis Variable : Y		
A2_discrete	N Obs	Mean
1	83	0.2891566
2	275	0.4290909
3	158	0.4303797
4	137	0.6277372

Analysis Variable : Y		
A3_discrete	N Obs	Mean
1	154	0.3766234
2	224	0.3392857
3	117	0.5470085
4	158	0.6202532

Analysis Variable : Y		
A8_discrete	N Obs	Mean
1	308	0.2662338
2	225	0.5422222
3	79	0.7468354
4	41	0.8048780

Analysis Variable : Y		
A11_discrete	N Obs	Mean
1	435	0.2896552
2	84	0.5833333
3	75	0.9200000
4	59	0.8813559

Analysis Variable : Y		
A14_discrete	N Obs	Mean
1	318	0.5220126
2	131	0.2748092
3	100	0.3800000
4	104	0.5384615

Analysis Variable : Y		
A15_discrete	N Obs	Mean
1	302	0.3642384
2	156	0.2628205
3	98	0.6530612
4	97	0.8350515

The following frequency tables can be used to evaluate our cut off points for the continuous variables and how those cut off points change the percentage of categorical values associated with the respective response variable categories.

Table of Y by A11_discrete					
Y	A11_discrete				
Frequency Percent Row Pct Col Pct	1	2	3	4	Total
0	309 47.32 86.55 71.03	35 5.36 9.80 41.67	6 0.92 1.68 8.00	7 1.07 1.96 11.86	357 54.67
1	126 19.30 42.57 28.97	49 7.50 16.55 58.33	69 10.57 23.31 92.00	52 7.96 17.57 88.14	296 45.33
Total	435 66.62	84 12.86	75 11.49	59 9.04	653 100.00

Table of Y by A15_discrete					
Y	A15_discrete				
Frequency Percent Row Pct Col Pct	1	2	3	4	Total
0	192 29.40 53.78 63.58	115 17.61 32.21 73.72	34 5.21 9.52 34.69	16 2.45 4.48 16.49	357 54.67
1	110 16.85 37.16 36.42	41 6.28 13.85 26.28	64 9.80 21.62 65.31	81 12.40 27.36 83.51	296 45.33
Total	302 46.25	156 23.89	98 15.01	97 14.85	653 100.00

We can also use PROC MEANS to evaluate the means of all of our discrete and categorical variables. The mean values found here are similar to those found on the frequency tables we created above. In particular, the mean values for each category are the equivalent to the percent of observations that are found when the response variable is 1. For example, for A15_discrete in the frequency table the percentage of category 1 observations that were also Y = 1 was 36.42%. In the PROC MEANS table for A15_discrete, for category 1 the means is .3642, which says that 36.42% of the time when the category is 1 for A15_discrete that the response variable is 1.

Below we have the means for each dummy variable. We are looking for a variable that shows a large difference in means for each response category. Also, we want these means to be either very close to 0 or 1 but opposing each other. For example, A9_t shows that it predicts a response of 1 almost 80% of the time and a response of 0 almost 95% of the time. Thus, it would be a reliable predictor of the response variable Y.

