# **MAT 303 Module Five Problem Set Report**

Logistic Regression

[Luna Saccoia]

[Luna.Saccoia@snhu.edu]

Southern New Hampshire University

## **1. Introduction**

*The data set that I will be exploring this week is a credit default data set that includes information about the demographics of people including education, age, sex, assets, ect. These variables will be used to determine the likelihood of someone defaulting on their credit line.*

*The type of analysis I will be running are generalized linear models, confusion matrixes, Hosmer-Lemeshow Goodness of Fit (GOF) Test, and Receiver Operating Characteristic (ROC) Curve.*

## **2. Data Preparation**

*The important variables in the data set that will be used in the models are age, sex, education, assets, marriage, missed payments in the last 3 months, credit utilization and whether or not they have defaulted on their credit line.*

*There are 8 columns and 600 rows in the data set.*

## **3. First Logistic Regression Model**

### **Reporting Results**

*The general form equation of a logistic regression model for defaulting on credit, using credit utilization and missed payments as independent variables is:*

*Y is 1 for defaulting on credit and 0 for not defaulting on credit. is credit utilization. is the dummy variable for missing payment in the last 3 months = true.*

*The equation transformed to form a model that is linear in the beta terms is:*

*The left side of the equation above is the natural log of odds, so this can be written as:*

*Where ‘odds’ is the odds of defaulting on the credit line (default = 1).*

*is the probability of an event happening, in this case defaulting on the credit line. is the ratio of the probability of an event happening, where is the probability of it happening.*

*The equation for this logistic regression model is:*

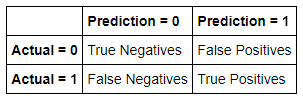
*The equation for this model in terms of natural log of odds is:*

*The estimated coefficient of credit utilization is 31.2090. This means that on average, the change in log odds for defaulting is 0.312090 for each percentage increase in credit utilization, holding all other variables constant. This is found by dividing the coefficient of credit utilization by 100 because credit utilization is expressed as a percentage.*

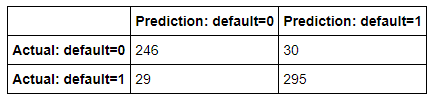
*An alternative way to express this is in terms of odds (and not log odds), would be to calculate:*

*The odds of defaulting increase by 36.63% for each percentage increase in credit utilization, holding all other variables constant. Again, the figure 0.312090 is found by dividing 31.2090 by 100 because credit utilization is expressed as a percentage. The figure 36.63% is obtained by multiplying the decimal value (0.3663) by 100, to turn the decimal into a percentage.*

*The general form table output of a confusion matrix is:*

**

*The output of the confusion matrix for this model is:*

**

*The confusion matrix results are:*

* *True positives: 295*
* *True negatives: 246*
* *False positives: 30*
* *False negatives: 29*
* *True Positive (TP): The actual value is 1 (default = 1) and the predicted value is 1 (default = 1). Hence a true positive.*
* *True Negative (TN): The actual value is 0 (default = 0) and the predicted value is 0 (default = 0). Hence a true negative.*
* *False Positive (FP): The actual value is 0 (default = 0) and the predicted value is 1 (default = 1). Hence a false positive. This is also a Type 1 Error.*
* *False Negative (FN): The actual value is 1 (default = 1) and the predicted value is 0 (default = 0). Hence a false negative. This is also a Type 2 Error.*

***Accuracy****is the ratio of the number of correct predictions to the total number of observations.*

***Precision****is the ratio of correct positive predictions to the total predicted positives.*

***Recall****is the ratio of correct positive predictions to the total positives examples.*

### **Evaluating Model Significance**

*The Hosmer-Lemeshow goodness of fit (GOF) test assesses whether the model predictions are close to the observed values of Y, which are either 0 or 1. In this model it is used to assess if the model fits the data or not.*

*The null and alternative hypotheses are:*

*The test statistic ( is 49.076. The P-value is 0.4298. The level of significance is 5%.*

*The P-value of 0.4298 is higher than the level of significance of 0.05. Thus the null hypothesis should not be rejected. The conclusion is that the model is appropriate for the data set.*

*The null and alternative hypotheses for calculating if credit utilization is significant based on Wald’s test with a 5% level of significance are:*

