Assignment #5

James Gray

Introduction:

This study is composed of three parts and explores the development of a single variable logistic regression model using multiple techniques. The objective is to predict a binary response of a credit approval using the credit_approval data set that is composed of 10 categorical variables and 6 continuous variables.

In part 1, an Exploratory Data Analysis (EDA) is conducted to evaluate the variables statistics and predictive accuracy of the attributes represented by the categorical variables. In part 2, a single variable logistic regression model is fit using a variable selected from the EDA as having the greatest predictive strength on credit approval. A second model is fit using the SCORE variable selection option to produce the best single variable logistic regression model. A comparison is then made to between the two approaches. In part 3, the optimal model is assessed using a Receiver Operating Characteristic (ROC) curve to evaluate the predictive power of the model and error tradeoffs. The optimal model is also compared with two other single variable models using ROC curves to evaluate if one model is always preferred with respect to the desired value of specificity.

Results:

Part 1 - Exploratory Data Analysis

The first step of the study analyzes the 16 independent variables of the credit_approval data set. Frequency distributions were produced to understand the values of the 10 categorical variables as shown in Figure 1.

A1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
?	12	1.74	12	1.74
a	210	30.43	222	32.17
b	468	67.83	690	100.00

A4	Frequency	Percent	Cumulative Frequency	Cumulative Percent
?	6	0.87	6	0.87
l	2	0.29	8	1.16
u	519	75.22	527	76.38
y	163	23.62	690	100.00

A5	Frequency	Percent	Cumulative Frequency	Cumulative Percent
?	6	0.87	6	0.87
g	519	75.22	525	76.09
gg	2	0.29	527	76.38
p	163	23.62	690	100.00

A6	Frequency	Percent	Cumulative Frequency	Cumulative Percent
?	9	1.30	9	1.30
aa	54	7.83	63	9.13
c	137	19.86	200	28.99
cc	41	5.94	241	34.93
d	30	4.35	271	39.28
e	25	3.62	296	42.90
ff	53	7.68	349	50.58
i	59	8.55	408	59.13
j	10	1.45	418	60.58
k	51	7.39	469	67.97
m	38	5.51	507	73.48
q	78	11.30	585	84.78
r	3	0.43	588	85.22
w	64	9.28	652	94.49
x	38	5.51	690	100.00

A7	Frequency	Percent	Cumulative Frequency	Cumulative Percent
?	9	1.30	9	1.30
bb	59	8.55	68	9.86
dd	6	0.87	74	10.72
ff	57	8.26	131	18.99
h	138	20.00	269	38.99
j	8	1.16	277	40.14
n	4	0.58	281	40.72
o	2	0.29	283	41.01
v	399	57.83	682	98.84
z	8	1.16	690	100.00

A9	Frequency	Percent	Cumulative Frequency	Cumulative Percent
f	329	47.68	329	47.68
t	361	52.32	690	100.00

A12	Frequency	Percent	Cumulative Frequency	Cumulative Percent
f	374	54.20	374	54.20
t	316	45.80	690	100.00

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

A16	Frequency	Percent	Cumulative Frequency	Cumulative Percent
+	307	44.49	307	44.49
-	383	55.51	690	100.00

Figure 1 - Frequency distributions for categorical variables

The 6 continuous variables were analyzed by producing a percentile distribution for each value of the response variable Y. This enables a visual inspection of how the values of the independent variables correspond to the binary output of Y. To contrast two variables, the distribution of A2 is fairly similar for each value of Y, while for A15 the output Y=0 has consistently lower values and Y=1 has significantly higher values. This highlights a potential strong relationship between A15 and Y. Lower values of A15 imply Y=0 while higher values of A15 imply Y=1.

