**Assignment #5**

Daniel S Prusinski

**Introduction:**

Logistic regression utilizes maximum likelihood to predict outcomes. This assignment analyzes 16 different variables, of which 6 are continuous and 10 are categorical level data. The overall goal is to find the best single variable to predict a positive (+) response for variable A16. Through utilizing SAS methods, I was able to conduct a relatively thorough EDA.

**Results Part 1:**

The variables in the data set range from continuous to categorical. Specifically, there are 6 continuous variables. To better understand the continuous variables they were discretized and analyzed against the response variable, coded Y, using the PROC MEANS statement in SAS. Each continuous variable was first discretized by its quantiles ranging from 5, 10, 25, 50, 75, 90 , and 95 percent, and then further discretized in an effort to create harsher cut points. The corresponding quantiles are cataloged as 1 equals the fifth percentile quartile, 2 equals the tenth percentile quartile, and so forth. Before analyzing the predictive accuracy of the attributes, explaining the response variables is necessary. A16 is the response variable and has been coded such that a “+” sign equals 1, thus I am assessing the predicting probability of 1.

Variable A2 discretized has desirable predictive quartiles. The vast majority of the observations fall within the 3,4, and 5 which follows a relative normal distribution. Before discretizing variable A2, it was negatively skewed to the left. Analyzing the mean, one can see that as the discretization of the variables increases the predictive probability also increases. It should be noted that the 95th percentile has the greatest predictive probability, but it also has the least amount of observations.

| **Analysis Variable : Y** | | |
| --- | --- | --- |
| **A2\_discrete** | **N Obs** | **Mean** |
| 1 | 52 | 0.2307692 |
| 2 | 41 | 0.3902439 |
| 3 | 100 | 0.4300000 |
| 4 | 173 | 0.4277457 |
| 5 | 166 | 0.4457831 |
| 6 | 78 | 0.6025641 |
| 7 | 43 | 0.6976744 |

| **Analysis Variable : Y** | | |
| --- | --- | --- |
| **A3\_discrete** | **N Obs** | **Mean** |
| 1 | 32 | 0.4375000 |
| 2 | 38 | 0.3947368 |
| 3 | 134 | 0.3059701 |
| 4 | 201 | 0.3880597 |
| 5 | 121 | 0.6115702 |
| 6 | 81 | 0.5432099 |
| 7 | 46 | 0.6521739 |

| **Analysis Variable : Y** | | |
| --- | --- | --- |
| **A8\_discrete** | **N Obs** | **Mean** |
| 3 | 284 | 0.2464789 |
| 4 | 147 | 0.5102041 |
| 5 | 125 | 0.6000000 |
| 6 | 59 | 0.7627119 |
| 7 | 38 | 0.8157895 |

Discretized A3 does not have the systematic increase in predictive probability as its quartiles increase. This variable also follows a rather normal distribution. The fifth quartile has a predictive probability of .61 and has a large amount of observations. Perhaps a further discretization of this catalog would enhance the predictive probability. A8 discrete has a positive skew. Over 85% of the observations fall below the 6th category. It can be seen that as the categories increase the predictive probability does as well. But, the last two quartiles have very few observations which could influence the statistical strength of the variable. A11 is very similar to A8 in that it is positively skewed. The vast majority of the observations are below the 50th percentile. There is a very strong predictive probability that as the categories increase so too does the predictive probability. Out of all the variables A11 has the strongest predictive probability. A14 has a strong positive skew, and the third category has a rather strong predictive probability. Given that there are many observations in this quartile, I would want to further divide this quartile to hone in on a stronger probability predictor quartile.