*is the credit utilization parameter.*

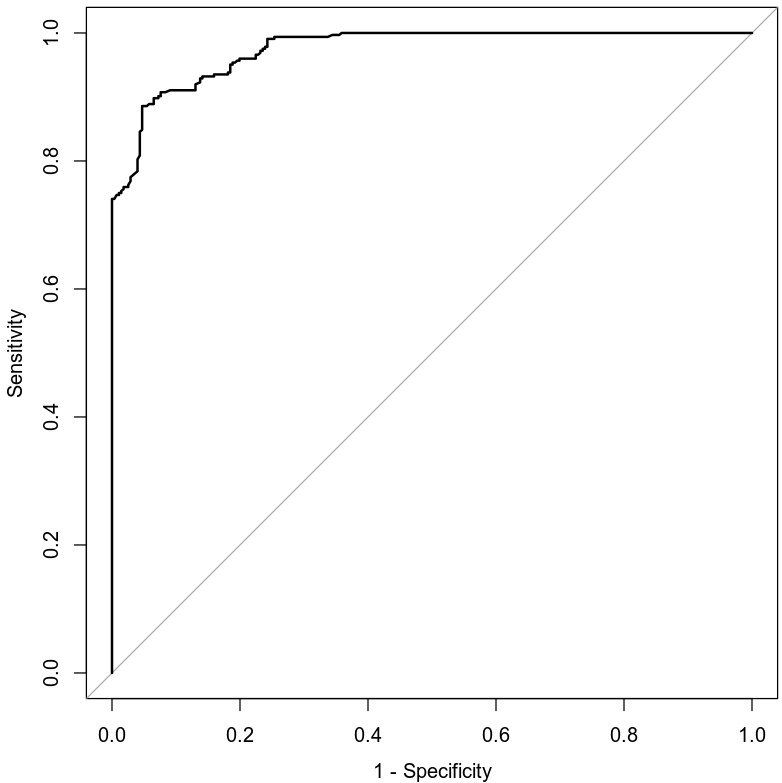
*The null and alternative hypotheses for calculating if missed payment is significant based on Wald’s test with a 5% level of significance are:*

*is the missed payment parameter.*

*The P-value for credit utilization is 2e-16. The P-value for missing payment is 1.16e-05.*

*The P-values for both parameters are well below the level of significance of 0.05. Each parameter, individually, is statistically significant at the level of significance of 0.05.*

*The Receiver Operating Characteristic (ROC) curve is displayed here:*

**

*The area under the curve (AUC) is 0.9746 or 97.46%. This is an indicator of how well the model distinguishes between Y = 0 and Y = 1. In general, the larger the AUC the better, because the larger the area under the curve, the better it is at predicting binary classes.*

### **Making Predictions Using Model**

***Prediction 1***

*Using the model to calculate the probability of an individual who has a credit utilization of 32% and has missed payments in the past three months defaulting on credit, the odds of this event occurring with these parameters is 0.75 or 75%. This means that an individual with a credit utilization of 32% and has missed payments in the last 3 months has a 75% chance of defaulting on their credit line.*

***Prediction 2***

*Using the model to calculate the probability of an individual who has a credit utilization of 32% and has not missed any payments in the past three months defaulting on credit, the odds of this event occurring with these parameters is* *0.4035 or 40.35%. This means that an individual with a credit utilization of 32% that has not missed any payments in the last 3 months has a 40.35% chance of defaulting on their credit line.*

## **4. Second Logistic Regression Model**

### **Reporting Results**

*The general form equation of a logistic regression model for defaulting on credit, using credit utilization, assets and education as independent variables is:*

*Y is 1 for defaulting on credit and 0 for not defaulting on credit. is the variable for credit utilization. are the dummy variables for car, house, and car + house. are the dummy variables for college and post graduate education.*

*The equation transformed to form a model that is linear in the beta terms is:*

*The left side of the equation above is the natural log of odds, so this can be written as:*

*Where ‘odds’ is the odds of defaulting on the credit line (default = 1).*

*The equation for this logistic regression model is:*

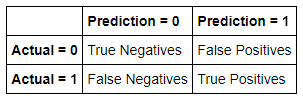
*The equation for this model in terms of natural log of odds is:*

