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

Logistic Regression

Amanda Loy

Amanda.Loy1@snhu.edu

Southern New Hampshire University

## **1. Introduction**

We're looking at the different factors in that defaulting on credit it's important to understand how credit is determined. The following data set that will be analyzed includes data on age sex education marriage assets missed payments and credit card utilization overall. From this, we’re hoping to determine if there is a way to predict which clients will default on their loans/credit accounts. Almost done the different analysis is an analysis that will be utilized logistical regressions and Z test and the Hosmer-Lemeshow goodness of fit(GOF)test.

## **2. Data Preparation**

The data set variables include age, sex, marital status, credit utilization, educational level, missed payments, and default status. The variables that will be assessed are credit utilization, missed payments, education level, and assets. The correlation between education and income is often factored into the ability to pay back debt. The data set includes six columns and 600 rows.

Table

Description automatically generated with low confidence

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

### **Reporting Results**

The first logistical regression model will look at defaulting as the response variable. We'll be using the predictors of credit utilization and education. Education is a quantitative variable so we will end up with two dummy variables. the general form is written as followers: E(y) = e(β0 + β1x1 + β2x2+ β3x3)/ 1+ e(β0 + β1x1 + β2x2+ β3x3). Education and credit utilization areas are prediction variables because it is assumed that a higher level of education is equal to a higher income and lower utilization of credit means that they have the financial ability to not default on a loan. The regression model will look slightly different as use the natural log of odds in the predictor in the response variable section. what this means is that we're going to take the natural log of the response variable and that will be the overall answer. After using the R output, the prediction model looks as follows: ln(odds) = -8.8488+0.3438x1-1.4975x2-4.2540x3.

Table

Description automatically generated

The estimated coefficient of credit utilization is interpreted as utilization percentage increases the log odds of defaulting increase by 0.3438.

Graphical user interface, text, application

Description automatically generated

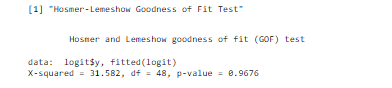
Report the following:

|  |  |  |
| --- | --- | --- |
|  | Variables | Results |
| Accuracy | TP+TN/TP+TN +FP+FN=  (303 +254)/(254+21+22+303) | 0.9283 |
| Precision | TP/TP +FP =  303/(303+22) | 0.9323 |
| Recall | TN/TN +FP  254/(254+22) | 0.9203 |

The confusion matrix is utilized to compare predictions with actual responses. TP is representative of true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives. From the confusion matrix we can test the accuracy which is the correct predictions to the total number of observations, the precision which is the ratio of correct positive predictions to the total predicted positives, and the sensitivity or recall which is the ratio of correct positive predictions to the total positive examples. overall this determines the accuracy of the model used, the precision of the model used, and the sensitivity of the model.

### **Evaluating Model Significance**

Evaluate model significance for the regression model. Address the following questions in your analysis:



The volume model significance and effectiveness a Hosmer-Lemeshow goodness of fit test is conducted. The P value of the test is 0.9676. The null hypothesis is that the model that was generated fits the provided data. The alternative hypothesis is that the model does not fit the data that was provided. The level of significance is 0.05. Based on the p-value and the level of significance we fail to reject the null hypothesis. This means that the model created does fit the data.

Graphical user interface, text, application

Description automatically generated

The Wald’s confidence intervals are used to calculate the slope parameters. This is another test to determine if the data is significant at a 0.05 level of significance. At a level of .05 significance, all variables reject the null that a beta value is equal to zero.

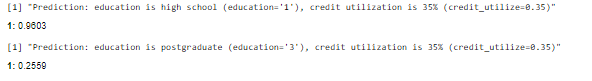
Chart

Description automatically generated

The Receiver Operating Characteristic (ROC) curve is a curve that determines the probability of predicting true positives and true negatives. The Area Under the Curve(AUC) is a representation of how well the model can distinguish between true positive and true negative values. The AUC value is 0.9859. This means that the model has a 99 chance of being able to distinguish between positive and negative predictions.

### **Making Predictions Using Model**

Make predictions using the regression model. Address the following questions in your analysis:



Using the prediction model with credit utilization of 35%, and a high school education results in a 0.9603 chance of defaulting.

