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

Logistic Regression

Ashley De Venuto

ashley.devenuto@snhu.edu

Southern New Hampshire University

## **1. Introduction**

For this problem set, a risk analysis will be run for a credit card company. The data set that is being used is a large historical data set. This data set will be used to analyze the relationship between the terms age, sex, education, marriage, assets, missed payments within the last three months, credit utilization, and default. The analysis will determine if any of these terms/characteristics have any way of determining if a customer using their service will default on payments. The results of the analysis can help credit card companies determine a customer’s credit limit and if any specific customer will default on their payments. For this problem set, an logistic regression model, wald test, the Hosmer-Lemeshow Goodness of Fit (GOF) test, an confusion matrix, and an ROC curve will be used to complete this analysis.

## **2. Data Preparation**

For this problem set, there are a lot of different variables that will be looked at during this analysis. The variables within the data set are age, sex, education, marriage, assets, missed payments, credit utilization, and default. The main variables being worked with are defaulting, credit utilization, education, assets, and missed payments. There are 8 columns and 600 rows. The 8 columns represent the 8 variables. The 600 rows represent the records within the data set.

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

### **Reporting Results**

The general form of a logistic regression model is:

The prediction equation of a logistic regression model is:

The pi

 in the equation equals the probability of an event happening. For this specific problem/equation, it determines the probability of defaulting on a credit card. The term Equation

A fraction with pi in the numerator, and one minus pi in the denominatoris the ratio of the probability at which the event happens.

The logistic regression model looks like this:

The estimated coefficient of credit utilization is 34.5288. This means that the change in log odds on average for defaulting is 0.345288 for each increase in credit utilization. This only happens if all the other variables are constant.

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

A screenshot of a computer

Description automatically generated

The true negatives is 251, the false positives is 25, the false negatives is 21, and the true positives is 303.

Accuracy = ()

Accuracy = (

Accuracy = or 92.3%

Precision = (

Precision = (

Precision = or 92.4%

Recall = (

Recall = (

Recall = = 0.9351 or 93.5%

### **Evaluating Model Significance**

For this portion of the analysis, the Hosmer-Lemeshow goodness of fit test is what’s going to be used. The null hypothesis is:

The alternative hypothesis is:

The test statistic is:

30.983

The P-value is:

0.956

The P-value is greater than the 0.05 or 5%. This means that the null hypothesis is not rejected. It also means that the model fits the data.

When using the Wald’s test, the null hypothesis is:

The alternative hypothesis is:

The p-value for credit utilization is 2e-16 which is less than the 0.05 or 5% significance level. This means that this variable is significant when it comes to determining if a customer will default on their credit card payments. The p-value for education is 7.62e-13 which is less than the 0.05 or 5% significance level. This means that this variable is significant when it comes to determining if a customer will default on their credit card payments. The results show that both the variables have a significant relationship with default which also means that null hypothesis is rejected, and the alternative hypothesis is accepted.

The Receiver Operating Characteristic (ROC) curve is used to measure the accuracy of the differential between Y = 0 and Y = 1. This model has the ability to accurately predict binary classes that are directly correlated to the size of the area under the curve. The bigger the AUC is means it is better accurate. The following image shows the ROC curve:

A graph of a function

Description automatically generated

The AUC is 0.9855. This means that the accuracy rate of predicting whether someone will default or not default is about 98.55%.

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

### **Making Predictions Using Model**

The probability of an individual defaulting on credit who has a credit utilization of 35% and has a high school education is about 97 %. The probability of an individual defaulting on credit who has a credit utilization of 35% and has a post-graduation education is about 32%.

**A white background with black text

Description automatically generated**

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

## **Reporting Results**

The general form of a logistic regression model is:

The prediction equation of a logistic regression model is:

The logistic regression model looks like this:

The estimated coefficient of credit utilization is 32.2559. This means that the change in log odds on average for defaulting is 0.322559 for each increase in credit utilization. This only happens if all the other variables are constant.

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

**A screenshot of a computer

Description automatically generated**

The true negatives is 262, the false positives is 14, the false negatives is 22, and the true positives is 302.

Accuracy = ()

Accuracy = (

Accuracy = or 94%

Precision = (

Precision = (

Precision = or 95.6%

Recall = (

Recall = (

Recall = = 0.9321 or 93.2%

### **Evaluating Model Significance**

For this portion of the analysis, the Hosmer-Lemeshow goodness of fit test is what’s going to be used. The null hypothesis is:

The alternative hypothesis is:

The test statistic is:

29.739

The P-value is:

0.9767

The P-value is greater than the 0.05 or 5%. This means that the null hypothesis is not rejected. It also means that the model fits the data.

When using the Wald’s test, the null hypothesis is:

The alternative hypothesis is:

The p-value credit utilization is 2.90e-16, which is less than the 0.05 or 5% significance level. The p-value for assets is 1.07e-11, which is less than the 0.05 or 5% significance level. The p-value for missed payments is 0.000149, which is less than the 0.05 or 5% significance level. The results show that reject the null hypothesis and accept the alternative hypothesis. All the variables have a significant relationship with default (response variable).

The Receiver Operating Characteristic (ROC) curve is used to measure the accuracy of the differential between Y = 0 and Y = 1. This model has the ability to accurately predict binary classes that are directly correlated to the size of the area under the curve. The bigger the AUC is means it is better accurate. The AUC is 0.9869. This means that the accuracy rate of predicting whether someone will default or not default is about 98.55%. The following image shows the ROC curve:

**A graph of a function

Description automatically generated**

### **Making Predictions Using Model**

The probability of an individual who has credit utilization of 35%, owns only a car, and has missed payments in the last 3 months is 0.9219 which is about 92% of those individuals. 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 is 0.1668 which is about 17% of those individuals.

## **5. Conclusion**

Based on the results, I would recommend using these models when assuming the sample size is big

enough. Both ROC curves have a lot of area under the curve. This points to a great level of accuracy. All

the variables tested here have a significant relationship with the response variable. The outcomes from

the analysis can be anticipated in the real world. When looking at the first model, the results show that a

person with only high school education is more likely to default on payments on their credit card while a

person that has post-graduation education, they are more likely to not default. In the second model, a

person that doesn’t have a lot of assets and has missed payments before is more likely to default on

payments later while a person that has multiple assets and hasn’t missed a payment is more likely to

not default on payments. These models can be used to test any combination of variables to predict the

risk of customer s defaulting on payments on their credit card. The purpose of the analysis was to be

able to help the credit card companies assess whether or not to give an individual a line of credit. Using

these models, the credit card companies can use any factor to see the likelihood of them having any

significance with defaulting on payments. They can use the results they got from using these models to

decide to give an individual a line of credit or a loan or not to because the risk is to high.