Analysis Variable : Y		
A1_b	**N Obs**	**Mean**
0	203	0.4679803
1	450	0.4466667

Analysis Variable : Y		
A4_u	**N Obs**	**Mean**
0	154	0.3051948
1	499	0.4989980

Analysis Variable : Y		
A5_g	**N Obs**	**Mean**
0	154	0.3051948
1	499	0.4989980

Analysis Variable : Y		
A6_aa	**N Obs**	**Mean**
0	601	0.4608985
1	52	0.3653846

Analysis Variable : Y		
A6_c	**N Obs**	**Mean**
0	520	0.4538462
1	133	0.4511278

Analysis Variable : Y		
A6_cc	**N Obs**	**Mean**
0	613	0.4355628
1	40	0.7250000

Analysis Variable : Y		
A6_ff	**N Obs**	**Mean**
0	603	0.4792703
1	50	0.1400000

Analysis Variable : Y

A6_i	N Obs	Mean
0	598	0.4715719
1	55	0.2545455

Analysis Variable : Y		
A6_k	N Obs	Mean
0	605	0.4677686
1	48	0.2708333

Analysis Variable : Y		
A6_m	N Obs	Mean
0	615	0.4552846
1	38	0.4210526

Analysis Variable : Y		
A6_q	N Obs	Mean
0	578	0.4273356
1	75	0.6533333

Analysis Variable : Y		
A6_w	N Obs	Mean
0	590	0.4457627
1	63	0.5238095

Analysis Variable : Y		
A6_x	N Obs	Mean
0	617	0.4311183
1	36	0.8333333

Analysis Variable : Y		
A7_bb	N Obs	Mean
0	600	0.4533333
1	53	0.4528302

Analysis Variable : Y		
A7_ff	N Obs	Mean

Analysis Variable : Y		
A1_b	**N Obs**	**Mean**
0	599	0.4808013
1	54	0.1481481

Analysis Variable : Y		
A7_h	**N Obs**	**Mean**
0	516	0.4050388
1	137	0.6350365

Analysis Variable : Y		
A7_v	**N Obs**	**Mean**
0	272	0.4889706
1	381	0.4278215

Analysis Variable : Y		
A9_t	**N Obs**	**Mean**
0	304	0.0592105
1	349	0.7965616

Analysis Variable : Y		
A10_f	**N Obs**	**Mean**
0	287	0.7073171
1	366	0.2540984

Analysis Variable : Y		
A12_f	**N Obs**	**Mean**
0	302	0.4801325
1	351	0.4301994

Analysis Variable : Y		
A13_g	N Obs	Mean
0	55	0.2909091
1	598	0.4682274

Analysis Variable : Y		
A13_s	N Obs	Mean
0	600	0.4683333
1	53	0.2830189

From our EDA I've concluded that A11_discrete and A15_discrete were the best continuous variables to predict the correct response value. I concluded that A9_t was the best dummy_variable created from the categorical variables. Overall I felt A9_t was the best for predicting the correct response variable value. I selected each of these variables based on the relatively high mean values each variable had in regard to the response variable value of 1 or the low mean value when the response variable value is 0.

------------- PART 2 --------------

I selected A9_t for the best predictor variable for the response variable values. I fit the model and the results are shown below.

Model Information	
Data Set	WORK.TEMPFILE
Response Variable	Y
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	653
Number of Observations Used	653

Response Profile		
Ordered Value	**Y**	**Total Frequency**
1	1	296
2	0	357

Probability modeled is Y=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

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

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

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

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
A9_t	62.213	36.148	107.071

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

Now, I used the selection "score" procedure to select the best variable. By looking for the variable with the highest chi-squares values, we can determine the best single variable for this model using the score selection procedure.

Model Information	
Data Set	WORK.TEMPFILE
Response Variable	Y
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	653
Number of Observations Used	653

Response Profile		
Ordered Value	Y	Total Frequency
1	1	296
2	0	357

Probability modeled is Y=1.

Note: The following variables are not used in the SCORE selection since they are a linear combination of other variables as shown.

A5_g =	A4_u

As we can see from the Chi-square score, similar to what we selected from our EDA, A9_t is also selected by the score procedure as the best individual variable for this regression model. When a dummy variable is dropped from the model, it is simply included in the base category from that point forward.

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	19.4908	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	2.3946	A7_v
1	1.7618	A6_aa
1	1.6332	A12_f
1	1.3991	A6_w
1	0.2564	A1_b
1	0.1692	A6_m

Regression Models Selected by Score Criterion		
Number of Variables	Score Chi-Square	Variables Included in Model
1	0.0032	A6_c
1	0.0000	A7_bb

Model Information	
Data Set	WORK.TEMPFILE
Response Variable	Y
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	653
Number of Observations Used	653

Response Profile		
Ordered Value	Y	Total Frequency
1	1	296
2	0	357

Probability modeled is Y=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

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 can be used to assess how well the model fits. In this case we are looking for the lowest value of the three, which is -2LogL.