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

Figure 2 - Percentile distributions for each continuous variable in relation to the binary response

The conditional distributions of Figure 2 were then used as insight to discretize the continuous variables into four or more categories for the purpose of creating an unbalanced distribution. Similar to described above, if there are similar values for each "segment" (category) of the continuous variable for the binary values of Y then variable is not a suitable predictor. If an imbalanced distribution exists, then this identifies that variable as a predictor candidate. The objective of the discretization is to capture non-linear relationships and to enable logistic regression using contingency tables. In this study, it would represent a cross-classification between a nominal level of the independent variable and the binary output variable Y as shown in Figure 3 for variable A15. The cut-point for each category of the continuous variable was manually selected and multiple iterations are often required for the purposes of creating an unbalanced distribution.

The contingency table in Figure 3 provides a similar view to the unbalanced distribution for variable A15 as seen in Figure 2. For example, for the category value=2 approximately 79% of the observation had Y=0, while when category value= 6 over 86% of the observation had Y=1. This confirms a relationship where small values of A15 predict Y=0 and larger values of A15 predict Y=1.

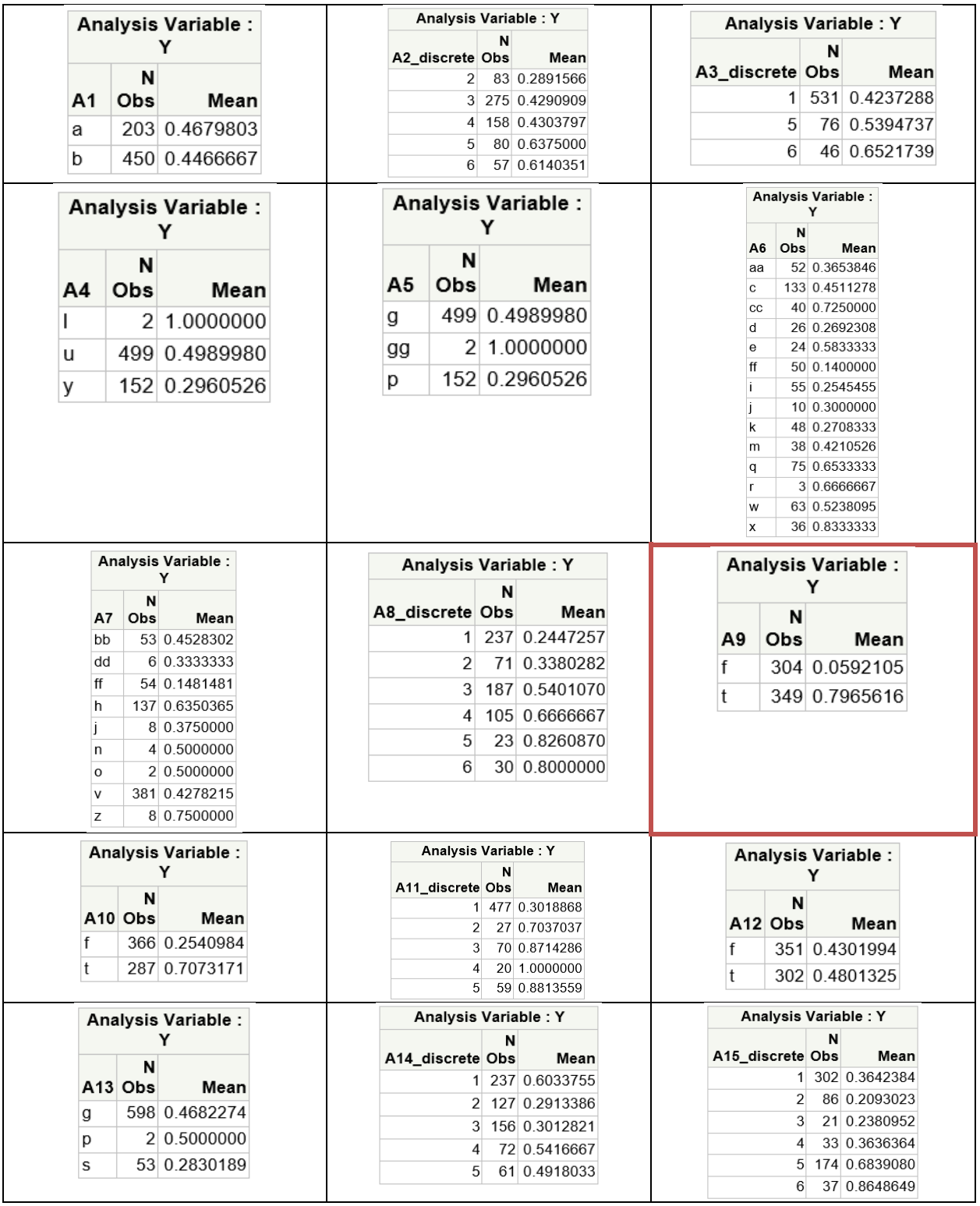
Table of Y by A15_discrete							
Y	A15_discrete						
Frequency Percent Row Pct Col Pct	1	2	3	4	5	6	Total
0	192 29.40 53.78 63.58	68 10.41 19.05 79.07	16 2.45 4.48 76.19	21 3.22 5.88 63.64	55 8.42 15.41 31.61	5 0.77 1.40 13.51	357 54.67
1	110 16.85 37.16 36.42	18 2.76 6.08 20.93	5 0.77 1.69 23.81	12 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