| **Analysis Variable : Y** | | |
| --- | --- | --- |
| **A11\_discrete** | **N Obs** | **Mean** |
| 1 | 366 | 0.2540984 |
| 2 | 69 | 0.4782609 |
| 3 | 42 | 0.4285714 |
| 4 | 176 | 0.8636364 |

| **Analysis Variable : Y** | | |
| --- | --- | --- |
| **A14\_discrete** | **N Obs** | **Mean** |
| 3 | 206 | 0.6213592 |
| 4 | 156 | 0.3269231 |
| 5 | 127 | 0.3149606 |
| 6 | 90 | 0.3888889 |
| 7 | 74 | 0.5675676 |

| **Analysis Variable : Y** | | |
| --- | --- | --- |
| **A15\_discrete** | **N Obs** | **Mean** |
| 4 | 275 | 0.3927273 |
| 5 | 123 | 0.1869919 |
| 6 | 91 | 0.4065934 |
| 7 | 164 | 0.7804878 |

A15 is strongly skewed to the left, and it predictive probability does not incrementally increase or decrease. Rather, the last category, 7, has a large number of observations with a strong predictive probability. Further EDA would lead me to divide category seven into a few more categories. A6 is the categorical variable that I decided to analyze the quartile distribution. Given the many different categories, I was hoping to find a strong predictive probability. This variable is one of the few that has a negative skew, and the 1st category has a strong predictive probability. The low number of observations leads me doubt the statistical significance. Out of all the categorical variables A9 had the strongest amount of observations producing the highest mean. This variable could be one of the strongest. Using Proc Means with a class statement has lead me to fit either A11, A9, or A15 given the strong predictive probability and number of observations within the category.

| **Analysis Variable : Y** | | |
| --- | --- | --- |
| **A9\_t** | **N Obs** | **Mean** |
| 0 | 304 | 0.0592105 |
| 1 | 349 | 0.7965616 |

| **Analysis Variable : Y** | | |
| --- | --- | --- |
| **A6\_discrete** | **N Obs** | **Mean** |
| 1 | 36 | 0.8333333 |
| 3 | 52 | 0.3653846 |
| 4 | 223 | 0.4932735 |
| 5 | 115 | 0.2086957 |
| 6 | 86 | 0.3372093 |
| 7 | 141 | 0.5957447 |

**Results Part 2:**

Utilizing PROC LOGISTIC with *start=1* and *stop=1* along with *selection=score*, the variables were ranked. The output below shows that variable A9 has the largest Score of Chi-Square. This test communicates that variable A9s coefficient is least likely to be zero.

| **Regression Models Selected by Score Criterion** | | |
| --- | --- | --- |
| **Number of Variables** | **Score Chi-Square** | **Variables Included in Model** |
| **1** | 349.7407 | A9\_t |
| **1** | 124.6199 | A10\_t |
| **1** | 102.8407 | A11 |
| **1** | 68.5069 | A8 |
| **1** | 26.5124 | A15 |
| **1** | 24.5293 | A7\_ff |
| **1** | 24.3184 | A2 |
| **1** | 23.8778 | A7\_h |
| **1** | 21.5046 | A4\_u |
| **1** | 21.3463 | A3 |
| **1** | 11.9678 | A6\_q |
| **1** | 5.4414 | A14 |
| **1** | 4.5695 | A6\_k |
| **1** | 2.3072 | A12\_t |
| **1** | 1.5869 | A7\_v |
| **1** | 0.8825 | A6\_m |
| **1** | 0.8600 | A6\_w |
| **1** | 0.0818 | A1\_a |
| **1** | 0.0000 | A7\_bb |

Based on my EDA utilizing the Proc Means statement, I chose to fit variable A11 because it had specific categories that had large predictive probabilities. I did not choose the optimal model. The best single variable logistic regression model using the *selection=score* option in PROC LOGISTIC was variable A9. I will compare the two models and assess the model adequacy between the two variables. Analyzing the inferential statistics AIC, SC, and -2 Log L are informal methods to assess the model fit. All of these statistics can be used to compare different sets of variables. Higher values for these statistics mean a worse fit to the data. It can clearly be seen that variable A9 has a lower values for all three statistics.

| **A11 Model Fit Statistics** | | |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 901.544 | 715.985 |
| **SC** | 906.025 | 733.911 |
| **-2 Log L** | 899.544 | 707.985 |

| **Model Fit Statistics A9** | | |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 901.544 | 493.254 |
| **SC** | 906.025 | 502.218 |
| **-2 Log L** | 899.544 | 489.254 |

