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

Logistic Regression

Kayla Sacks

kayla.sacks@snhu.edu

Southern New Hampshire University

## **1. Introduction**

For this analysis I will be exploring multiple variables concerning the relationship between how likely an individual is to default on their credit and their characteristics. The characteristics that I will be looking at are an individual’s credit utilization, education, the assets they own, and if they have missed any payments in the last 3 months. To explore this relationship, I will be using two logistic regression models and reviewing the results using confusion matrix, the Hosmer-Lemeshow goodness of fit test, Wald’s test, and creating a Receiver Operating Characteristic (ROC) curve for each model. The results of this analysis can help credit companies to learn what customers are more or less likely to default on their credit to better manage their risk.

## **2. Data Preparation**

The data set that I will be using has a total of 8 columns and 600 rows. For my analysis I will be using five of the columns as my variables. The variables I will be using are:

* default – If the individual defaulted on their credit. 0 = did not default, 1 = default. This will be the response variable for both models
* credit utilize – Percentage of how much credit is being used out of credit allowed. This value is a decimal that represents a percentage. This is the first predictor variable in both models.
* education – The highest level of education attained. 1 = high school, 2 = college, 3 = postgraduate. This is the second predictor variables for the first model.
* assets – The assets owned by the individual. 0 = none, 1 = car only, 2 = house only, 3 = car and house. This is the second predictor variable for the second model.
* missed\_payment – If the individual missed any payments in the last 3 months. 0 = none, 1 = missed. This is the third predictor variable for the second model.

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

### **Reporting Results**

The general form for this regression model is:

The prediction regression equation is:

This can then be transformed into a model that is linear in beta terms:

Since the left side of the transformed equation is the natural log of odds, the equation can be written as:

For this model, “odds” is the odds of defaulting on credit (default = 1).

is the probability that the individual will default while is the ratio of the probability that the individual will default, or the odds that they will default.

The prediction regression equation of this model in terms of the natural log of odds is:

The logistic regression model was created using credit utilization and education as predictors and the summary is below.

*A screenshot of a computer

Description automatically generated*

The prediction model equation using the outputs from the logistic regression model is:

The prediction model equation in terms of the natural log of odds using the outputs from the regression model is:

The estimated coefficient of credit utilization is 34.3869. Since credit utilization is expressed as a percentage, we will divide the estimated coefficient by 100 to get 0.343869. This tells us that, on average, the change in log odds for defaulting is 0.343869 for each percent increase in credit utilization if the other variables remain constant. We can express this in terms of odds instead of log odds by calculating:

Since credit utilization is expressed as a percentage, we will multiply the result by 100 to get 41.04%. From this we can say that if there is no change in other variables, for each percentage increase in credit utilization, the odds of defaulting increase by 41.04%.

The confusion matrix was created to find the number of true positives, true negatives, false positives, and false negatives.

*A white background with black text

Description automatically generated*

The outcome from the Confusion Matrix is:

* True Positives – the actual and predicted value are both 1: 303
* True Negatives – the actual and predicted value are both 0: 254
* False Positives – the actual value is 0 but the predicted value is 1: 22
* False Negatives – the actual value is 1 but the predicted value is 0: 21

Accuracy was calculated to find the ratio of the correct predictions to the total number of observations.

Precision was calculated to find the ratio of correct positive predictions to the total predicted positives.

Finally, sensitivity (recall) was calculated to find the ratio of correct positive predictions to the total amount of positives.

### **Evaluating Model Significance**

The Hosmer-Lemeshow goodness of fit test will be used to see if the model is appropriate for the data set. We will start with the null and alternate hypotheses which are:

I used several different group sizes for the Hosmer-Lemeshow test to see if the p-value of the group sizes changed significantly. When running the test, once the group value was 27 or higher a warning stated that “the data did not allow for the requested number of bins.” With this information I ran two full Hosmer-Lemeshow to get the test statistic (X-squared) and df values for the group sizes of 10 and 20.The test statistics for each of these tests are 4.4518 and 13.275 respectively. The lowest p-value from the test is 0.1832411 which is higher than the 0.05 level of significance and will not reject the null hypothesis.

Next will be Wald’s test to determine which terms are statistically significant using the null and alternate hypotheses with a 5% level of significance.

is the term for credit utilization and has a p-value < 2e-16 which is less than the 0.05 level of significance and will reject the null hypothesis.

is the first term for education (college level education) and has a p-value of 0.00134 which is less than the 0.05 level of significance and will reject the null hypothesis.

is the second term for education (post graduate level education) and has a p-value of 9.72e-13 which is less than the 0.05 level of significance and will reject the null hypothesis.

With the conclusion of this test, we can see that both credit utilization and education are both statistically significant in the model when using a 5% level of significance.

The Receiver Operating Characteristic (ROC) curve can be seen below and shows the accuracy of the test by visualizing the ratio of true positives to false positives.

*A graph of a function

Description automatically generated*

The area under the curve (AUC) for the ROC is 0.9859 or 98.59% and indicates how well the model distinguishes between Y= 0 and Y = 1 *(zyBooks, a Wiley brand, 2016)*. The high value shows that the model is good at predicting if an individual will default on their credit.