Figure 3 - Contingency Table for Y by A15_discrete

In this next step of the EDA we prepare the discrete, nominal scale variables for use in the model development. The nominal values for each discrete variable have no numerical meaning so it is would be inappropriate to include these values in the model together with interval scale variables (continuous). For example, the nominal values of "a" and "b" for variable A1 are just names. A collection of "design variables" are used to transform the nominal scale variables into data set representation that can then be used in the model fitting. The number of design variables required to represent a nominal variable with *k* possible values is *k*-1. For example, a discrete variable with three categories would require two design variables.

Missing values for the predictor variables can impact model results and adequacy. The entire credit_approval data set of 690 observation was reviewed and missing values for categorical variables are represented as "?" and "-" values for continuous variables. This data cleansing step removed 37 observations from the analysis.

The next step in the EDA evaluated the predictive accuracy of each categorical variable. This included the categorical variables (A1, A4, A5, A6, A7, A9, A10, A12, A13) as well as the discretized continuous variables. The analysis was facilitated by calculating a mean distribution of Y for level of the categorical variable as shown in Figure 4. The most significant imbalance is shown by variable A9. One could conclude that a value of "f" corresponds to Y=0 and a value of "t" corresponds to a value of Y=1.



Analysis Variable : Y

A1	N Obs	Mean
a	203	0.4679803
b	450	0.4466667

Analysis Variable : Y

A2_discrete	N Obs	Mean
2	83	0.2891566
3	275	0.4290909
4	158	0.4303797
5	80	0.6375000
6	57	0.6140351

Analysis Variable : Y

A3_discrete	N Obs	Mean
1	531	0.4237288
5	76	0.5394737
6	46	0.6521739

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

A8_discrete	N Obs	Mean
1	237	0.2447257
2	71	0.3380282
3	187	0.5401070
4	105	0.6666667
5	23	0.8260870
6	30	0.8000000

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

A11_discrete	N Obs	Mean
1	477	0.3018868
2	27	0.7037037
3	70	0.8714286
4	20	1.0000000
5	59	0.8813559

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

A14_discrete	N Obs	Mean
1	237	0.6033755
2	127	0.2913386
3	156	0.3012821
4	72	0.5416667
5	61	0.4918033

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

Figure 4 - Mean distributions for categorical variables

Part 2 – Model Building

In this part of the study two single variable logistic regression models are fit using the EDA results and an automated selection method. The purpose is to compare the model results to determine if the EDA provided a reliable process to select the optimal model or not.

Model Fitting Using EDA

Given that the value A9=t showed a strong correlation to predicting Y=1, the design variable A9_t was used to fit the logistic regression model. The fitted logit is given by:

$$g(x) = -2.756 + 4.1306x$$

Model Information	
Data Set	WORK.TEMP2
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	0	357
2	1	296

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

Figure 5 - Fitted Logistic Regression Model using A9_t

Model Selection using Score

The second single variable logistic regression model was fit using an automated selection method to find the optimal model. The selection method is called the SAS "score" option. This process also selected A9_t as the variable for the optimal single logistic regression model. This confirms that the optimal regression model was chosen by the EDA in part 1.