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

When testing if the model is significant, we must check if each of the three values (Likelihood ratio, score, and Wald) is significant. If so, then we can say that the model has more statistically significant predictive power with the variable(s) than without the variables. As we can see, each metric has a probability of less than .0001 which says that the model is statistically significant.

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

The maximum likelihood estimates are used to determine the coefficients/estimates, the odd-ratio, probability or fitted values, and the test statistics to assess each parameter and the model. We can test if the individual variables have significant predictive power, which is significant in this case because p<.0001 for A9_t.

The g(x) or logit or log-odds or estimate of the logistic regression equation can be found from this table and is as follows: g(x)= -2.7656 + 4.1306*A9_t. This numeric coefficient of the variable A9_t can be interpreted as the expected change in the logit for every unit change in A9_t with the other variables held constant.

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
A9_t	62.213	36.148	107.071

The odds-ratio for A9_t is interpreted as the odds that the response value is 1 is 62.213 times greater when the predictor variable A9_t is 1 than when it is zero. It is calculated from the estimate or coefficient of the predictor variable A9_t…. e^(4.1306) = 62.213.

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

These measures are used to evaluate the association between the predicted values versus the observed values. These measures rely on concordant and discordant pairs. Concordant pairs are those pairs where the lower ordered response value (often 0) has a lower predicted mean score than the observation with the higher ordered response value. In other words, it is the percent of correctly classified pairs. This is desirable, while discordant pairs have a higher predicted mean score for lower order response values (less desirable).

Somers' D is used to determine the strength and direction of relation between pairs of variables. It has a value of -1 to +1 with +1 meaning that all pairs agree or are concordant. A Somers' D value of .740 shows fairly strong concordance with between the predicted and observed responses.

Gamma is similar to Somers' D except that it does not penalize for ties and therefore (using the same scale of -1 to +1) is usually higher value than Somers' D, which is what we see here as well.

Tau-a is similar to a generalized value of R-square that is derived from the likelihood ratio. It is defined to be the ratio of the difference between the number of concordant pairs minus the discordant pairs divided by the total number of possible pairs.

C is used to determine how well the model can discriminate the response. It's value ranges from 0.5 to 1, where 0.5 is randomly guessing (no predictive power). Thus we want a higher number and our number of .870 shows us that our model is strong at discriminating the response value.

Thus, we can see that, taken together, based on our concordant/discordant values, Somers' D, Gamma, Tau-a, and c values that we have a fairly strong model for predicting the response variable values correctly.

----------PART 3 -------------

Here we evaluate our model with a ROC curve and then compare it to another model with an additional variable using the ROC curve. The area under the ROC curve ranges from 0.5 to 1.0. This provides a measure of the model's ability to discriminate between those subjects who experience the "outcome of interest" versus those who do not. We would like to use a cutpoint that maximizes both sensitivity (probability of detecting a true value) and specificity (the probability of detecting a false value). We also use an arbitrary cutoff point, above which we consider the test to be abnormal and below to be normal. This cutoff point determines how many true positives, true negatives, false positives, and false negatives are shown. There is a tradeoff between sensitivity and specificity so different cutoff points will affect the sensitivity and specificity values. Also, the closer the curve comes to the 45-degree diagonal, the less accurate the test. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

Model Information	
Data Set	WORK.TEMPFILE
Response Variable	Y
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	653
Number of Observations Used	653