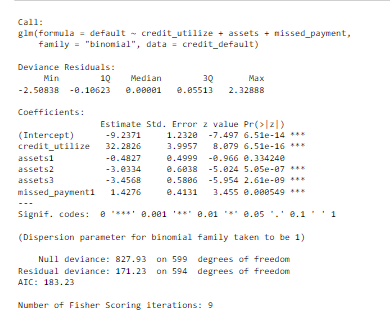
Using the prediction model with credit utilization of 35%, and a postgraduate education results in a 0.2559 chance of defaulting

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

### **Reporting Results**

The second logistical regression model will look at defaulting as the response variable. We'll be using the predictors of credit utilization, assets, and missed payments. Missed payments is looked at as missed payments in the past three months. Assets are quantitative variables so we will end up with three dummy variables. the general form is written as followers: E(y) = e(β0 + β1x1 + β2x2+ β3x3+ β4x4+β5x5)/ 1+ e(β0 + β1x1 + β2x2+ β3x3+ β4x4+β5x5) . X1 represents credit utilization, x2 represents assets, x3 represents assets, x4 represents assets, and x5 represents missed payments.

The prediction equation of this model in terms of the natural log of odds to express the beta terms in linear form. Ln(1/1-n) = e(-9.2371 + 32.2826x1 -0.48272x2-3.0334x3 -3.4568x4+1.4276x5)/ 1+ e(-9.2371+ 32.2826x1 -0.4827x2 -3.0334x3+-3.4568x4+1.4276x5). The logistic regression model that is created using outputs obtained from your R script is ln(odds) = -9.2371 + 32.2826x1 -0.48272x2-3.0334x3 -3.4568x4+1.4276x5



Graphical user interface, text, application, email

Description automatically generated

Report the following:

|  |  |  |
| --- | --- | --- |
|  | Variables | Results |
| Accuracy | TP+TN/TP+TN +FP+FN=  (303 +262)/(262+21+14+303) | 0.9417 |
| Precision | TP/TP +FP =  303/(303+14) | 0.9558 |
| Recall | TN/TN +FP  262/(262+14) | 0.9493 |

TP is representative of true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives. From the confusion matrix we can test the accuracy which is the correct predictions to the total number of observations, the precision which is the ratio of correct positive predictions to the total predicted positives, and the sensitivity or recall which is the ratio of correct positive predictions to the total positive examples. The model is accurate with predicting .9417 of the data. The model is .9558 precise with true positive predictions. The model is able to correct negative predictions at a rate of 0.9493.

### **Evaluating Model Significance**

A screenshot of a computer

Description automatically generated with medium confidence

The volume model significance and effectiveness a Hosmer-Lemeshow goodness of fit test is conducted. The P value of the test is 0.9945. The null hypothesis is that the model that was generated fits the provided data. The alternative hypothesis is that the model does not fit the data that was provided. The level of significance is 0.05. Based on the p-value and the level of significance we fail to reject the null hypothesis. This means that the model created does fit the data.

Graphical user interface, text, application

Description automatically generated

The Wald’s confidence intervals are used to calculate the slope parameters. This is another test to determine if the data is significant at a 0.05 level of significance. At a level of .05 significance, all variables reject the null that a beta value is equal to zero.

Chart

Description automatically generated

The Receiver Operating Characteristic (ROC) curve is a curve that determines the probability of predicting true positives and true negatives. The Area Under the Curve(AUC) is a representation of how well the model can distinguish between true positive and true negative values. The AUC value is 0.9874. This means that the model has a .9874 chance of being able to distinguish between positive and negative predictions.

### **Making Predictions Using Model**

Text

Description automatically generated with medium confidence

Using the prediction model with credit utilization of 35%, owns only a car, and has missed payments in the last 3 months results in a 0.9603 chance of defaulting.

Using the prediction model with a credit utilization of 35%, owns a car and house, and has no missed payments in the last 3 months results in a 0.2559 chance of defaulting.

## **5. Conclusion**

After analyzing both model one and model 2, I believe that model 2 is a more accurate description of how to predict if someone will default on a loan. The assets as well as the utilization seem to play a huge part in determining the significance of or the odds of default. When looking into model 2 the Hosmer-Lemeshow goodness of fit test is statistically higher with .9945 P value. This means that with 99% of the data provided can be predicted by model 2. This means that the data is a good fit for use in terms of using the predictor values of credit utilization assets and missed payments to determine if a loan will default. For example, this is useful in terms of extending a loan out to someone for a vehicle or a home. The predictor variables can help the financial institution assess if the person is likely to default on that loan and determine if they would be a candidate for a loan.