Model Information	
Data Set	WORK.TEMP2
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	0	357
2	1	296

Probability modeled is Y=1.

used in the SCORE selection since they are a linear combina

A5_gg =	Intercept - A4_u - A4_y
A5_p =	A4_y
A6_s =	0

Regression Models Selected by Score Criterion		
Number of Variables	Score Chi-Square	Variables Included in Model
1	356.4519	A9_t
1	107.6653	A11
1	72.2924	A8
1	28.0037	A3
1	23.1084	A7_h
1	22.1186	A7_ff

Figure 6 - Fitted Regression Model using the SCORE selection

Model Fitting Interpretations and Adequacy

The estimated coefficient of x in the logit model g(x) = -2.756 + 4.1306x provides a useful measure of association between the response variable and the independent variable. This value can be used to calculate the odds ratio that approximates how much more likely (or unlikely) for Y=1 to exist when X=1 as compared to observations with X=0. The odds ratio is:

$$\widehat{OR} = e^{4.1306} = 62.2215$$

In the context of this study, where Y denotes credit approval or denial and X denotes the presence of a nominal variable A9, the odds of credit being approved (y=1) is approximately 62 times greater when the variable A9 is equal to "t". This calculation is also provided in the SAS output as shown below.

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

Effect	Unit	Estimate	95% Confidence Limits	
A9_t	1.0000	62.213	37.062	110.311

Figure 7 - Odds Ratio Estimate

If the design variable A9_t is removed from the regression model, the results would only represent the impact of the constant. In other words, the inclusion of the design variable is a measure of the variable's predictive power.

Assessing the adequacy of a logistic regression model includes evaluating a number of key diagnostics. The model fit statistics provide useful measures to compare competing models. AIC and SC adjust for the number of predictors and the preferred model has a lower value.

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

Figure 8 - Model Fit Statistics

Testing the null hypothesis (beta=0) is evaluated by multiple methods as shown in Figure 9. The hypothesis is rejected given that $p < 0.05$ for all tests. This confirms there is a relationship between the response variable and the predictor variable. The global chi-square addresses the question, "Is this model better than nothing?" (Allison, 2012)

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

Figure 9 -Null hypothesis tests

There are also four diagnostics that can be used to assess the model's predictive power. These statistics use pairs of observations to calculate an association to response variable. In this model there are 105,672 pairs of observations in which one observation has Y=1 and one observation has Y=0. Each pair is then evaluated to determine if the observation with y=1 has a higher predicted value (probability) than the observation with y=0. If the condition holds that the observation with the higher predicted probability is consistent where the observed value is also Y=1 then this pair is concordant. Using the optimal model measures in Figure 10 and odds ratio, the predicted value of credit approval is higher when A9="t" and in 75.2% of the cases (653) there is consistency in which credit was approved when A9="t" than not. In 23.6% of the cases the observations had the same predicted value and there was only 1.2% of the cases where the model predicted a probability higher for the case where Y=0 than where Y=1. The Somer's D, Gamma and Tau-a are calculated from the concordant, discordant and tied pairs. They are measures of predictive power where larger values correspond to stronger associations between the predicted and observed values (Allison, 2012).

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

Figure 10 - Measures of Predictive Power

Part 3 – Model Assessment using the ROC Curve

In this part of the study a ROC curve is produced to evaluate the predictive power of the optimal model. The two cut points on the ROC curve represent the probability where there is an overall balance of events correctly predicted and non-events correctly predicted. The Y axis (sensitivity) is the true positive rate and the X axis represents the false positive rate. The cut point at 0.80 represents the probability that is highest point where there is the optimal balance between true positives and false positives. This model is produces 94% true positives and 6% false positives based on the ROC output in Figure 12.

