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

Logistic Regression

Nicholas Kreuziger

nicholas.kreuziger@snhu.edu

Southern New Hampshire University

## **1. Introduction**

The dataset being explored is the “default of credit card clients Data Set” hosted by The University of California Irvine. This dataset is intended for training a machine earning model to estimate the probability of someone defaulting on their credit commitment. Since the response variable is whether someone defaults, it is a binary measure and requires the use of a Logistic Regression Model. Results from analyses such as these can be used to assess the likelihood a credit applicant would default on their credit cards.

## **2. Data Preparation**

Within the entire dataset there are 24 columns and 600 rows. For our purposes we will utilize 8 columns and the full 600 rows. The variables we are interested in are represented by the table below:

Text, email

Description automatically generated

Most of these data points are quantitative. As such a Multiple Regression Model would not work as efficiently, Multiple Regression Models output values below 0 and above 1. A Logistic function produces an S-shaped curve and values within the range of 0 to 1.

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

### **Reporting Results**

We will first create a model that estimates the probability of defaulting on credit using credit utilization and missed payments as predictor variables.

E(y) = Probability of Default

β0 = Intercept Parameter; indicates when the logistic curve crosses the vertical axis.

X1 = Credit Utilization Ratio

β1 = Coefficient of Credit Utilization Ratio

X2 = Missed Payment

β2 = Coefficient of Missed Payment

The general form of this model appears:

Written in the natural log of odds to express beta terms in linear form:

In the natural log of odds, π represents the probability for default, not the value of pi as it normally constitutes. represents the odds of success. In this model it represents the probability of defaulting.

Applying the General Model, we use statistical software to arrive at a summary, shown below:

Text

Description automatically generated with low confidence

According to these coefficients, for every percent of credit utilization (assuming all other fixed values), you will see an increase in 0.312090 of default likelihood. This can be expressed in terms of odds by,

Every percent of credit utilization carries with it a 36.63% increase in likelihood for default assuming all other variables remain constant. Thus, Credit Utilization carries with it a heavy influence on the likelihood for credit default.

Using these Coefficients, our General Model and natural log of odds become:

A confusion matrix is used as a diagnostic for Logistic Models. A confusion matrix is a table output as:

|  |  |  |
| --- | --- | --- |
|  | Prediction = 0 | Prediction =1 |
| Actual = 0 | True Negatives | False Positives |
| Actual = 1 | False Negatives | True Positives |

Using a confusion matrix we can assess Accuracy, Precision and Recall. Defined as,

**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.

The Confusion Matrix for Model 1 is computed:

Graphical user interface, text

Description automatically generated with medium confidence

For simplicity, True Negative = 246, False Negative = 29, True Positive = 295, False Positive = 30, thus:

### **Evaluating Model Significance**

In the interests of assessing this model’s validity we will apply a Hosmer-Lemeshow goodness of fit test. A Hosmer-Lemeshow test assesses whether a model is appropriate for the dataset by assessing whether model predictions are close to the observed values of the response variable (in this case the actual instances of credit default vs predicted instances).

The Null Hypothesis is,

H0: The model fits the data

The Alternative Hypothesis is,

Ha: The model does not fit the data

We will test at a 5% level of Significance,

α = 0.05

Text

Description automatically generated

The P-Value of 0.4298 is larger than the Significance Level 0.05, we fail to reject the Null Hypothesis. The Model fits the data.

Following a Hosmer-Lemeshow test we can further validate the model by testing its individual factors with a Wald Test. Each individual Wald Test can be summed as such,

Null Hypothesis, Coefficient of the predictor variable is 0

βi = 0

Alternative Hypothesis, Coefficient of the predictor variable does not equal 0

βi ≠ 0

The p-values of these individual tests can be found in the PR(>|z|) column of the Summary for this model:

Text

Description automatically generated

According to the Wald Test every term is significant in this model.

An additional test to assess this model is the ROC curve. An ROC curve measures the performance for a classifier at various threshold settings. The area under the curve is an indicator of how well the model distinguishes between Y = 0 or Y =1 (in this model, whether someone defaults on their credit card). In the ROC curve graph below you can see there is a large area beneath the line, indicating this model does a good job of distinguishing between Y = 0 or Y = 1.

