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

Logistic Regression

Camilo Hoyos

camilo.hoyos@snhu.edu

Southern New Hampshire University

## **1. Introduction**

The data set that will be exploring is related to customers and their credit risk. Example attributes include, age, sex, education, assets, and missed payments. The results can be used to identify if a customer is likely to default on their credit or make all payments to keep a low balance. The analysis that will be performed are logistic regression models, one in particular contains a qualitative variable.

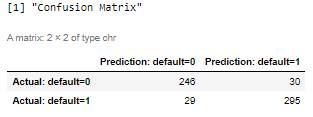
## **2. Data Preparation**

The important variables in this data set are credit utilization, missed payments, assets, and education. All of these are qualitative variables with the exception of credit utilization which is quantitative. For missed payments, it determines if they have missed a payment in the past 3 months (0-no,1-yes). Assets determine what they own (0-none, 1-car, 2-house, 3-car and house). And education determines the highest education level attained (1-high school, 2-college, 3-post graduate). The data set contains 8 columns and 600 rows.

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

### **Reporting Results**

The general form of logistic regression model is (enlarged for readability): , where X1 is credit utilization, and X2 is if the customer has missed a payment in the past three months. The model in natural log of odds is: . is the odds that someone defaults on their loan, and are the inverse of odds that someone defaults on their loan, or the odds that someone maintains their credit account. The logistic regression model is: and in the terms of natural log of odds it is: . The estimated coefficient for credit utilization is 31.209. This means that on average the change in log odds for defaulting is .31209 for each percentage increase in credit utilization. The confusion matrix is below:



We see 295 true positives, and 246 true negatives. We see 30 false positives (type 1 error) and 29 false negatives (type 2 error).

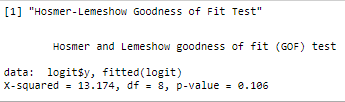
Accuracy: = =or 90.17% of correct predictions to total number of observations.

Precision: = = or 90.77% of correct positive predictions to the total predicted positives.

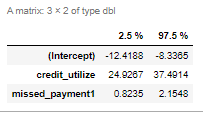
Recall: = = or 91.05% of correct positive predictions to the total positives examples.

### **Evaluating Model Significance**

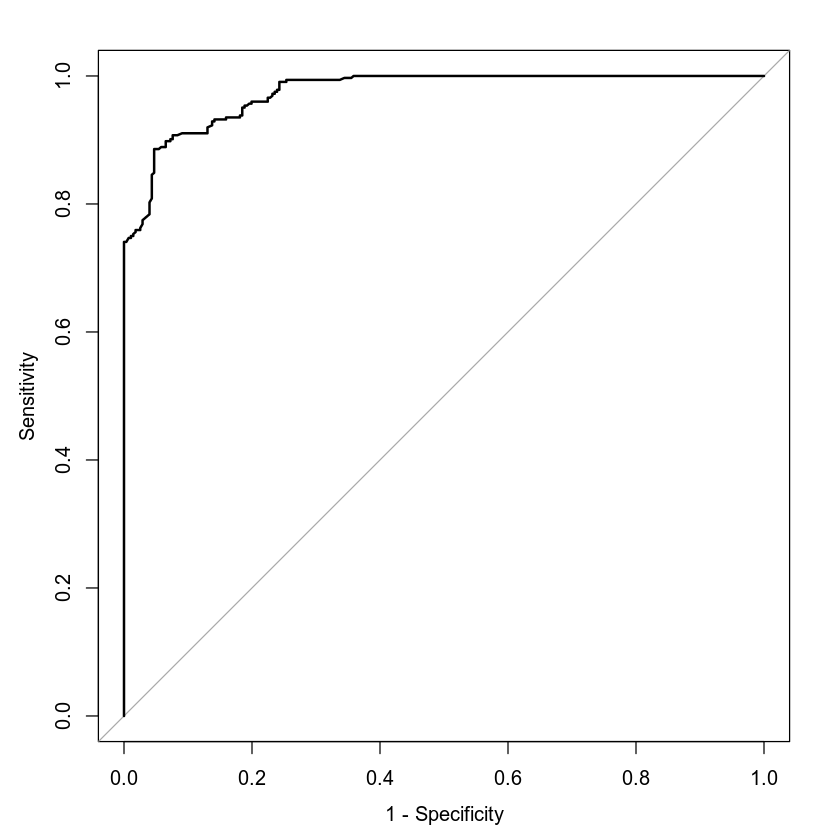
The null hypothesis (H0) is that there is no relationship between a customer defaulting on their loan, credit utilization, and if they’ve missed a payment in the past 3 months. The alternative hypothesis (Ha) is that there is a relationship between a customer defaulting on their loan, credit utilization, and if they’ve missed a payment in the past 3 months such that a prediction could be made on whether or not the customer will default on their loan. The Hosmer-Lemeshow goodness of fit test shows a p-value of 0.106 which does not pass the 5% level of significance test.