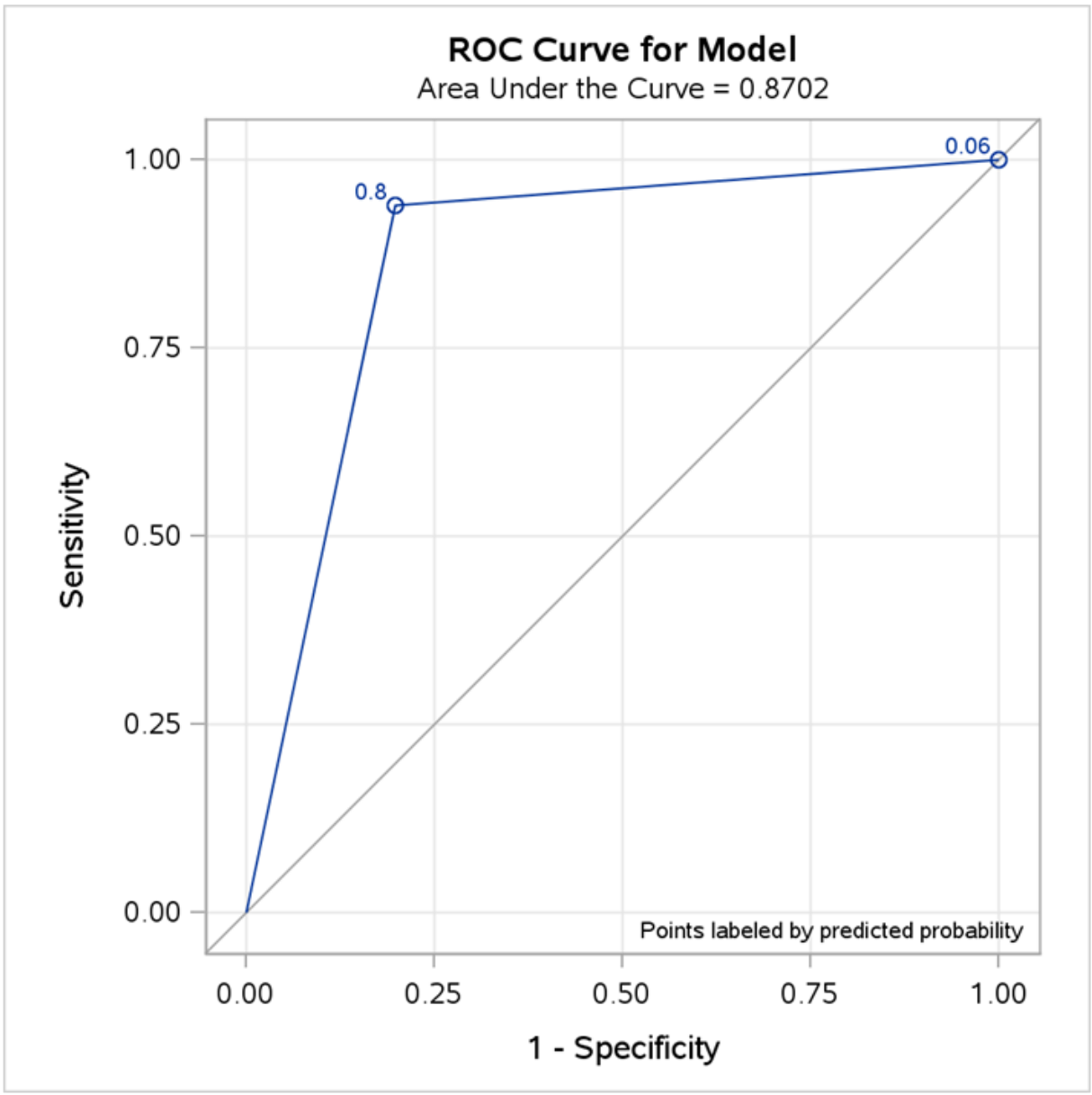
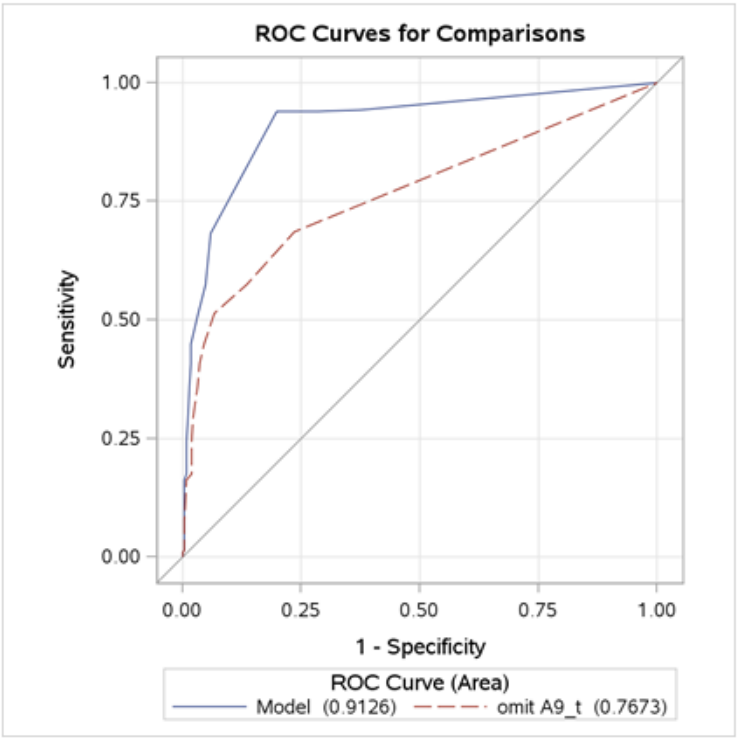