Response Profile		
Ordered Value	**Y**	**Total Frequency**
1	1	296
2	0	357

Probability modeled is Y=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

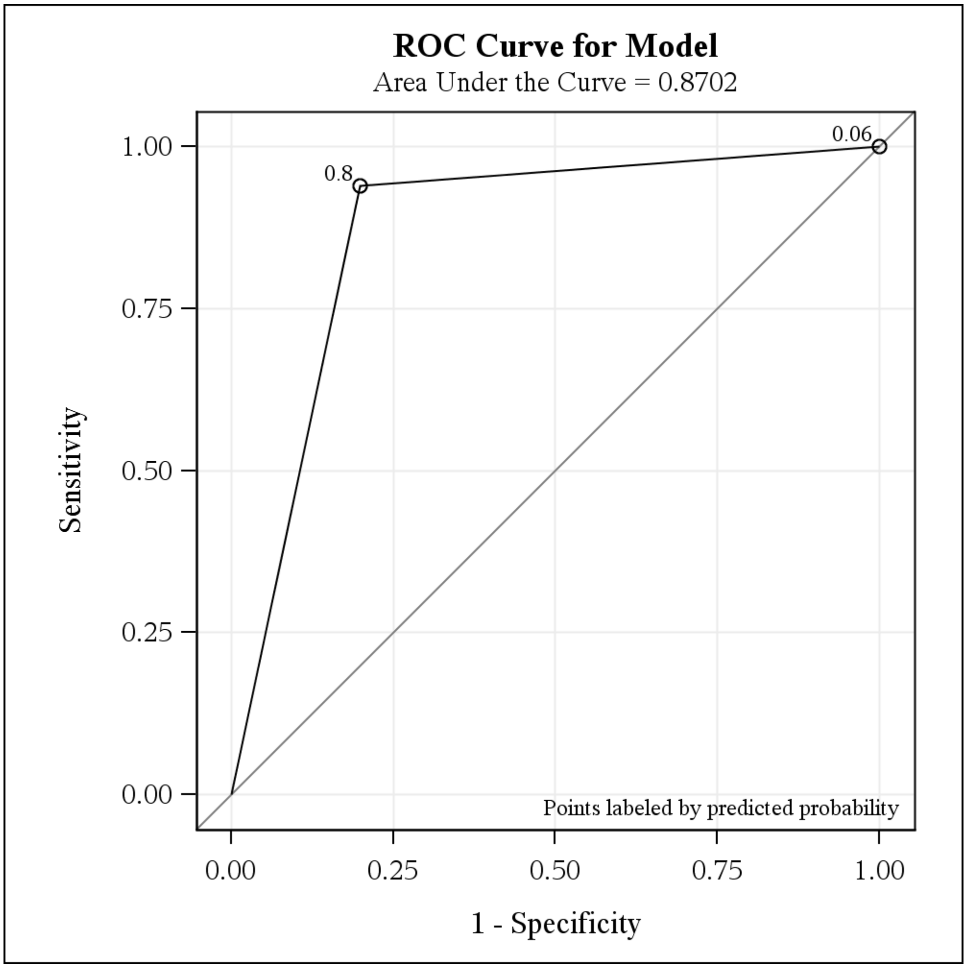
Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	901.544	493.254
SC	906.025	502.218
-2 Log L	899.544	489.254

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

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

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
A9_t	62.213	36.148	107.071

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



Obs	_PROB_	_POS_	_NEG_	_FALPOS_	_FALNEG_	_SENSIT_	_1MSPEC_
1	0.79656	278	286	71	18	0.93919	0.19888
2	0.05921	296	0	357	0	1.00000	1.00000

In this case the cutoff points are 0.8 and 0.06 which coincide with the mean values associated with our predictor variable A9_t with each response variable value (1 and 0). We can also see that the total area under the curve is .8702. This measures how well the test separates the variable being testing into those with and without the response value in question. Thus, the more area under the curve covered the better the test, while an area coverage of 0.5 indicates a completely useless test (same as guessing).

Model Information	
Data Set	WORK.TEMPFILE
Response Variable	Y
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	653
Number of Observations Used	653

Response Profile		
Ordered Value	Y	Total Frequency
1	1	296
2	0	357

Probability modeled is Y=1.