Based on Wald’s test none of the terms would be eligible for a slope of 0. All terms pass based on a 5% level of significance.



The Receiver Operating Characteristic curve is displayed below. This calculation measures the performance of a logistic regression model. Specifically it identifies how well the model predicts the classes of 0 or 1. The larger the area under the curve, the more accurate it is. This curve has an AUC of .9746 which states that 97.46% of the fitted values results fall within the curve.



### **Making Predictions Using Model**

The probability of an individual who has a credit utilization of 32% and has missed a payment in the past 3 months is 0.75 or 75%. Their odds of defaulting on a loan is 1 to 3 or the customer is 3 times as likely to default on their credit.

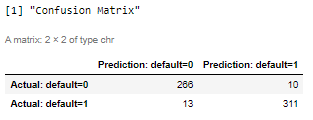
The probability of an individual who has a credit utilization of 32% and has not missed a payment in the past 3 months is .4035 or 40.35%. Their odds of defaulting on a loan is 1 to 0.68 or the customer is 0.68 times as likely to default on their credit.

One missed payment has a concerningly high probability of defaulting whereas by paying it drastically reduces the probability of defaulting by roughly 35%.

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

### **Reporting Results**

The general form of logistic regression model is: , where X1 is credit utilization, X2 is assets (car), X3 is assets (house), X4 is assets (house and car), X5 is education (graduate), and X6 is education (postgraduate). The model in natural log of odds is: . is the odds that someone defaults on their loan, and are the inverse of odds that someone defaults on their loan, or the odds that someone maintains their credit account. The logistic regression model is: and in the terms of natural log of odds it is: . The confusion matrix is below:



We see 311 true positives, and 266 true negatives. We see 10 false positives (type 1 error) and 13 false negatives (type 2 error).

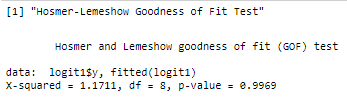
Accuracy: = =or 96.17% of correct predictions to total number of observations.

Precision: = = or 96.88% of correct positive predictions to the total predicted positives.

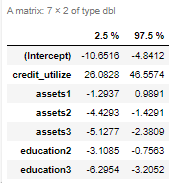
Recall: = = or 95.99% of correct positive predictions to the total positives examples.