Figure 11 - ROC Curve for A9_t

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

Figure 12 - ROC Plot output

Another ROC plot was generated to compare the optimal model (A9_t) with alternate model with a second variable A11 (blue line). The preferred model is the alternate model with the additional predictor as it has a larger cumulative area under the curve. Selecting the preferred model by using this method is not always a hard and fast rule. There are two types of errors and the tradeoffs between these need to be evaluated given the context of the problem. A false positive exists when the model predicts an event when it does not exist. A false negative exists when the model does not predict an event when in fact is the event exists. A model that generates more false positives would be preferred to a model that generates false negatives for detect a harmful disease.



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 A9_t	0.7673	0.0175	0.7330	0.8015	0.5345	0.7211	0.2653

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Reference = Model	1	105.3415	<.0001

ROC Contrast Estimation and Testing Results by Row						
Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
Model - omit A9_t	0.1454	0.0142	0.1176	0.1731	105.3415	<.0001

Figure 13 - Multiple ROC Curves

Conclusions:

In Part 1 on this study we demonstrated how to conduct an EDA to evaluate the credit_approval data set and select a single predictor for developing an optimal logistic regression model. The continuous variables were discretized to identify non-linear relationships and create contingency tables. Given that categorical variables cannot be combined with continuous variables in a logistic regression model, the values were transformed into a collection of design variables. Finally, mean distributions were calculated for each categorical variables and the A9 variable was identified as the most extreme case that showed predictive power.

In Part 2, a single logistic regression model was fit using the A9 variable (A9_t design variable) as the optimal model. A second approach using automated selection (score) was employed to find the best single variable logistic model and A9_t was identified as the top predictor. This confirmed that the EDA and the score selection method selected the same model. The optimal found that credit_approval is over 62 times more likely when the A9 =t when compared to A9=f (the meaning of A9 is not known).

In Part 3, ROC curves were used to evaluate predictive power of the model and the tradeoff of false positive errors (false alarms) with false negatives. A alternate logistic regression model was fit using a second predictor and it was found to preferred given its larger area under the ROC curve. One should also evaluate what error type of errors are preferred when selected a model. False alarms (false positives) may be preferred to undetected events (false negative) or the reverse depending on the context of the study.