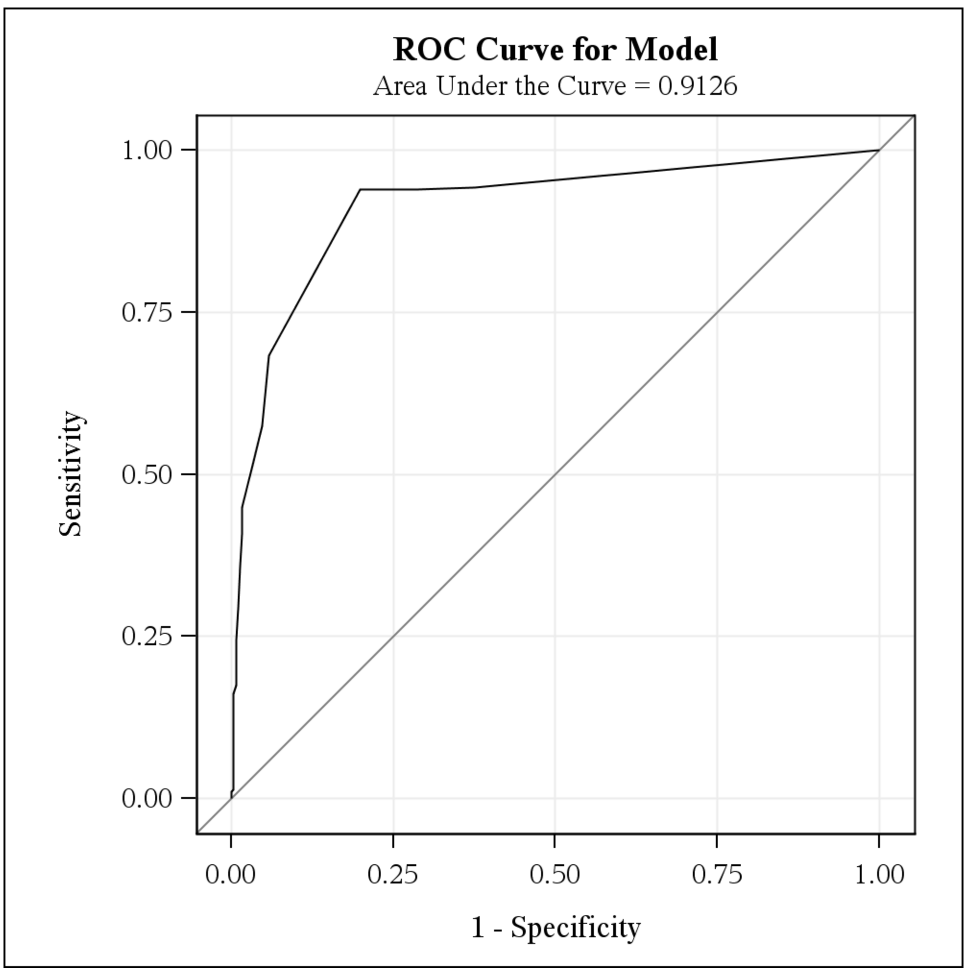
Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	901.544	461.750
SC	906.025	475.195
-2 Log L	899.544	455.750

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

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

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
A9_t	40.641	23.305	70.874
A11	1.231	1.132	1.338