The Global Null hypothesis tests that all the explanatory variables have coefficients equal to zero. It can be seen that both variables have at least one coefficient that does not equal zero. Both models also have a significant p-value. A9 has much higher scores, and it should be noted that this variable was considered the best based on its Score of 356.4519.

| **A11 Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 191.5589 | 3 | <.0001 |
| **Score** | 178.4621 | 3 | <.0001 |
| **Wald** | 137.8205 | 3 | <.0001 |

| **A9 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 |

Both models have coefficients that are statistically significant. A9 has a much higher Wald Chi-Square, but all the coefficients have a low enough p-value to statistically warrant using each one. In regression analysis, choosing a model based on parsimony is desired. While both models have intercepts statistically significant, A9 has fewer variables and is more desirable to use. Through analyzing these statistical outputs one can assess the goodness of-fit for each model as well as compare different models.

| **Analysis of Maximum Likelihood Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** |  | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** |  | 1 | -1.0769 | 0.1201 | 80.4439 | <.0001 |
| **A11\_discrete** | **2** | 1 | 0.9899 | 0.2693 | 13.5155 | 0.0002 |
| **A11\_discrete** | **3** | 1 | 0.7892 | 0.3341 | 5.5789 | 0.0182 |
| **A11\_discrete** | **4** | 1 | 2.9227 | 0.2503 | 136.3229 | <.0001 |

| **A9 Analysis of Maximum Likelihood Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** |  | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** |  | 1 | -1.3649 | 0.1330 | 105.3672 | <.0001 |
| **A9\_a** | **1** | 1 | 4.1306 | 0.2770 | 222.3474 | <.0001 |

After assessing the goodness of fit, it is desirable to analyze the statistics that measure the predictive power of specific variables. The approaches used in SAS from the PROC LOGISTIC command are ordinal measures of association that produce model-free measures of predictive power. The percent concordant mean is interpreted as a pair of observations with different responses, and the observation with the lower ordered response value has a lower predicted mean score than the observation with the higher ordered response value (UCLA.edu). Percent discordant is the opposite of concordant, and if there is a tie it is displayed in the third row. Somer’s, Gamma, Tau-a, and C are all calculations based on the percent calculations. Higher scores relate to greater predictive power. The C calculation is desirable in that it relates to the ROC curve, and the Tau-a is most closely associated to the R squared of linear regression. Comparing the two models, both have mixed calculations, the C value will directly relate to the ROC curve. Again, A9 has similar values to A11, but it is more desirable based on parsimony and goodness-of-fit values.

| **A11 Association of Predicted Probabilities and Observed Responses** | | | |
| --- | --- | --- | --- |
| **Percent Concordant** | 61.8 | **Somers' D** | 0.527 |
| **Percent Discordant** | 9.2 | **Gamma** | 0.742 |
| **Percent Tied** | 29.0 | **Tau-a** | 0.261 |
| **Pairs** | 105672 | **c** | 0.763 |

| **A9 Association of Predicted Probabilities and Observed Responses** | | | |
| --- | --- | --- | --- |
| **Percent Concordant** | 49.4 | **Somers' D** | 0.456 |
| **Percent Discordant** | 3.9 | **Gamma** | 0.855 |
| **Percent Tied** | 46.7 | **Tau-a** | 0.226 |
| **Pairs** | 105672 | **c** | 0.728 |