Code:

```
/*  James Gray
        2013.07.26
        graymatter@u.northwestern.edu
        Assignment5_JG.sas
*/

/*      This code is for PREDICT 410 Assignment #5 - Binary Response EDA */

********************************************************************************************;
* Get the data on the SAS server - mydata.credit_approval -

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

********************************************************************************************;
* Review credit_approval dataset metadata and 5 observations;
********************************************************************************************;
proc contents data=mydata.credit_approval; run; quit;
proc print data=mydata.credit_approval(obs=5); run; quit;

********************************************************************************************;
* EDA - create a binary response variable Y by classifying the A16 var
```

```
*************************************************************************************************;
data temp;
	set mydata.credit_approval;

	if (A16='+') then Y=1;
		else Y=0;
run;

/* proc print data=temp;run; this was used to assess data quality for missing values */

*************************************************************************************************;
* EDA - evaluate counts of categorical variables using FREQ
*************************************************************************************************;
proc freq data = temp;
	tables A1 A4 A5 A6 A7 A9 A12 A13 A16;
run;
*************************************************************************************************;
* EDA - evaluate continuous predictor variables
*************************************************************************************************;
proc means data = temp p5 p10 p25 p50 p75 p90 p95;
	class Y; * produce stats for each unique value of Y;
	var A2 A3 A8 A11 A14 A15;
run;

*************************************************************************************************;
* EDA - categorize continuous variables
*************************************************************************************************;
proc print data=temp(obs=5); run;

data temp2;
	set temp;

	if (A2 < 10) then A2_discrete=1;
		else if (A2 < 20) then A2_discrete=2;
		else if (A2 < 30) then A2_discrete=3;
		else if (A2 < 40) then A2_discrete=4;
		else if (A2 < 50) then A2_discrete=5;
		else A2_discrete=6;

	if (A3 < 10) then A3_discrete=1;
		else if (A3 < 0.5) then A3_discrete=2;
		else if (A3 < 2) then A3_discrete=3;
		else if (A3 < 10) then A3_discrete=4;
		else if (A3 < 13) then A3_discrete=5;
		else A3_discrete=6;

	if (A8 < 0.5) then A8_discrete=1;
		else if (A8 < 1) then A8_discrete=2;
		else if (A8 < 3) then A8_discrete=3;
		else if (A8 < 7) then A8_discrete=4;
		else if (A8 < 10) then A8_discrete=5;
		else A8_discrete=6;
```

```
        if (A11 < 3) then A11_discrete=1;
                else if (A11 < 4) then A11_discrete=2;
                else if (A11 < 8) then A11_discrete=3;
                else if (A11 < 10) then A11_discrete=4;
                else A11_discrete=5;

        if (A14 < 105) then A14_discrete=1;
                else if (A14 < 170) then A14_discrete=2;
                else if (A14 < 300) then A14_discrete=3;
                else if (A14 < 400) then A14_discrete=4;
                else A14_discrete=5;

        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;

/* show cross categorization of Y by values of levels of A15;
proc freq data = temp2;
        tables Y*A15_discrete;
run;
*/
        * create design vars for A1 with 2 categories (a,b);

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

        * create design vars for A4 with 3 categories (l,u,y);

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

        * create design vars for A5 with 3 categories (g,gg,p);

        if (A5='gg') then A5_gg=1; else A5_gg=0;
        if (A5='p') then A5_p=1; else A5_p=0;

        * create design vars for A6 with 14 categories (?,aa,c,cc,d,e,ff,i,j,k,m,q,r,w,s);

        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='r') then A6_r=1; else A6_r=0;
        if (A6='w') then A6_w=1; else A6_w=0;
        if (A6='s') then A6_s=1; else A6_s=0;
```

```
        * create design vars for A7 with 9 categories (?,bb,dd,ff,h,j,n,o,v,z);

        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='o') then A7_o=1; else A7_o=0;
        if (A7='v') then A7_v=1; else A7_v=0;
        if (A7='z') then A7_z=1; else A7_z=0;