*The estimated coefficient of credit utilization is 36.3201. This means that on average, the change in log odds for defaulting is 0.363201 for each percentage increase in credit utilization, holding all other variables constant. This is found by dividing the coefficient of credit utilization by 100 because credit utilization is expressed as a percentage.*

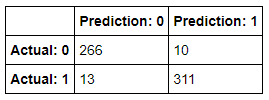
*An alternative way to express this is in terms of odds (and not log odds), would be to calculate:*

*The odds of defaulting increase by 43.79% for each percentage increase in credit utilization, holding all other variables constant. Again, the figure 0.363201 is found by dividing 36.3201 by 100 because credit utilization is expressed as a percentage. The figure 43.79% is obtained by multiplying the decimal value (0.4379) by 100, to turn the decimal into a percentage.*

*The general form table output of a confusion matrix is:*

**

*The output of the confusion matrix for this model is:*

**

*The confusion matrix results are:*

* *True positives: 311*
* *True negatives: 266*
* *False positives: 10*
* *False negatives: 13*
* *True Positive (TP): The actual value is 1 (default = 1) and the predicted value is 1 (default = 1). Hence a true positive.*
* *True Negative (TN): The actual value is 0 (default = 0) and the predicted value is 0 (default = 0). Hence a true negative.*
* *False Positive (FP): The actual value is 0 (default = 0) and the predicted value is 1 (default = 1). Hence a false positive. This is also a Type 1 Error.*
* *False Negative (FN): The actual value is 1 (default = 1) and the predicted value is 0 (default = 0). Hence a false negative. This is also a Type 2 Error.*

***Accuracy****is the ratio of the number of correct predictions to the total number of observations.*

***Precision****is the ratio of correct positive predictions to the total predicted positives.*

***Recall****is the ratio of correct positive predictions to the total positives examples.*

### **Evaluating Model Significance**

*The Hosmer-Lemeshow goodness of fit (GOF) test assesses whether the model predictions are close to the observed values of Y, which are either 0 or 1. In this model it is used to assess if the model fits the data or not.*

*The null and alternative hypotheses are:*

*The test statistic ( is 18.423. The P-value is 1. The level of significance is 5%.*

*The P-value of 1 is higher than the level of significance of 0.05. Thus the null hypothesis should not be rejected. The conclusion is that the model is appropriate for the data set.*

*The null and alternative hypotheses for calculating if credit utilization is significant based on Wald’s test with a 5% level of significance are:*

*is the credit utilization term.*

*The null and alternative hypotheses for calculating if asset=car owned is significant based on Wald’s test with a 5% level of significance are:*

*is the assets=car owned term.*

*The null and alternative hypotheses for calculating if asset=house owned is significant based on Wald’s test with a 5% level of significance are:*

*is the assets=house owned term.*

*The null and alternative hypotheses for calculating if asset=car + house owned is significant based on Wald’s test with a 5% level of significance are:*

*is the assets=car + house owned term.*

*The null and alternative hypotheses for calculating if education=college is significant based on Wald’s test with a 5% level of significance are:*

