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

Logistic Regression

Sarah Steinbaum

sarah.steinbaum@snhu.edu

Southern New Hampshire University

## **1. Introduction**

The following analyses will be a risk analysis, used to help a credit card company determine the likelihood of an individual defaulting on their credit based on certain factors such as credit utilization, the types of assets an individual has, education level, and whether or not an individual has missed a payment within the last three months. We will use a set of historical data to conduct the analyses which includes logistic regression using quantitative and qualitative variables.

## **2. Data Preparation**

There are several different variables that are important in this problem set. The response variable is, *default*, and represents how likely an individual is to default on their credit. The value 0 indicates that an individual did not default on their credit and 1 indicates that the individual did default on their credit. The quantitative predictor variable, *credit\_utilization*, represents credit utilization by the individual, essentially, how much of the credit allowed is being used. The qualitative predictor variable, *education*, represents the highest education level attained by the individual with 1 representing high school, 2 representing college, and 3 representing post-grad. The qualitative predictor variable, *assets*, represents the assets owned by the individual with 0 representing no assets, 1 representing car only, 2 representing house only, and 3 representing that the individual owns a car and a house. The variable, *missed\_payment*, represents whether the individual has missed a payment within the last three months with 0 indicating no missed payment and 1 indicating there has been a missed payment. There are eight columns that each contain a variable and 600 rows.

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

### **Reporting Results**

The general form of a logistic regression model:

This model can be transformed to form a model that is linear in beta terms:

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

The prediction equation of a logistic regression model:

In this model, y and *odds* represent defaulting on credit as the response variable, represents credit utilization, and and are dummy variables for education. Education is a qualitative predictor variable that represents the highest level of education attained. High school is characterized by 1, college is characterized by 2, and post grad is characterized by 3. The symbol, , represents the probability an individual will default on credit. is the ratio of the probability an individual will default on credit.

The logistic regression model:

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

The estimated coefficient of credit utilization is 34.3869. On average, this means that the change in log odds is 0.343869 for each percentage increase in credit utilization, holding all other variables constant.

The values for the confusion matrix are listed below:

* True positive = 303
* True negative = 254
* False positive = 22
* False negative = 21

Accuracy is the ratio of the number of correct predictions to the total number observations. The equation to for accuracy is:

Accuracy =

We can use the confusion matrix to solve the equation:

Accuracy = = 0.9283333

Precision is the ratio of correction predictions to the total predicted positives. The equation for precision is:

Precision =

We can use the confusion matrix to solve the equation:

Precision = = 0.93230769

Recall is the correct positive predictions to the total positive examples. The equation for recall is:

Recall =

We can use the confusion matrix to solve the equation:

Recall = = 0.93518519

### **Evaluating Model Significance**

We will now conduct the Hosmer-Lemeshow Goodness of Fit (GOF) test to determine whether or not the model is appropriate for the data set. The null hypothesis is that the model does fit the data set. The alternative hypothesis is that the model does not fit the data set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test statistic | P-value | Conclusion |
| First Logistic Regression Model | = The model fits the data  = The model does not fit the data | *t* = 31.582 | *p* = 0.9676 | Fail to reject the null hypothesis |

The P-value is greater than the 5% level of significance which indicates that there is sufficient evidence to fail to reject the null hypothesis. We can conclude that the model is appropriate for this data set. We will now conduct the Wald’s test in order to find out which terms in the model are significant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test statistic | P-value | Conclusion |
| Credit utilization  (credit\_utilize) |  | *t* = 8.527 | *p* = < 2E-16 | Reject the null hypothesis |
| Education – college  (education2) |  | *t* = -3.208 | *p* = 0.00134 | Reject the null hypothesis |
| Education – post grad  (education3) |  | *t* = - 7.134 | *p* = 9.7E-13 | Reject the null hypothesis |

Each variable has a P-value that is less than the 5% level of significance, which indicates that we have sufficient evidence to reject the null hypothesis. We can conclude that the predictor variables *credit\_utilize*, *education2*, and *education3* are statistically significant.

Chart

Description automatically generated

The graph above shows theReceiver Operating Characteristic (ROC) curve. This can be considered a probability curve and is a measurement on the performance of a classifier at different threshold settings. The value of the Area Under the Curve (AUC) is 0.9874. The AUC represents the measure of separability. Essentially, this chart tells us how accurate it is at distinguishing the binary responses, 0s as 0s and 1s as 1s. A higher AUC indicates increased accuracy.

### **Making Predictions Using Model**

The probability of an individual defaulting on credit who has a credit utilization of 35% and a high school education is 0.9603. This indicates that there is a 96% likelihood that an individual with these characteristics would default on their credit. The probability of an individual defaulting on credit who has a credit utilization of 35% and has a post grad education is 0.2559. This indicates that there is roughly a 25% chance that an individual would default on their credit, given these variables. Both scenarios use a credit utilization of 35% but the different education levels gave a drastic difference in probability for defaulting on credit. Based on this analysis, we can conclude an individual’s level of education can heavily impact the likelihood of defaulting on credit.