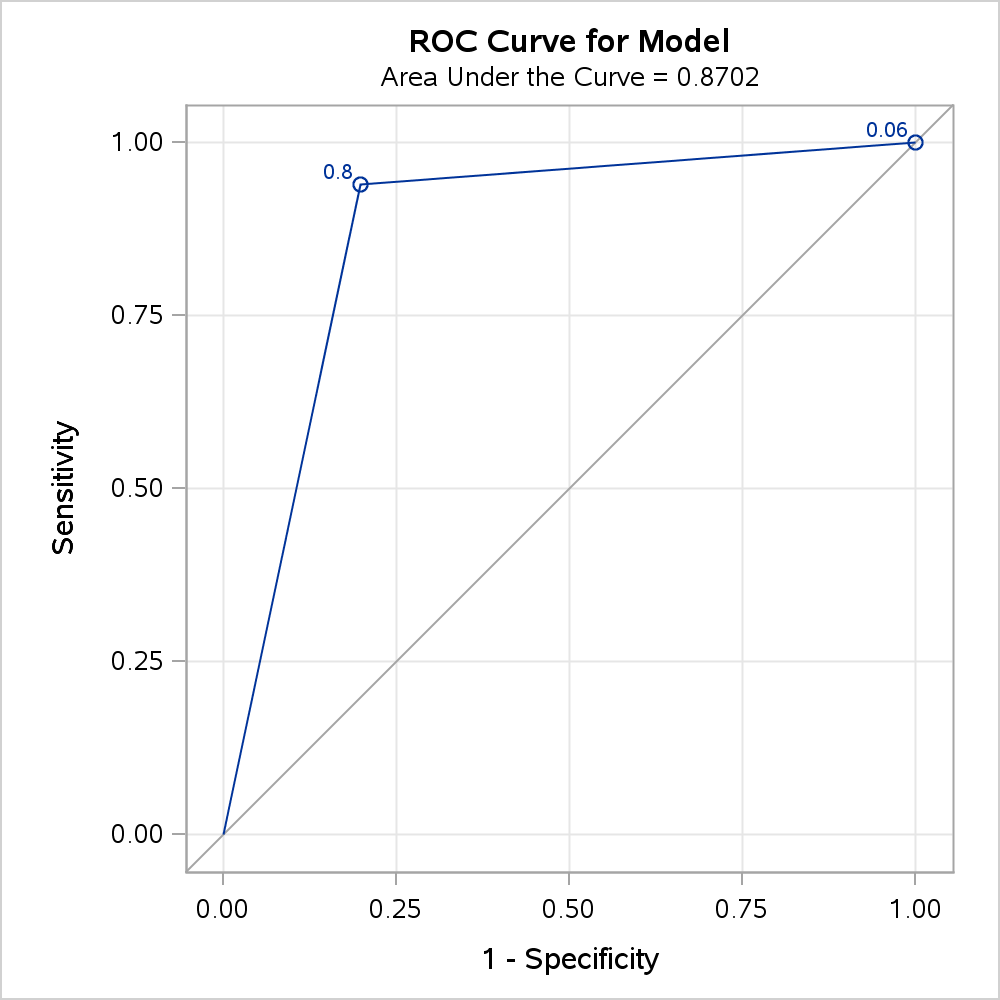
Model adequacy is assessed through analyzing the goodness-of-fit statistics along with the statistics that measure predictive power. In conducting an EDA, there may be reasons that delineate using a certain model based on specific inferential statistics. The estimated coefficients for dummy variables are interpreted as the log odds increase by x for every 1 unit increase in the explanatory variable. This is extremely hard to understand because the logistic model assumes a non-linear relationship. Rather, assessing the odds ratio estimates for the coefficients is a better way to understand the importance of a specific variable. The greater the odds the more significant the coefficient, but the p-value for the Chi Square needs to be statistically significant in order to warrant using the coefficient and the corresponding odds. A dummy variable is dropped from a model when it is the same as another variable.

