**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.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

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.

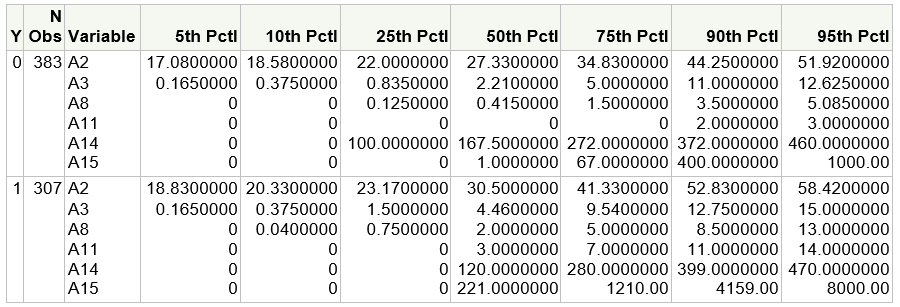


Figure 2 - Percentile distributions for each continuous variable in relation to the binary response

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.

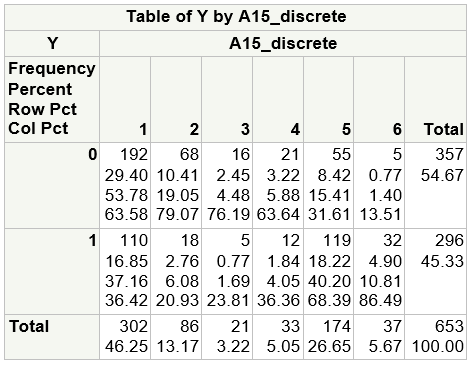


Figure 3 - Contingency Table for Y by A15\_discrete

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.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

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

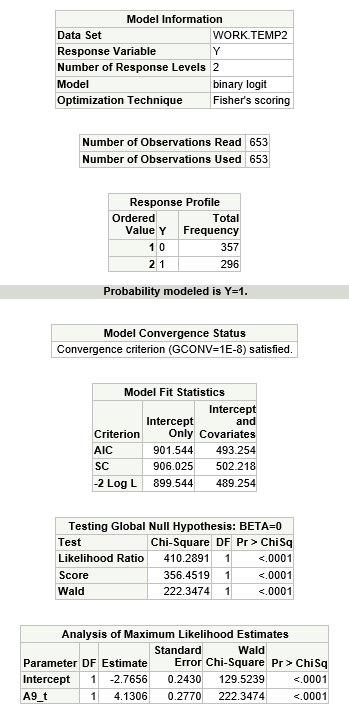


Figure 5 - Fitted Logistic Regression Model using A9\_t

#### 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.

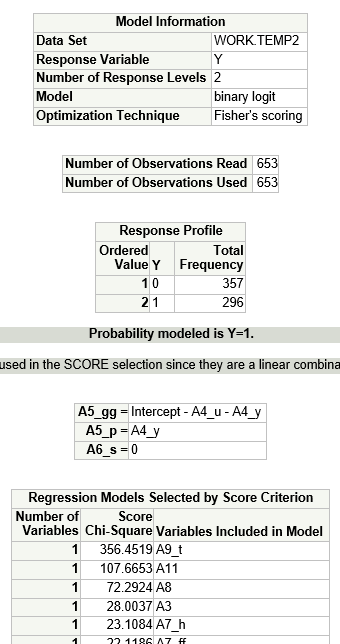


Figure 6 - Fitted Regression Model using the SCORE selection

#### 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:

= 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.

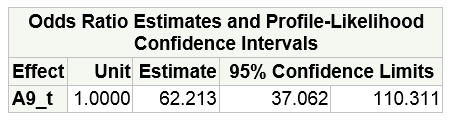


Figure 7 - Odds Ratio Estimate

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.

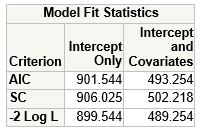


Figure 8 - Model Fit Statistics

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)

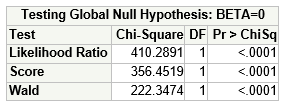


Figure 9 -Null hypothesis tests

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).

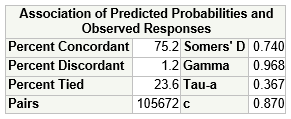


Figure 10 - Measures of Predictive Power

### 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.

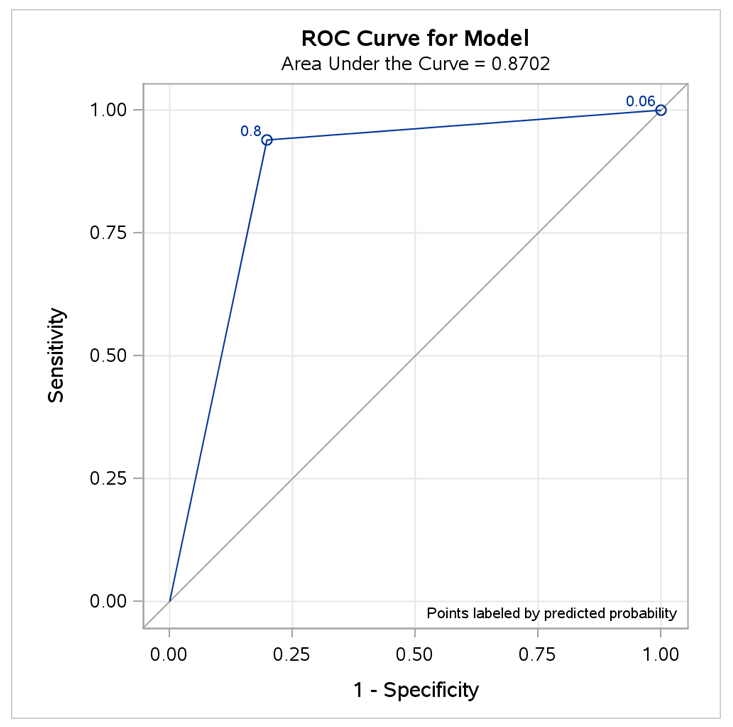


Figure 11 - ROC Curve for A9\_t

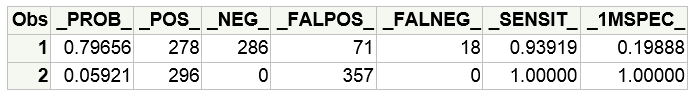


Figure 12 - ROC Plot output

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.

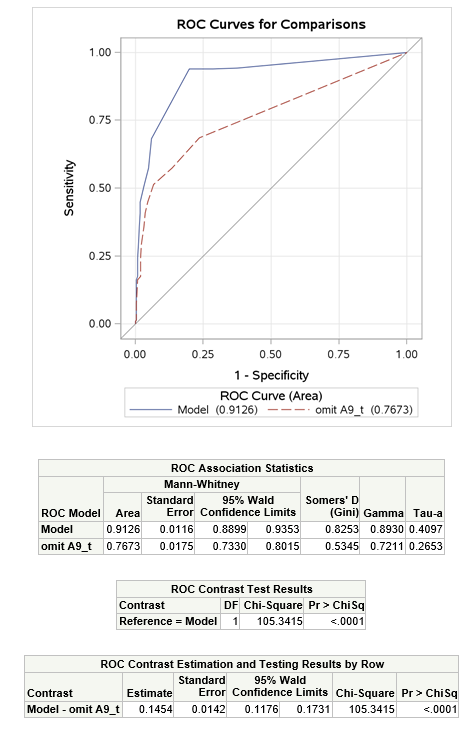


Figure 13 - Multiple ROC Curves

**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.