Chart

Description automatically generated

A picture containing diagram

Description automatically generated

The value of AUC (Area under curve) assessed is 0.9746. This indicates that 97.46% of the area of this plot is within the curve. This is exceptional!

### **Making Predictions Using Model**

Now that we have thoroughly tested the validity of this model, we can make predictions for a few scenarios.

1. What is the probability of an individual who has a credit utilization of 32% and has missed payments in the past three months defaulting on credit?
   1. 
   2. There is a 75% likelihood of default with this combination of factors. In odds this means .75 / (1 - .75) = 3. There is a 3:1 (3 in 1) chance of credit default.
2. What is the probability of an individual who has a credit utilization of 32% and has not missed payments in the past three months defaulting on credit?
   1. 
   2. There is a 40.35% likelihood of default with this combination of factors. In odds this means .4035 / (1 - .4035) = 0.6764. There is an 6764:10,000 (6764 in 10,000) chance of credit default.

This model, in the event the applicants both have a 32% credit utilization, gauges a 34.65% difference in likelihood of default depending upon if a payment was missed in the last 3 months. Intuitively this makes sense, although I would’ve expected lower than 75%.

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

### **Reporting Results**

We will now create a second model that estimates the probability of defaulting on credit using credit utilization, assets, and education as predictor variables. Note that due to Assets having 4 possible inputs (0, 1, 2 or 3) and Education having 3 possible inputs (1, 2 or 3) they each have dummy variables. Categorical variables table is below the named variables to help clarify the meaning of each input.

E(y) = Probability of Default

β0 = Intercept Parameter; indicates when the logistic curve crosses the vertical axis.

X1 = Credit Utilization Ratio

β1 = Coefficient of Credit Utilization Ratio

X2 = Assets 1

β2 = Coefficient of Assets 1

X3 = Assets 2

β3 = Coefficient of Assets 2

X4 = Assets 3

β4 = Coefficient of Assets 3

X5 = Education 2

β5 = Coefficient of Education 2

X6 = Education 3

β6 = Coefficient of Education 3

Categorical Variables are represented:

|  |  |  |  |
| --- | --- | --- | --- |
| Assets Value | X2 | X3 | X4 |
| 0 (None) | 0 | 0 | 0 |
| 1 (Car Only) | 1 | 0 | 0 |
| 2 (House Only) | 0 | 1 | 0 |
| 3 (Car and House) | 0 | 0 | 1 |

|  |  |  |
| --- | --- | --- |
| Education Value | X5 | X6 |
| 1 (High School) | 0 | 0 |
| 2 (College) | 1 | 0 |
| 3 (Post-Graduate) | 0 | 1 |

The general form of this model appears:

Written in the natural log of odds to express beta terms in linear form:

Applying the General Model, we use statistical software to arrive at a summary, shown below:

Table

Description automatically generated

Using these Coefficients, our General Model and natural log of odds become:

The Confusion Matrix for Model 1 is computed:

Table

Description automatically generated with medium confidence

For simplicity, True Negative = 266, False Negative = 13, True Positive = 311, False Positive = 10, thus:

### **Evaluating Model Significance**

In the interests of assessing this model’s validity we will apply a Hosmer-Lemeshow goodness of fit test. A Hosmer-Lemeshow test assesses whether a model is appropriate for the dataset by assessing whether model predictions are close to the observed values of the response variable (in this case the actual instances of credit default vs predicted instances).

The Null Hypothesis is,

H0: The model fits the data

The Alternative Hypothesis is,

Ha: The model does not fit the data

We will test at a 5% level of Significance,

α = 0.05

Text

Description automatically generated

The P-Value of 1 is larger than the Significance Level 0.05, we fail to reject the Null Hypothesis. The Model fits the data.

Following a Hosmer-Lemeshow test we can further validate the model by testing its individual factors with a Wald Test. Each individual Wald Test can be summed as such,

Null Hypothesis, Coefficient of the predictor variable is 0

βi = 0

Alternative Hypothesis, Coefficient of the predictor variable does not equal 0

βi ≠ 0

The p-values of these individual tests can be found in the PR(>|z|) column of the Summary for this model:

Text

Description automatically generated

According to the Wald Test every term is significant in this model except for Assets 1. Recall that Assets 1 value indicates the credit applicant only possesses a car. This indicates that possession of a car makes minimal difference in likelihood of default.