**Results Part 3:**

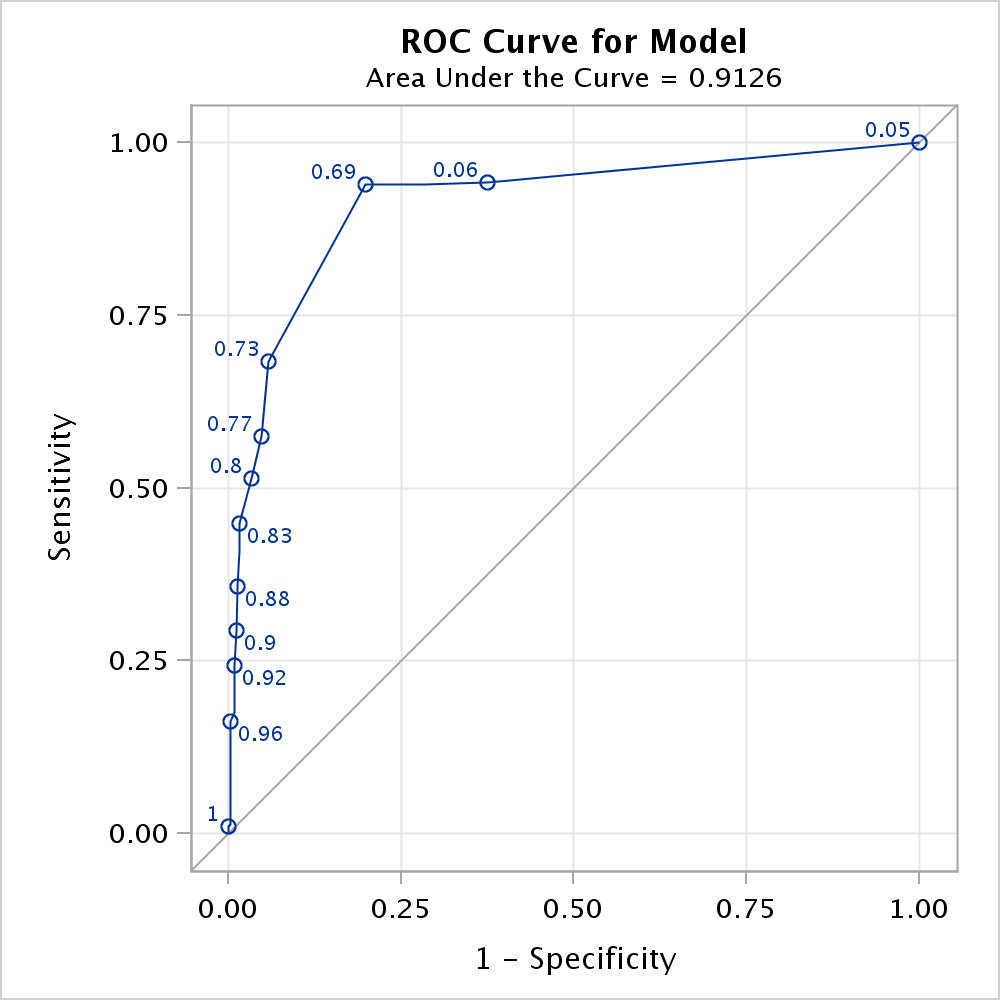
Cut-points are used to classify test results as positive. Sensitivity and specificity are terms used to discern how accurate certain cut-points are when analyzing a dichotomous variable. Ideally for classification purposes, the optimal cut-point is where specificity and sensitivity are maximized. The variable A9 produces two cut-points, as displayed below. The probabilities between these two cut-points discriminate the optimum experience for the outcome of interest. In other words the area under the ROC curve provides a probability that that 9\_t will be 1 versus 0.

