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

Logistic Regression

Yaqub Mohamud

yaqub.mohamud@snhu.edu

Southern New Hampshire University

**1. Introduction**

For this week's problem set, we will do a risk analysis for a credit card company. The relationship between customer attributes like age, gender, highest level of education attained, marital status, assets owned, credit usage, whether they missed payments in the last three months, whether they defaulted on their credit, and the likelihood of defaulting on credit payments is being examined using a sizable historical dataset. Credit card companies can use the data to determine a person's credit limit and whether they are at higher risk of defaulting. In addition to statistical methods like logistic regression, my analysis makes use of Wald confidence intervals for slope parameters, the Hosmer-Lemeshow goodness of fit (GOF) test, and the receiver operating characteristic curve (ROC curve).

## **2. Data Preparation**

The data collection that will be utilized in the models includes crucial characteristics such as age, gender, education, assets, marriage, missing payments in the previous three months, credit use, and whether they have defaulted on their credit line.  
  
The data set has eight columns and 600 rows.

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

### **Reporting Results**

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

Y = e^(𝛽0 + 𝛽1X1 + 𝛽2X2 + 𝛽3X3) / 1 + e^(𝛽0 + 𝛽1X1 + 𝛽2X2 + 𝛽3X3)

where y is 1 for defaulting on credit and 0 for not defaulting on credit, x1 is credit utilization, and x2 and x3 are dummy variables for education.

The model's prediction equation, which expresses the beta terms in linear form using the natural log of odds, is as follows:

ln (π/1−π )=β0+β1x1+β2x2+β3x3

π represents individual defaulting. While π/1−π represents individuals not defaulting.

This may also be rewritten as follows: ln ( odds)=β0+β1 x1+β2 x2+β3 x3.

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

The prediction model is:

ln( odds )=−8.8488 + 34.3869x1 − 1.4975x2 − 4.2540x3

The estimated coefficient for variable credit utilization is 34.3869. This suggests that the log probability of defaulting may, on average, change when a customer's credit goes into default. Assuming all other factors stay the same, there is a 34.38% likelihood that credit utilization will rise.

The calculated coefficient for education (college) is -1.4975, meaning that for every percentage increase in college education, the average change in the log probabilities for defaulting falls by 0.014975. The calculated coefficient for education (postgraduate) is -4.2540, meaning that for every percentage increase in postgraduate education, the average change in the log chances for defaulting falls by 0.0425.

Results from the matrix of confusion: There were 303 true positives, 21 true negatives, 22 false positives, and 254 false negatives.

According to the confusion matrix, the accuracy value is 0.9279, meaning that the accuracy level is 92.79%; the precision value is 0.9352, meaning that the precision level is 93.52%; and the recall value is 0.9323, meaning that the recall level is 93.23%.

### **Evaluating Model Significance**

We use the Hosmer-Lemeshow Goodness of Fit test (GOF) to determine whether this model is appropriate. The p-value is 0.9676, and the test statistic is 31. 582.Since the p-value is greater than the 5% significance level, we fail to reject the null hypothesis.

The P-value are 2e-16 for credit utilization, 0.00134 for college education, and 9.72e-13 for post-graduation education. All these numbers fall below the significance level of 0.05, which is the 5% level. This indicates that every value in this model is statistically significant.

The curve for the Receiver Operating Characteristic (ROC) is shown here:

A graph of a function

AI-generated content may be incorrect.

This ROC's area of 0.9859 is quite close to 1, meaning that 98.59% of examples can be accurately classified by the model. With a high AUC value, the model can accurately forecast whether credit would be defaulted on.

### **Making Predictions Using Model**

Using the model to calculate the probability of an individual who has a credit utilization of 35% and has a high school education, the odds of this event occurring with these parameters is 0.9603 or 96.03%. This means that this individual has a 96.03% chance of defaulting on their credit line.