An additional test to assess this model is the ROC curve. An ROC curve measures the performance for a classifier at various threshold settings. The area under the curve is an indicator of how well the model distinguishes between Y = 0 or Y =1 (in this model, whether someone defaults on their credit card). In the ROC curve graph below you can see there is a large area beneath the line, indicating this model does a good job of distinguishing between Y = 0 or Y = 1.

Chart

Description automatically generated

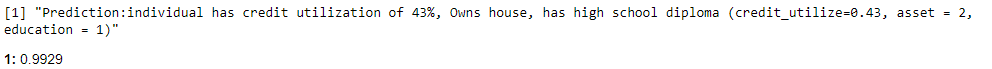
A picture containing chart

Description automatically generated

The value of AUC (Area under curve) assessed is 0.9936. This indicates that 99.36% of the area of this plot is within the curve. This is better than the AUC value for model 1.

### **Making Predictions Using Model**

Now that we have thoroughly tested the validity of this model, we can make predictions for a few scenarios.

1. What is 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?
   1. 
   2. There is a 99.29% likelihood of default with this combination of factors. In odds this means .9929 / (1 - .9929) =139.85. There is a 13,985:100 (13,985 to 100) chance of credit default.
2. What is 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?
   1. Text

      Description automatically generated
   2. There is a 34.68% likelihood of default with this combination of factors. In odds this means .3468 / (1 - .3468) = 0.5309. There is an 5309:10,000 (5309 in 10,000) chance of credit default.

The only difference in these scenarios was the difference between a high school education and a postgraduate degree. There is a much higher likelihood of default when you have only a high school education vs a postgraduate education. This would make sense because experientially speaking a borrower with a postgraduate degree is likely to have more income and experience remaining consistent on their credit obligations.

## **5. Conclusion**

With the analyses performed in this report two logistic regression models were made to predict the likelihood of defaulting on credit card debt.

Below is a table summarizing the variables utilized:

|  |  |  |
| --- | --- | --- |
| Variables | Model 1 | Model 2 |
| Credit utilization | X | X |
| Missed payment | X |  |
| Assets |  | X |
| Education |  | X |

Below is a table summarizing model testing results:

|  |  |  |
| --- | --- | --- |
| Test | Model 1 | Model 2 |
| Accuracy (ratio of the number of correct predictions to the total number of observations) | 90.17% | 96.17% |
| Precision (ratio of correct positive predictions to the total predicted positives) | 90.77% | 96.88% |
| Recall (ratio of correct positive predictions to the total positives examples) | 91.05% | 95.99% |
| Hosmer-Lemeshow p-value | 0.4298 | 1 |
| Wald Test (individual predictors determined as statistically significant) | 2/2 | 5/6 |
| AUC Curve (ROC test) | 97.46% | 99.36% |

These results indicate that while missed payments in conjunction with credit utilization can bring a model to within 90% accuracy. Assets and education in conjunction with credit utilization can bring a model to within 95% or more accuracy. It’s impressive that either of these models can approach that level of accuracy. Furthermore, when you consider the number of these variables that passed the Wald Test you can further conclude this Models dependability. I would be comfortable recommending both models in concert to provide two probability outputs. Averaging between the two probabilities would provide a reasonable gauge of likelihood to default on a credit obligation. Due to the nature of qualitative data upping the number of inputs helps to better assess the dynamic nature of the real-world. It would also be beneficial to see this model applied to more real-world datasets to get more stress testing of accuracy.

Practically speaking this model has a lot of importance. It helps to better assess what factors are related to credit default. This model is a good example of a concept that could be applied to a wider net of situations, whether it’s mortgage default or other economic situations. You could use these results to argue for the need to emphasize and fund public education further. It also begs the question of what other factors may affect credit default. Could access to public transportation be a factor? When you consider the minimal impact of solely owning a car, it may be because there was access to public transportation.