| **Classification Table** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Prob Level** | **Correct** | | **Incorrect** | | **Percentages** | | | | |
| **Event** | **Non- Event** | **Event** | **Non- Event** | **Correct** | **Sensi- tivity** | **Speci- ficity** | **False POS** | **False NEG** |
| **0.001** | 296 | 0 | 357 | 0 | 45.3 | 100.0 | 0.0 | 54.7 | . |
| **0.050** | 296 | 0 | 357 | 0 | 45.3 | 100.0 | 0.0 | 54.7 | . |
| **0.059** | 278 | 0 | 357 | 18 | 42.6 | 93.9 | 0.0 | 56.2 | 100.0 |
| **0.060** | 278 | 286 | 71 | 18 | 86.4 | 93.9 | 80.1 | 20.3 | 5.9 |
| **0.100** | 278 | 286 | 71 | 18 | 86.4 | 93.9 | 80.1 | 20.3 | 5.9 |
| **0.500** | 278 | 286 | 71 | 18 | 86.4 | 93.9 | 80.1 | 20.3 | 5.9 |
| **0.550** | 278 | 286 | 71 | 18 | 86.4 | 93.9 | 80.1 | 20.3 | 5.9 |
| **0.750** | 278 | 286 | 71 | 18 | 86.4 | 93.9 | 80.1 | 20.3 | 5.9 |
| **0.780** | 278 | 286 | 71 | 18 | 86.4 | 93.9 | 80.1 | 20.3 | 5.9 |
| **0.850** | 0 | 357 | 0 | 296 | 54.7 | 0.0 | 100.0 | . | 45.3 |
| **0.900** | 0 | 357 | 0 | 296 | 54.7 | 0.0 | 100.0 | . | 45.3 |
| **0.999** | 0 | 357 | 0 | 296 | 54.7 | 0.0 | 100.0 | . | 45.3 |

| **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 |

****

ROC curves can be used to compare different models. For example, using variables A11 and A9\_t produces the following ROC curve and results. This ROC curve has a .91 discrimination probability. This is larger than A9, but in my opinion both models have great discrimination probabilities.



Analyzing the cut-points for the model A9 and A11 produces the following results. Utilizing a cut-point with .06 probability will maximize the sensitivity and specialty values. Depending on the desired model specificities, different models are desired for different situations.

| **Obs** | **\_PROB\_** | **\_POS\_** | **\_NEG\_** | **\_FALPOS\_** | **\_FALNEG\_** | **\_SENSIT\_** | **\_1MSPEC\_** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 1.00000 | 1 | 357 | 0 | 295 | 0.00338 | 0.00000 |
| **2** | 0.99989 | 2 | 357 | 0 | 294 | 0.00676 | 0.00000 |
| **3** | 0.99615 | 3 | 357 | 0 | 293 | 0.01014 | 0.00000 |
| **4** | 0.99285 | 4 | 356 | 1 | 292 | 0.01351 | 0.00280 |
| **5** | 0.99121 | 5 | 356 | 1 | 291 | 0.01689 | 0.00280 |
| **6** | 0.98675 | 7 | 356 | 1 | 289 | 0.02365 | 0.00280 |
| **7** | 0.98374 | 10 | 356 | 1 | 286 | 0.03378 | 0.00280 |
| **8** | 0.98006 | 14 | 356 | 1 | 282 | 0.04730 | 0.00280 |
| **9** | 0.97557 | 22 | 356 | 1 | 274 | 0.07432 | 0.00280 |
| **10** | 0.97011 | 23 | 356 | 1 | 273 | 0.07770 | 0.00280 |
| **11** | 0.96346 | 30 | 356 | 1 | 266 | 0.10135 | 0.00280 |
| **12** | 0.95540 | 48 | 356 | 1 | 248 | 0.16216 | 0.00280 |
| **13** | 0.94567 | 52 | 354 | 3 | 244 | 0.17568 | 0.00840 |
| **14** | 0.93396 | 62 | 354 | 3 | 234 | 0.20946 | 0.00840 |
| **15** | 0.91994 | 72 | 354 | 3 | 224 | 0.24324 | 0.00840 |
| **16** | 0.90325 | 87 | 353 | 4 | 209 | 0.29392 | 0.01120 |
| **17** | 0.88352 | 106 | 352 | 5 | 190 | 0.35811 | 0.01401 |
| **18** | 0.86039 | 121 | 351 | 6 | 175 | 0.40878 | 0.01681 |
| **19** | 0.83354 | 133 | 351 | 6 | 163 | 0.44932 | 0.01681 |
| **20** | 0.80270 | 152 | 345 | 12 | 144 | 0.51351 | 0.03361 |
| **21** | 0.76774 | 170 | 340 | 17 | 126 | 0.57432 | 0.04762 |
| **22** | 0.72868 | 202 | 336 | 21 | 94 | 0.68243 | 0.05882 |
| **23** | 0.68574 | 278 | 286 | 71 | 18 | 0.93919 | 0.19888 |
| **24** | 0.39348 | 278 | 285 | 72 | 18 | 0.93919 | 0.20168 |
| **25** | 0.34517 | 278 | 284 | 73 | 18 | 0.93919 | 0.20448 |
| **26** | 0.29986 | 278 | 282 | 75 | 18 | 0.93919 | 0.21008 |
| **27** | 0.15728 | 278 | 280 | 77 | 18 | 0.93919 | 0.21569 |
| **28** | 0.13167 | 278 | 279 | 78 | 18 | 0.93919 | 0.21849 |
| **29** | 0.10969 | 278 | 276 | 81 | 18 | 0.93919 | 0.22689 |
| **30** | 0.09100 | 278 | 274 | 83 | 18 | 0.93919 | 0.23249 |
| **31** | 0.07522 | 278 | 255 | 102 | 18 | 0.93919 | 0.28571 |
| **32** | 0.06199 | 279 | 223 | 134 | 17 | 0.94257 | 0.37535 |
| **33** | 0.05096 | 296 | 0 | 357 | 0 | 1.00000 | 1.00000 |