        * create design vars for A9 with 2 categories (f,t);

        if (A9='t') then A9_t=1; else A9_t=0;

        * create design vars for A12 with 2 categories (f,t);

        if (A12='t') then A12_t=1; else A12_t=0;

        * create design vars for A13 with 3 categories (g,p,s);

        if (A13='p') then A13_p=1; else A13_p=0;
        if (A13='s') then A13_s=1; else A13_s=0;

        * delete missing values - categorical="?", continuous="-";
        if (A1='?') then delete;
                else if (A2='-') then delete;
                else if (A3='-') then delete;
                else if (A4='?') then delete;
                else if (A5='?') then delete;
                else if (A6='?') then delete;
                else if (A7='?') then delete;
                else if (A8='-') then delete;
                else if (A9='?') then delete;
                else if (A10='?') then delete;
                else if (A11='-') then delete;
                else if (A12='?') then delete;
                else if (A13='?') then delete;
                else if (A14='-') then delete;
                else if (A15='-') then delete;
run;


/*proc print data=temp2(obs=5); run; */

*************************************************************************************;
* EDA - Contingency table for discretization of continuous var A15
*************************************************************************************;
* show cross categorization of Y by values of levels of A15;
proc freq data = temp2;
        tables Y*A15_discrete;
run;
```

```
*************************************************************************************************;
* EDA - Assess predictive accuracy of categorical
*************************************************************************************************;

%macro class_mean(c);
proc means data=temp2 mean;
*class A1 A4 A5 A6 A7 A9 A10 A12 A13;
class &c. ;
var Y;
run;
%mend class_mean;

* execute macro;
%class_mean (C=A1);
%class_mean (C=A2_discrete);
%class_mean (C=A3_discrete);
%class_mean (C=A4);
%class_mean (C=A5);
%class_mean (C=A6);
%class_mean (C=A7);
%class_mean (C=A8_discrete);
%class_mean (C=A9);
%class_mean (C=A10);
%class_mean (C=A11_discrete);
%class_mean (C=A12);
%class_mean (C=A13);
%class_mean (C=A14_discrete);
%class_mean (C=A15_discrete);


*************************************************************************************************;
* Single variable logistic regression model using var A9_t selected by EDA.
* In the scenario below, the model will predict Y=1.
*************************************************************************************************;
title 'Logistic regression with one categorical predictor from EDA (A9_t)';
proc logistic data=temp2;
        model Y (event='1') = A9_t / clodds=pl;
run;

*************************************************************************************************;
* Single variable selection logistic regression model using SCORE
*************************************************************************************************;
title 'Logistic regression using variable selection';
proc logistic data=temp2;
        model Y (event='1') = A2 A3 A8 A11 A14 A15
        A1_b
        A4_u A4_y
        A5_gg A5_p
        A6_c A6_cc A6_d A6_e A6_ff A6_i A6_j A6_k A6_m A6_q A6_r A6_w A6_s
        A7_dd A7_ff A7_h A7_j A7_n A7_o A7_v A7_z
        A9_t
        A12_t
```

```
        A13_p A13_s
        / selection=score start=1 stop=1 ;
run;
quit;

*************************************************************************************************;
* Model assessment using the ROC curve;
*************************************************************************************************;
title 'ROC curve for optimal model';
ods graphics on;
proc logistic data=temp2 descending plots(only)=roc(id=prob);
        model Y = A9_t / outroc=roc1;
run;
ods graphics off;

proc print data=roc1; run;quit;

title 'ROC curve for optimal model and alternate model';
ods graphics on;
proc logistic data=temp2;
        model Y (event='1') = A9_t A11;
        roc 'omit A9_t' A11;
        roccontrast / estimate=allpairs;
run;
ods graphics off;


*************************************************************************************************;
* END
*************************************************************************************************;
```

References:

Allison, P. (2012). *Logistic regression using sas: theory and application*. (2nd ed., p. 63). Cary: SAS Publishing.