### **Evaluating Model Significance**

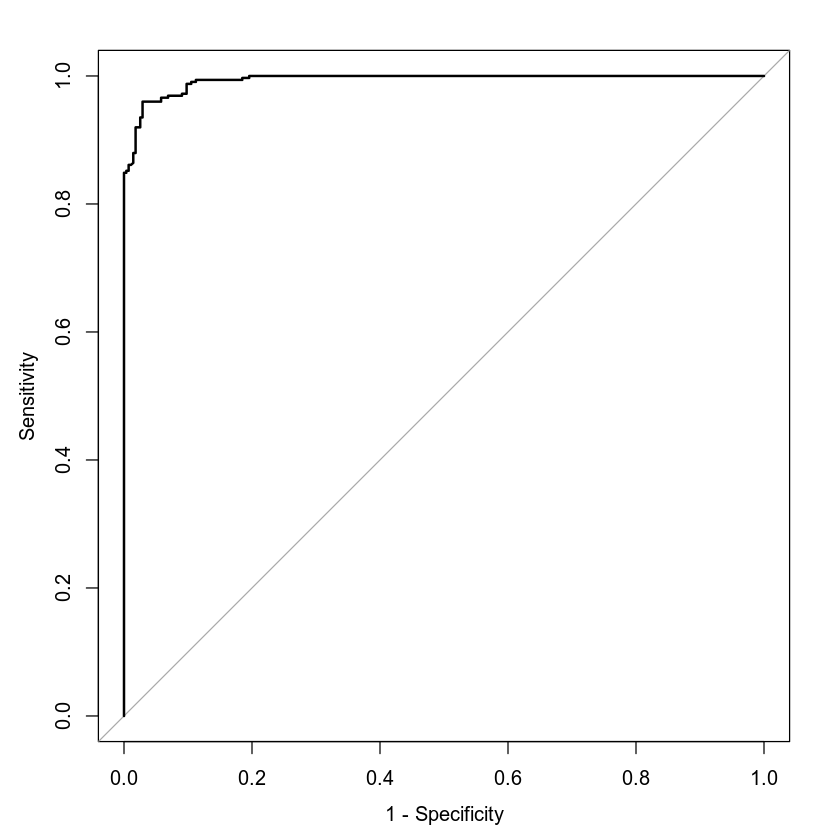
The null hypothesis (H0) is that there is no relationship between a customer defaulting on their loan, credit utilization, their assets, and their education. The alternative hypothesis (Ha) is that there is a relationship between a customer defaulting on their loan, credit utilization, their assets, and their education such that a prediction could be made on whether or not the customer will default on their loan. The Hosmer-Lemeshow goodness of fit test shows a p-value of 0.9969 which does not pass the 5% level of significance test.



Based on Wald’s test all terms pass with one exception, if the customer has a car only as an asset. There is a possibility it can have a slope of 0 between the range of (-1.2937, 0.9891).



The Receiver Operating Characteristic curve is displayed below. This calculation measures the performance of a logistic regression model. Specifically it identifies how well the model predicts the classes of 0 or 1. The larger the area under the curve, the more accurate it is. This curve has an AUC of .9936 which states that 99.36% of the fitted values results fall within the curve.



### **Making Predictions Using Model**

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 is 0.984 or 98.4%. Their odds of defaulting on a loan is 1 to 61.5 or the customer is 61.5 times as likely to default on their credit.

The probability of an individual who has a credit utilization of 43%, owns a car and a house, and has attained a postgraduate degree is 0.3468 or 34.68%. Their odds of defaulting on a loan is 1 to 0.53 or the customer is 0.53 times as likely to default on their credit.

Gaining an education up to postgraduate drastically reduces the probability of defaulting on a loan by over 60%.

## **5. Conclusion**

Of the two models evaluated, the second model that includes assets, education, and credit utilization is the one I would recommend, . The reason for that is because it contains a large portion of fitted values under the curve from the ROC. What this model does is utilize some customer attributes, assets, education, and credit utilization to identify the likelihood that they default on their loan using a logistic regression model. This model needs to be utilized because the response variable is a binary result, they either are likely or not likely to default. The only concerning result is that neither of the tested LR passed a Hosmer-Lemeshow goodness of fit test.

The practical importance of the analyses that have been performed are to identify which attributes best contributed to a logistic regression model. Using ROC as a means to evaluate the fitted values also helps determine how well the model fits. The selected model has a very high AUC at 99.36%. The Wald’s test did identify a term to remove, which was specifically any customer who has an asset of a car only. It is possible that only the ownership of a car has a minimal impact on the fit of the model due to Wald’s test.