**Conclusion**

Through this assignment, the PROC MEANS, PROC LOGISTIC, and ROC Curve statements helped me to better understand the 15 variables predictive relationship with A16 codded as Y. This assignment felt as though I was drinking from a fire hose given the new output needing to be interpreted and understood as well as the different criteria for assessing a logistic model. I look forward to learning more about interpreting and assessing models in logistic regression.

**Code:**

\*Daniel Prusinski Assignment 5 Version 1\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\*\*\*\*Statement to access where the data is stored\*\*\*\*\*;

**libname** mydata '/courses/u\_northwestern.edu1/i\_833463/c\_3505/SAS\_Data/'**;**

ods graphics on**;**

\*\*\*\*\*Setting the Temp Data\*\*\*\*\*;

**data** temp**;**

set mydata.credit\_approval**;**

if **(**A16='+'**)** then Y =**1;**

else Y=**0;**

**run;**

**proc** **freq** **data**=temp**;**

tables A1 A4 A5 A6 A7 A9 A10 A12 A13 A16**;**

**run;**

**proc** **means** **data**=temp p5 p10 p25 p50 p75 p90 p95**;**

class Y**;**

var A2 A3 A8 A11 A14 A15**;**

**run;**

\*\*\*\*\*Discrete Variables\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**data** temp**;**