ROC Model: omit a11

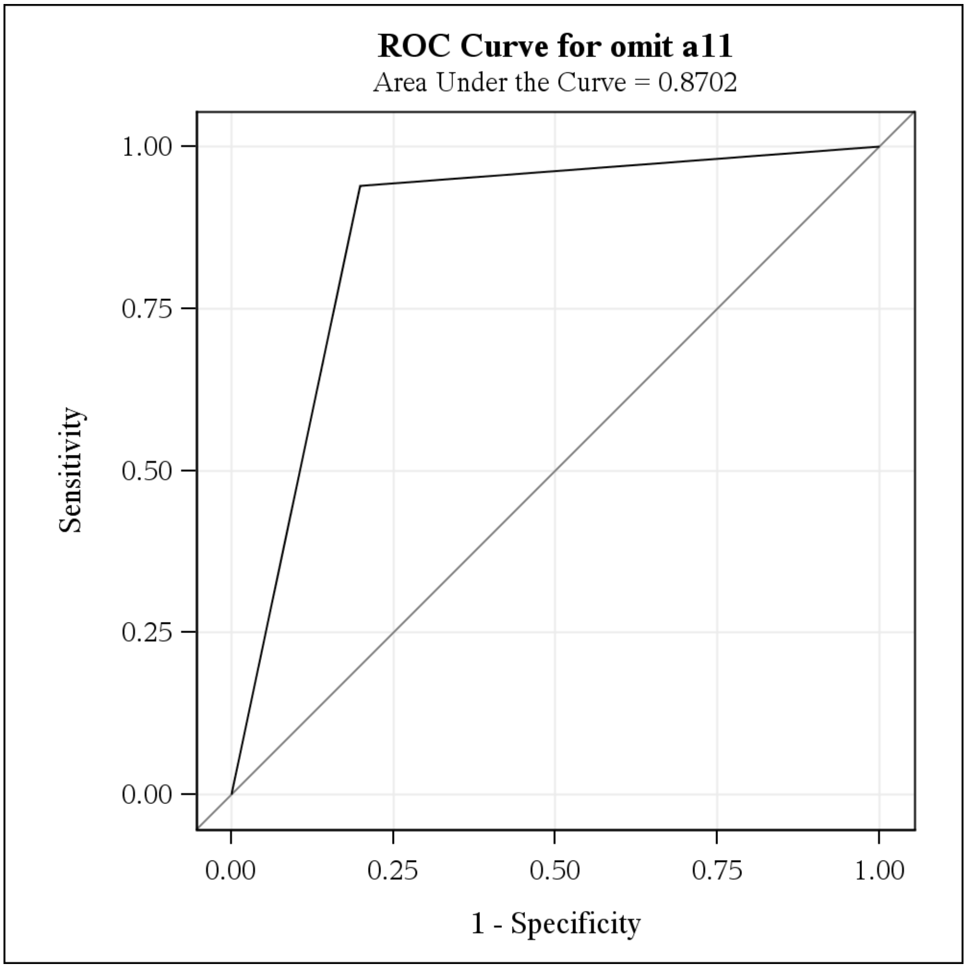
Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

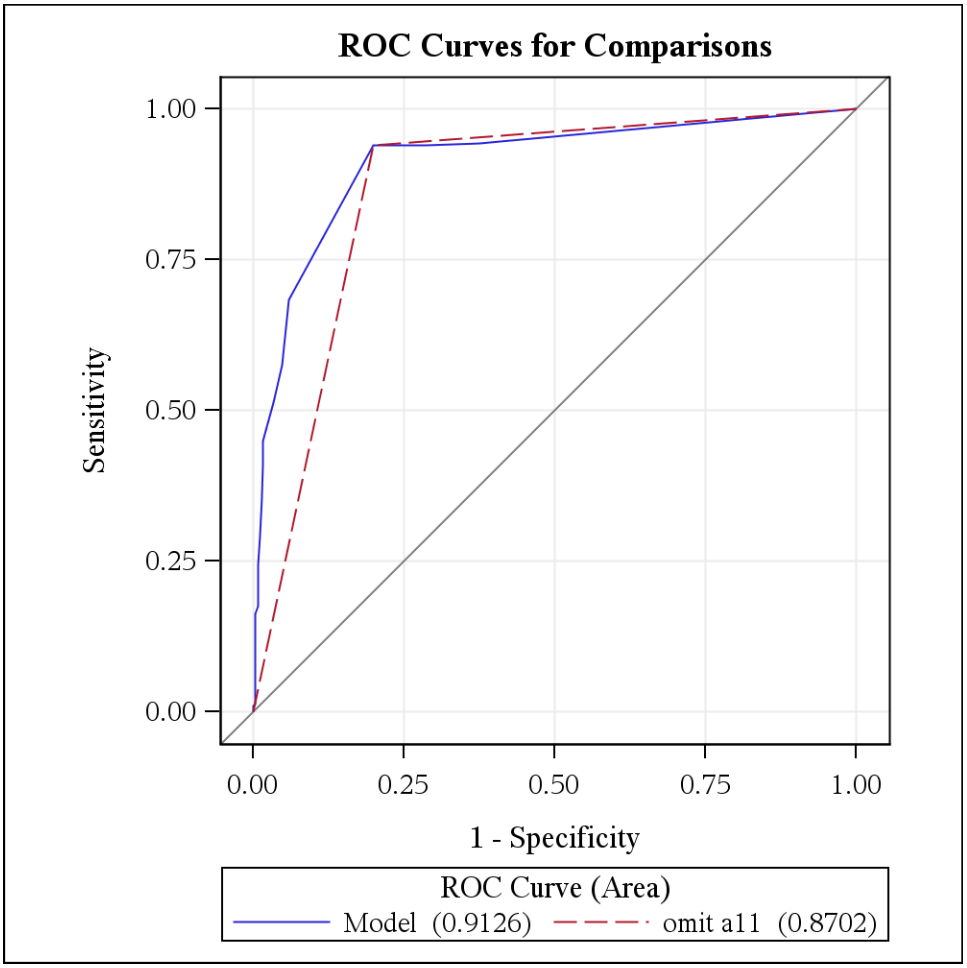
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	901.544	493.254
SC	906.025	502.218
-2 Log L	899.544	489.254