*is the education=college term.*

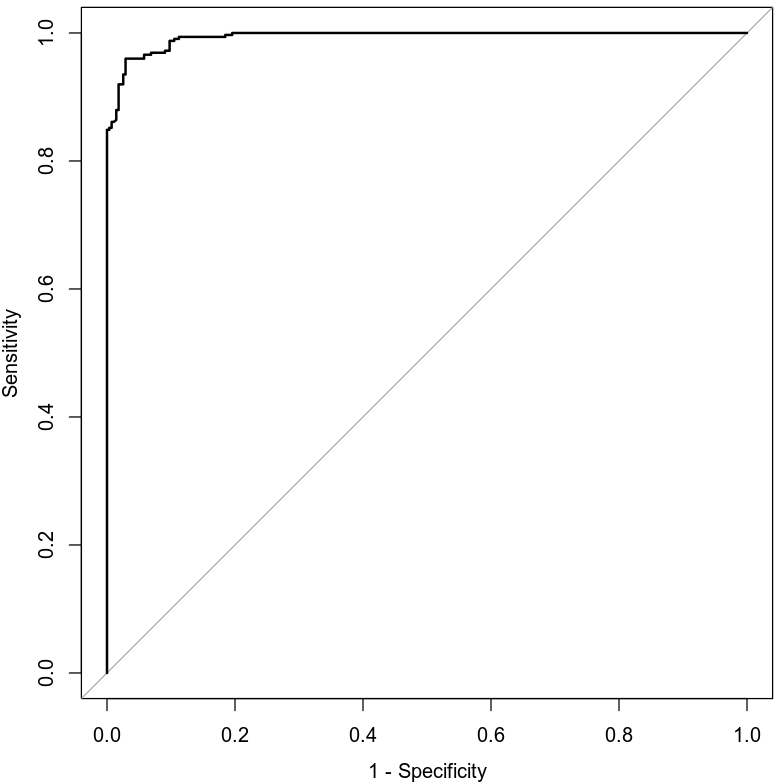
*The null and alternative hypotheses for calculating if education=post graduate is significant based on Wald’s test with a 5% level of significance are:*

*is the education=post graduate term.*

*The P-value for assets=car owned is 0.79365. The P-value for assets=house owned is 0.00013. The P-value for assets=car + house owned is 8.43e-08. The P-value for education=college is 0.00128. The P-value for education=post graduate is 1.68e-09.*

*The P-values for assets=car owned is not below the level of significance of 0.05. The P-values for the other terms (assets=house owned, assets=car + house owned, education=college, and education=post graduate) are all below the level of significance of 0.05. The term for assets=car owned is not statistically significant at this level of significance. The other terms are each statistically significant at this level of significance.*

*The Receiver Operating Characteristic (ROC) curve is displayed here:*

**

*The area under the curve (AUC) is 0.9936 or 99.36%. This is an indicator of how well the model distinguishes between Y = 0 and Y = 1. In general, the larger the AUC the better, because the larger the area under the curve, the better it is at predicting binary classes.*

### **Making Predictions Using Model**

***Prediction 1***

*Using the model to calculate the probability of an individual who has a credit utilization of 43%, owns a car and a house, and has attained a high school diploma defaulting on credit, the odds of this event occurring with these parameters is* 0.984 *or 98.40%. This means that an individual with a credit utilization of 43%, owns a car and a house, and has attained a high school diploma has a 98.40% chance of defaulting on their credit line.*

***Prediction 2***

*Using the model to calculate the probability of an individual who has a credit utilization of 43%, owns a car and a house, and has attained a postgraduate degree defaulting on credit, the odds of this event occurring with these parameters is* 0.3468 *or 34.68%. This means that an individual with a credit utilization of 43%, owns a car and a house, and has attained a postgraduate degree has a 34.68% chance of defaulting on their credit line.*

## **5. Conclusion**

*Based on the analyses performed and assuming that the sample size is sufficiently large, I would recommend using these models. They both have sufficient area under the curve (AUC), every term is statistically significant in each model (with the exception of assets=car owned), and the models show a strong correlation between the variables being analyzed.*

*In each scenario, the results are as expected in a real world scenario. In the first one, it is determined that if an individual misses a payment, they are more likely to default on the credit line than someone who has not missed any payments. In the second scenario, it is determined that if an individual has attained a post graduate degree, they are less likely to default on their credit line than someone with only a high school diploma. Both of these scenarios had other factors that were the same, and only had one factor that was different. The factor being tested in each scenario was significant in determining credit worthiness of an individual with these same factors.*

*The practical importance of the analyses performed, is that credit companies can use the models to determine the risk involved in extending credit to individuals using the variables in this data set. They can test to see which variables have the largest impact on the risk of defaulting on the loan and include those variables in their calculations. They can use the solution of these calculations to determine if they should extend credit to the individual who has applied with them, or if the risk is too great.*

## **6. Citations**

*Zybooks MAT 303: Applied Statistics II for Science, (2016, August).*

*Retrieved April 1, 2020, from https://learn.zybooks.com/zybook/SNHUMAT303v1*

*Codio MAT 303: Applied Statistics II STEM.*

*Retrieved April 1, 2020, from https://codio.com/lsaccoia/module-five-problem-set*