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

### **Reporting Results**

The general form of a logistic regression model:

This model can be transformed to form a model that is linear in beta terms:

+ +

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

+ +

The prediction equation of a logistic regression model:

In this model, y and *odds* represents the response variable which is defaulting on credit, represents credit utilization. Assets is a qualitative variable where an individual can have no assets (0), a car only (1), house only (2), or a car and house (3). Assets is represented by the dummy variables , , and . represents missed payment where 0 indicates no missed payments and 1 indicates that there has been at least one missed payment within the last three months.

Once the R script has ran and we have our model, we can add the beta estimates to the equation:

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

The values for the confusion matrix are listed below:

* True positive = 303
* True negative = 262
* False positive = 14
* False negative = 21

Accuracy is the ratio of the number of correct predictions to the total number observations. The equation to for accuracy is:

Accuracy =

We can use the confusion matrix to solve the equation:

Accuracy = = 0.94166667

Precision is the ratio of correction predictions to the total predicted positives. The equation for precision is:

Precision =

We can use the confusion matrix to solve the equation:

Precision = = 0.95583596

Recall is the correct positive predictions to the total positive examples. The equation for recall is:

Recall =

We can use the confusion matrix to solve the equation:

Recall = = 0.93518519

**Evaluating Model Significance**

We will now conduct the Hosmer-Lemeshow Goodness of Fit (GOF) test to determine whether or not the model is appropriate for the data set. The null hypothesis is that the model does fit the data set. The alternative hypothesis is that the model does not fit the data set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test statistic | P-value | Conclusion |
| Second Logistic Regression Model | = The model fits the data  = The model does not fit the data | *t* = 26.733 | *p* = 0.9945 | Fail to reject the null hypothesis |

Since the P-value is greater than the 5% level of significance, we have sufficient evidence and fail to reject the null hypothesis. We can conclude that the model is appropriate for the data set. We will not conduct Wald’s test in order to find which terms are statistically significant in the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test statistic | P-value | Conclusion |
| Credit Utilization  (credit\_utilization) |  | *t* = 8.079 | *p* = 6.51E-16 | Reject the null hypothesis |
| Assets – Car Only  (assets1) |  | *t* = -0.966 | *p* = 0.334240 | Fail to reject the null hypothesis |
| Assets – House Only  (assets2) |  | *t* = -5.024 | *p* = 5.05E-07 | Reject the null hypothesis |
| Assets – Car and House  (assets3) |  | *t* = -5.806 | *p* = 2.61E-09 | Reject the null hypothesis |
| Missed Payment  (missed\_payment1) |  | *t* = 3.455 | *p* = 0.000549 | Reject the null hypothesis |

The variables *credit\_utilization*, *assets2*, *assets3*, and *missed\_payment1* each have a P-value less than the 5% level of significance which indicates that these variables are statistically significant and have a correlation to defaulting on credit. The variable, *assests2*, has a P-value greater than the 5% level of significance, making it not statistically significant. This indicates that an individual owning a house only does not have a correlation to defaulting on credit.

Chart

Description automatically generated

TheReceiver Operating Characteristic (ROC) curve is presented on the graph above. This is a measurement on the performance of a classifier at different threshold settings. The value of the Area Under the Curve (AUC) is 0.9529. The AUC represents the measure of separability. The ROC tells us how accurate the model is at distinguishing the binary responses, 0s as 0s and 1s as 1s. A higher AUC indicates increased accuracy.

### **Making Predictions Using Model**

The probability of an individual who has a credit utilization of 35%, only owns a car, and has missed payments in the last three months is 0.9529. This means that an individual with these characteristics has a 95% likelihood to default on their credit. 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 three months is 0.1986. This indicates that there is roughly a 20% chance that someone with these characteristics would default on their credit.

## **5. Conclusion**

Both models helped us determine the correlation certain variables have with how likely an individual is to default on their credit. The first logistic regression model used the predictor variables for credit utilization and highest level of education attained. We conducted a Hosmer-Lemeshow Goodness of Fit test and determined that the model was appropriate for the data set as the P-value was 0.9676 and greater than the 5% level of significance. We then conducted Wald’s test in order to find which terms were statistically significant in the model. It was determined that each predictor variable was statistically significant which indicates that there is a correlation to an individual’s highest level of education and credit utilization to defaulting on credit. Next, we created a Receiver Operating Characteristic (ROC) curve and determined that the value of the Area Under the Curve (AUC) was 0.9874 which suggests that the model is highly accurate at distinguishing binary responses.

The second logistic model used predictor variables for credit utilization, assets, and missed payment. We conducted a Hosmer-Lemeshow Goodness of Fit test and determined that the model was appropriate for the data set as the P-value was 0.9945 and greater than the 5% level of significance. After conducting Wald’s test to find which terms were statistically significant, it was determined that the variables for credit utilization, assets for car only, and assets for house and car were statistically significant and therefore have a correlation to defaulting on credit. The variable assets for house only was not statistically significant. The ROC curve was then created and the value for AUC was 0.9529 which suggests that the model is highly accurate. Overall, based on the analyses that were performed and assuming that the sample size is sufficiently large, I would recommend using the second logistic model.

## **6. Citations**

Berrier, J. (2016). MAT 303: Applied Statistics 2 for Science. Zyante Inc. (zyBooks.com)