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

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

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
A9_t	62.213	36.148	107.071


ROC Curve for omit a11
Area Under the Curve = 0.8702
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificity



ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D (Gini)	Gamma	Tau-a
Model	0.9126	0.0116	0.8899	0.9353	0.8253	0.8930	0.4097
omit a11	0.8702	0.0127	0.8453	0.8950	0.7403	0.9684	0.3675

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Reference = Model	1	31.0489	<.0001

ROC Contrast Estimation and Testing Results by Row						
Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
Model - omit a11	0.0425	0.00762	0.0275	0.0574	31.0489	<.0001

We can see when comparing the models that the model that includes both variables (A9_t and A11) encompasses more area under the ROC curve comparatively speaking. This indicates that it is a better model for our purposes. We can also see that it is further to the left-hand side of the graph and almost exactly on the same plane at the top of the graph. As with any ROC analysis, if we want more specificity then we will sacrifice sensitivity which can change how much the model is preferred.

CONCLUSION-----

Using the EDA and PROC LOGISTIC with the score selection procedure we determined that A9_t was the best single variable we could use to predict the values of the response variable. Initially, we selected A9_t via EDA by comparing the difference between its mean values for each response variable value. Then we discovered that the selection score procedure selected the same variable so we fit our model and evaluated the results. We found that it was statistically significant as an individual variable and the model was statistically significant for predictive power as well. Finally, we created a ROC curve from which we could determine the sensitivity and specificity of the model and, again, found that it had strong predictive power. We also compared it to another model that included an additional variable and found the new model was slightly better in terms of its measures of area under the curve, specificity, and sensitivity.