### **Making Predictions Using Model**

**Prediction 1**

We will use this model to predict the probability of an individual defaulting on credit who has a credit utilization of 35% and has a high school education.

The probability of an individual defaulting on credit who has a credit utilization of 35% and has a high school education 96.03%.

**Prediction 2**

We will use this model to predict the probability of an individual defaulting on credit who has a credit utilization of 35% and has a post graduate education*.*

The probability of an individual defaulting on credit who has a credit utilization of 35% and has a post graduate education 25.59%.

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

### **Reporting Results**

The general form for this regression model is:

The prediction regression equation is:

*.*

The prediction regression equation of this model in terms of the natural log of odds is:

The second logistic regression model was created using credit utilization assets, and missed payment as predictors and the summary is below.

*A screenshot of a computer

Description automatically generated*

Using the output from the model, we can update the prediction equation in terms of natural log odds to:

The confusion matrix was created to find the number of true positives, true negatives, false positives, and false negatives.

*A white rectangular box with black text

Description automatically generated*

The outcome from the Confusion Matrix is:

* True Positives – the actual and predicted value are both 1: 303
* True Negatives – the actual and predicted value are both 0: 262
* False Positives – the actual value is 0 but the predicted value is 1: 14
* False Negatives – the actual value is 1 but the predicted value is 0: 21

Accuracy was calculated to find the ratio of the correct predictions to the total number of observations.

Precision was calculated to find the ratio of correct positive predictions to the total predicted positives.

Finally, sensitivity (recall) was calculated to find the ratio of correct positive predictions to the total amount of positives.

### **Evaluating Model Significance**

The Hosmer-Lemeshow goodness of fit test will be used to see if the model is appropriate for the data set. We will start with the null and alternate hypotheses which are:

I again used several different group sizes for the Hosmer-Lemeshow test to see if the p-value of the group sizes changed significantly. I ran two full Hosmer-Lemeshow to get the test statistic (X-squared) and df values for the group sizes of 10 and 20.The test statistics for each of these tests are 5.5238 and 9.5484 respectively. The lowest p-value from the test is 0.1244927 which is higher than the 0.05 level of significance and will not reject the null hypothesis.

Next will be Wald’s test to determine which terms are statistically significant using the null and alternate hypotheses with a 5% level of significance.

is the term for credit utilization and has a p-value of 6.51e-16 which is less than the 0.05 level of significance and will reject the null hypothesis.

is the first term for assets (own car only) and has a p-value of 0.334240 which is higher than the 0.05 level of significance and will not reject the null hypothesis.

is the second term for assets (own house only) and has a p-value of 5.05e-07 which is less than the 0.05 level of significance and will reject the null hypothesis.

is the third term for assets (own car and house) and has a p-value of 2.61e-09 which is less than the 0.05 level of significance and will reject the null hypothesis.

is the second term for missed payments and has a p-value of 0.000549 which is less than the 0.05 level of significance and will reject the null hypothesis.

With the conclusion of this test, we can see that credit utilization, two of the three asset variables, and missed payment are statistically significant in the model when using a 5% level of significance.

The ROC curve can be seen below and shows the relationship between the true positives and false positives for this model.

*A graph of a function

Description automatically generated with medium confidence*

The AUC for the ROC is 0.9874 or 98.74%. This indicates that this is a good model for predicting if an individual will default on their credit.

### **Making Predictions Using Model**

**Prediction 1**

We will use this model to predict the probability of an individual who has a credit utilization of 35%, owns only a car, and has missed payments in the last 3 months.

The probability of an individual defaulting on credit who has a credit utilization of 35%, owns only a car, and has missed payments in the last 3 months is 95.29%.

**Prediction 2**

We will use this model to predict the probability of an individual who has a credit utilization of 35%, owns a car and a house, and has not missed payments in the last 3 months.

The probability of an individual defaulting on credit who has a credit utilization of 35%, owns a car and a house, and has not missed payments in the last 3 months is 19.86%.

## **5. Conclusion**

Assuming a there is a sufficiently large sample size, I would recommend using these models based on the analysis that was conducted. There are multiple reasons to use these models such as the accuracy, precision, and sensitivity being above 92% for both models. Both models also failed to reject the null hypothesis for the Hosmer-Lemeshow goodness of fit test showing that the model fits the data. The only term that was not significant during the Wald test is the first dummy variable for assets, which is if the individual owns a car only, with the other two dummy variables for assets being statistically significant to the second model. The area under the curve (AUC) is at 0.9858 for model one and 0.9874 for model two which is statistically sufficient.

Based on these results, we can see that each of the predicter variables used in both models are sufficient in predicting if an individual will default on their credit. The practical importance of this analysis is that credit companies can use these models to measure the risk of an individual defaulting on their credit. This can help the companies manage customers with a higher risk to default as well as help with the decision if a customer wants to extend their credit line based on the customers characteristics that were used in the analysis.

## 

## **6. Citations**

zyBooks, a Wiley brand. (2016). *Applied Regression Analysis (R)*. ZyBooks. <https://www.zybooks.com/catalog/applied-regression-analysis-r/>