set mydata.credit\_approval**;**

if **(**A16='+'**)** then Y =**1;**

else Y=**0;**

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;**

if **(**A2 < **24)** then A2\_discrete=**1;**

else if **(**A2 < **31)** then A2\_discrete=**2;**

else if **(**A2 < **42)** then A2\_discrete=**3;**

else if **(**A2 < **53)** then A2\_discrete=**4;**

else if **(**A2 < **59)** then A2\_discrete=**5;**

else A2\_discrete=**6;**

if **(**A3 < **1.6)** then A3\_discrete=**1;**

else if **(**A3 < **4.5)** then A3\_discrete=**2;**

else if **(**A3 < **9.6)** then A3\_discrete=**3;**

else if **(**A3 < **12)** then A3\_discrete=**4;**

else if **(**A3 < **15)** then A3\_discrete=**5;**

else A3\_discrete=**6;**

if **(**A8 < **1)** then A8\_discrete=**2;**

else if **(**A8 < **3)** then A8\_discrete=**3;**

else if **(**A8 < **5)** then A8\_discrete=**4;**

else A8\_discrete=**6;**

if **(**A11 < **.5)** then A11\_discrete=**1;**

else if **(**A11 < **1.5)** then A11\_discrete=**2;**

else if **(**A11 < **3)** then A11\_discrete=**3;**

else A11\_discrete=**4;**

if **(**A6\_a < **1)** then A6\_discrete=**1;**

else if **(**A6\_a < **1.5)** then A6\_discrete=**2;**

else if **(**A6\_a < **2)** then A6\_discrete=**3;**

else if **(**A6\_a < **6)** then A6\_discrete=**4;**

else if **(**A6\_a < **9)** then A6\_discrete=**5;**

else if **(**A6\_a < **11)** then A6\_discrete=**6;**

else A6\_discrete=**7;**

if **(**A14 < **101)** then A14\_discrete=**1;**

else if **(**A14 < **170)** then A14\_discrete=**2;**

else if **(**A14 < **281)** then A14\_discrete=**3;**

else if **(**A14 < **400)** then A14\_discrete=**4;**

else if **(**A14 < **471)** then A14\_discrete=**5;**

else A14\_discrete=**6;**

\*\*\*\*\*Categorical Variables\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

if **(**A1='a'**)** then A1\_a=**1;** else A1\_a=**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='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;**

\*\*\*\*\*I left off a few of the small variables, I want to see what this does\*\*\*\*\*;

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='t'**)** then A10\_t=**1;** else A10\_t=**0;**

if **(**A12='t'**)** then A12\_t=**1;** else A12\_t=**0;**

if **(**A13='g'**)** then A13\_g=**1;** else A13=\_g=**0;**

\*\*\*\*\*This purges the Data, 90 LSB\*\*\*\*\*;

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**;**

%**macro** class\_mean**(**c**);**

**proc** **means** **data**=temp mean**;**

class &c. **;**

var Y**;**

**Run;**

%**mend** class\_mean**;**

%class\_mean **(**c=A1\_a**);**

%class\_mean **(**c=A2\_discrete**);**

%class\_mean **(**c=A3\_discrete**);**

%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\_d**);**

%class\_mean **(**c=A6\_e**);**

%class\_mean **(**c=A6\_ff**);**

%class\_mean **(**c=A6\_i**);**

%class\_mean **(**c=A6\_j**);**

%class\_mean **(**c=A6\_k**);**

%class\_mean **(**c=A6\_m**);**

%class\_mean **(**c=A6\_q**);**

%class\_mean **(**c=A6\_r**);**

%class\_mean **(**c=A6\_w**);**

%class\_mean **(**c=A7\_bb**);**

%class\_mean **(**c=A7\_ff**);**

%class\_mean **(**c=A7\_h**);**

%class\_mean **(**c=A7\_v**);**

%class\_mean **(**c=A8\_discrete**);**

%class\_mean **(**c=A9\_t**);**

%class\_mean **(**c=A10\_t**);**

%class\_mean **(**c=A11\_discrete**);**

%class\_mean **(**c=A12\_t**);**

%class\_mean **(**c=A13\_g**);**

%class\_mean **(**c=A14\_discrete**);**

%class\_mean **(**c=A15\_discrete**);**

title "Logistic Regression with One Categorical Predictor Variable LRUS p35"**;**

**proc** **logistic** **data**=temp**;**

class A11\_discrete **(**param=ref ref='1'**);**

model Y **(**event= '1'**)** = A11\_discrete /**;**

**run;**

**proc** **logistic** **data** =temp descending**;**

model Y **(**event ='1'**)** = A1\_a A2 A3 A4\_u A5\_g A6\_k A6\_m A6\_q A6\_w A7\_bb A7\_ff A7\_h A7\_v

A8 A9\_t A10\_t A11 A12\_t A13\_g A14 A15 / selection=score start=**1** stop=**1;**

**run;**

**proc** **logistic** **data** =temp**;**

class A9\_t **(**param=ref ref='0'**);**

model Y **(**event ='1'**)** = A9\_t /**;**

**run;**

**proc** **logistic** **data**=temp descending plots**(**only**)**=roc**(**id=prob**);**

model Y = A9\_t / outroc=roc1**;**

**run;**

**proc** **logistic** **data**=temp descending plots**(**only**)**=roc**(**id=prob**);**

model Y = A9\_t A11 / outroc=roc1**;**

**run;**

**proc** **print** **data**=roc1**;**

**run**