CODE ----------

```
* creating the macro for data set;
%let PATH = /courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/;
%let NAME = MYDATA;
%let LIB = &NAME..;


libname &NAME. "&PATH." access=readonly;

%let INFILE = &LIB.credit_approval;
%let TEMPFILE = TEMPFILE;

* setting database to temp data file;
data &TEMPFILE.;
set &INFILE.;

        * set target variable ;
        if A16='+' then Y=1;
```

```
        else if A16='-' then Y=0;
        else Y = .;


* subsection 2 - discretize continuous variables;
        if (A2<20) then A2_discrete=1;
        else if (A2<30) then A2_discrete=2;
        else if (A2<40) then A2_discrete=3;
        else A2_discrete=4;

        if (A3<1) then A3_discrete=1;
        else if (A3<4) then A3_discrete=2;
        else if (A3<8) then A3_discrete=3;
        else A3_discrete=4;

        if (A8<1) then A8_discrete=1;
        else if (A8<4) then A8_discrete=2;
        else if (A8<8) then A8_discrete=3;
        else A8_discrete=4;

        if (A11<2) then A11_discrete=1;
        else if (A11<5) then A11_discrete=2;
        else if (A11<10) then A11_discrete=3;
        else A11_discrete=4;

        if (A14<150) then A14_discrete=1;
        else if (A14<225) then A14_discrete=2;
        else if (A14<325) then A14_discrete=3;
        else A14_discrete=4;

        if (A15 < 1.5) then A15_discrete=1;
        else if (A15 < 250) then A15_discrete=2;
        else if (A15 < 1001) then A15_discrete=3;
        else A15_discrete=4;

* subsection 3 - change variables to appropriate formats
        (continuous to discrete and categorical to dummy_variables)
  I'm combining any category with less than 30 observations and
        using them for the base;
  if (A1 ='b') then A1_b=1; else A1_b=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 ='ff') then A6_ff=1; else A6_ff=0;
        if (A6 ='i') then A6_i=1; else A6_i=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;

        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 ='f') then A10_f=1; else A10_f=0;

        if (A12 ='f') then A12_f=1; else A12_f=0;

        if (A13 ='g') then A13_g=1; else A13_g=0;
        if (A13 ='s') then A13_s=1; else A13_s=0;



* subsection 4 - delete missing values;
        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;

* to "fix" this data we could create some code that takes the mean or median values and substitutes it for these
variables;


run; quit;


* subsection 1 - proc freq to view details of categorical variables;
proc freq data=&TEMPFILE.;
tables A1 A4 A5 A6 A7 A9 A10 A12 A13 A16;
run;

* subsection 1 - proc means to view details of continuous variables;
proc means data=&TEMPFILE. p5 p10 p25 p50 p75 p90 p95;
class y;
```

```
var A2 A3 A8 A11 A14 A15;
run;



* subsection 5 - macro for class_mean(c);
%macro class_mean(c);
proc means data=&tempfile. mean;
*class a1 a4 a5 a6 a7 a8 a10 a12 a13;
class &c.;
var Y;
run;
%mend class_mean;


* discretized continuous variables;
%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);

* I selected A15_discrete and a11_discretes as best variabes from continuous variables;
proc freq data=&tempfile.;
tables Y*a11_discrete;
run;

proc freq data=&tempfile.;
tables y*a15_discrete;
run;

* categorical variables;
%class_mean(c=a1_b);
%class_mean(c=a4_u);
%class_mean(c=a5_g);
%class_mean(c=a6_aa);
%class_mean(c=a6_c);
%class_mean(c=a6_cc);
%class_mean(c=a6_ff);
%class_mean(c=a6_i);
%class_mean(c=a6_k);
%class_mean(c=a6_m);
%class_mean(c=a6_q);
%class_mean(c=a6_w);
%class_mean(c=a6_x);
```

```
%class_mean(c=a7_bb);
%class_mean(c=a7_ff);
%class_mean(c=a7_h);
%class_mean(c=a7_v);
%class_mean(c=a9_t);
%class_mean(c=a10_f);
%class_mean(c=a12_f);
%class_mean(c=a13_g);
%class_mean(c=a13_s);


* I selected A9 as best variable from categorical variables;
* I selected A9 as best categorical variable and a11_discrete and a15_discrete as best dummy variables;

******************* PART 2 *************************;
* fit logistic regression model to the variables- each dummy variable must be included;

* subsection 1 - proc logistic on variable selected from EDA;
proc logistic data=&tempfile. descending;
model Y = A9_t;
run;


* subsection 2 - proc logistic on all variables;
proc logistic data=&tempfile. descending;
model Y = 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_f a12_f a13_g a13_s/ selection=score start=1 stop=1;
run;


* model selects A9_t as best predictive variables;
proc logistic data=&tempfile. descending;
model y = a9_t;
run;



************* PART 3 *************;
ods graphics on;
proc logistic data=&tempfile. descending plots(only)=roc(id=prob);
model Y= A9_t / outroc=roc1;
run;
ods graphics off;
```

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


*** comparison model ****;
ods graphics on;
proc logistic data=&tempfile. descending;
model y = a9_t a11;
roc 'omit a11' a9_t;
roccontrast/ estimate=allpairs;
run;
ods graphics off;
```