Using the model to calculate the probability of an individual who has a credit utilization of 35% and has a post graduate education, the odds of this event occurring with these parameters is 0.2559 or 25.59%. This means that this individual has a 25.59% 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 missed payments as independent variables is:

Y = e^(𝛽0 + 𝛽1X1 + 𝛽2X2 + 𝛽3X3+ 𝛽4X4+ 𝛽5X5) / 1 + e^(𝛽0 + 𝛽1X1 + 𝛽2X2 + 𝛽3X3+ 𝛽4X4+ 𝛽5X5)

where x1 is credit utilization, x2, x3, and x4 are dummy variables for assets, x5 is missing payments, and y is 1 for credit default and 0 for non-default.

The model's prediction equation, which expresses the beta terms in linear form using the natural log of odds, is as follows:

ln (π/1−π )=β0+β1x1+β2x2+β3x3+β4x4+β5x5

π represents individual defaulting. While π/1−π represents individuals not defaulting.

This may also be rewritten as follows: ln ( odds)=β0+β1 x1+β2 x2+β3 x3+β4 x4+β5 x5.

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

The prediction model is:

ln( odds )=−9.2371 + 32.2826x1 − 0.4827x2 − 3.0334x3 − 3.4568x4 + 1.4276x5

A screenshot of a graph

AI-generated content may be incorrect.

True negatives: 262, false positives: 14, false negatives: 21, true positives: 303

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Accuracy = (303 + 262) / (303 + 262 + 14 + 21)

Accuracy = 0.9416 or 94.2%

Precision = TP / (TP + FP)

Precision = 303 / (303 + 14)

Precision = 0.9558 or 95.6%

Recall = TP / (TP + FN)

Recall = 303 / (303 + 21)

Recall = 0.9351 or 93.5%

### **Evaluating Model Significance**

We use the Hosmer-Lemeshow Goodness of Fit test (GOF) to determine whether this model is appropriate. The p-value is 0.9924, and the test statistic is 26.733. Since the p-value is greater than the 5% significance level, we fail to reject the null hypothesis.

The P-value are 6.51e-16 for credit utilization, 0.3342 for car only asset, 5.05e-7 for house only asset, 2.61e-9 for car and house asset, and 0.0005 for missed payment. All these numbers except house only asset fall below the significance level of 0.05, which is the 5% level. This indicates that every value except house only asset in this model is statistically significant.

The curve for the Receiver Operating Characteristic (ROC) is shown here:

A graph of a function

AI-generated content may be incorrect.

This ROC's area of 0.9874 is quite close to 1, meaning that 98.74% of examples can be accurately classified by the model. With a high AUC value, the model is a significant tool for predicting whether a person will default on their credit.

### **Making Predictions Using Model**

Using the model to calculate the probability of an individual who has a credit utilization of 35%, who owns a car and has missed payments in the last three months, the odds of this event occurring with these parameters is 0.9529 or 95.29%. This means that this individual has a 95.29% chance of defaulting on their credit line.

Using the model to calculate the probability of an individual who has a credit utilization of 35% %, who owns a car and a house and has missed payments in the last three months, the odds of this event occurring with these parameters is 0.1986 or 19.86 %. This means that this individual has a 19.86% chance of defaulting on their credit line.

## **5. Conclusion**

In conclusion, I would suggest using these models, assuming that the sample size is appropriately large depending on the analyses performed. Both models have an adequate area under the curves (AUC), each term is statistically significant, and there is a strong correlation between the variables being studied.

The outcomes of each case are consistent with what would be expected in a real-world situation. According to the first one, a person who misses a payment has a higher chance of going into default on the credit line than a person who hasn't missed any. According to the second scenario, a person with a postgraduate degree has a lower chance of defaulting on their credit line than someone with just a high school degree. Only one component differed between the two scenarios; all other factors were the same. Each scenario's tested component was important in assessing a person's credit worthiness based on those same variables.

Credit companies may use the models to assess the risk of giving credit to individuals based on the factors in this data set, which makes the analyses conducted practically significant. They can conduct tests to determine which factors most significantly affect the likelihood of loan default and include those factors into their computations. By using the results of these computations, they can decide whether to give credit to the applicant or if the risk is too high.