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

Logistic Regression

Justin Farquhar

Justin.farquhar@snhu.edu

Southern New Hampshire University

## **1. Introduction**

The data set that we are exploring is for a credit card company through our role as a risk analyst. We are tasked with determining the relationships between varying customer characteristics and what effect they have on whether the customer is likely to default on their credit card or not. This could be used within the company to determine customer eligibility to qualify for a credit card, a credit line increase, and even adjusting API rates for the credit cards. For the most part within this problem set we will be utilizing logistic regression models utilizing wald confidence intervals, hosmer-lemeshow goodness of fit test, a confusion matrix, and obtaining a receiver operating characteristic curve.

## **2. Data Preparation**

Many of these variables within the data set are going to be important, but I believe the most important to be default as it is the main response variable. I also believe assets, missed\_payment, and credit\_utilize are all just as important or will at least play a larger role within the predictions. Assets will give a good initial indicator to the overall financial health of the individual, there are always going to be some outside factors that fall into it, but with predictions it is important to base off the known. The missed payments variable immediately shows the responsibility and timeliness of the customer as if they are missing payments already, the odds of defaulting would likely increase. Credit utilize shows how much of their available credit that is being used each much. Generally, the higher the percent utilization, the harder it is going to be to fully pay off your balance each month, which can easily snowball and build up until it is unmanageable, leading to defaulting. There are 8 columns and 600 rows within this data set.

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

### **Reporting Results**

General Form

Prediction Equation

Linear Form

Within the linear form we have the presence of the π, which represents the probability of the event occurring. We also have 1- π, which represents the probability of the event not occurring. The instance of function could realistically be replaced with odds in the linear form that was provided, but I wanted to leave them in to reference their meaning here.

To then move onto interpreting the estimated coefficient of credit utilization, the model gives us the value of 34.3869, which means that for every unit increase of credit utilization, the odds of defaulting increase by 34.3869. After obtaining the confusion matrix we can get values for accuracy, precision, and recall. In this model Accuracy is equal to 0.9416667, Precision is equal to 0.955836, and Recall is equal to 0.9351852.

### **Evaluating Model Significance**

The null hypothesis for a logistic regression model is always going to be that the model fits the data well and the alternative hypothesis will be that the model does not fit the data well. The P value that was returned was 0.9588, which is higher than the 5% level of significance, indicating that there is insufficient evidence to reject the null hypothesis, meaning that the model fits the data well. The test also returned a test statistic of 31.582 which is relatively lower than the Chi-square distribution of 47 degrees, which shows that there is no significant deviation from the observed and predicted value. Looking through each individual predictors significance level, we have a value of less than 2e-16 for credit\_utilize, a value of 0.00134 for education2, and a value of 9.72e-13 for education3, indicating that each predictor is significant at the 5% level of significance.

A graph with a line

Description automatically generated with medium confidence

The above image is the ROC curve for the model. Looking at the curve, we can see its quick rise to the top left corner, which suggests a high sensitivity and specificity, which indicates that is distinguishes between the classes well. We also get a AUC(area under the curve) from this, which was 0.9859 in this case. In the case of a perfect classifier, this value would be 1 while a random classifier would have a value of 0.5. Seeing as our AUC value is 0.9859, we have a classifier that is fairly close to perfect.

### **Making Predictions Using Model**

A white background with black text

Description automatically generated

The above image provides the results of both predictions that were made with this model. The first being the probability of an individual defaulting on a credit who has a credit utilization of 35% and has high school education. The model returned a 96.03% chance of this customer defaulting. This is an alarmingly high value, but a 35% utilization is extremely high especially if this is an average value and not just a one-off situation. The second prediction is the probability of an individual defaulting on credit who has a utilization of 35% and has post graduate education. The model returned a 25.59% chance of this customer defaulting. This is still a fairly high number, clearly better than the previous situation, but again looking at the very high 35% can be extremely hard to pay off in full unless it is properly accounted for ahead of time, or the individual has a much lower total line of credit in the first place.

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

### **Reporting Results**

General Form

Prediction Equation

Linear Form

Looking at the confusion matrix, the true positive is 303, the true negative is 262, the false positive is 14, and the false negative is 21. Plugging these values into the calculations gives us an accuracy of 0.9416667, a precision of 0.955836, and a recall of 0.9351852.

### **Evaluating Model Significance**

We then perform the Hosmer-Lemeshow goodness of fit test so that we can assess whether the model is appropriate for the data set. The P value that was returned was 0.9924, which is higher than the 5% level of significance, indicating that there is insufficient evidence to reject the null hypothesis, meaning that the model fits the data well. The test also returned a test statistic of 26.733 which is relatively lower than the Chi-square distribution of 47 degrees, which shows that there is no significant deviation from the observed and predicted value. Looking through each individual predictors significance level, we have a value of 6.51e-16 for credit\_utilize, a value of 0.3342 for assets1, a value of 5.05e-07 for assets2, a value of 2.61e-09 for assets3, and a value of 0.000549 for missed\_payment1. These values indicate that every predictor is significant at the 5% level of significance other than assets1.

A graph of a function

Description automatically generated

The above image is the ROC curve for the model. Looking at the curve, we can see its quick rise to the top left corner, which suggests a high sensitivity and specificity, which indicates that is distinguishes between the classes well. We also get a AUC(area under the curve) from this, which was 0.9874 in this case. In the case of a perfect classifier, this value would be 1 while a random classifier would have a value of 0.5. Seeing as our AUC value is 0.9874, we have a classifier that is fairly close to perfect.

### **Making Predictions Using Model**

The first prediction being the probability of an individual defaulting on a credit who has a credit utilization of 35, owns only a car, and has missed payments within the last three months. The model returned a 95.29 % chance of this customer defaulting. This is an alarmingly high value, but a 35% utilization is extremely high especially if this is an average value and not just a one-off situation. The second prediction is the probability of an individual defaulting on credit who has a utilization of 35%, owns a car and a house, and has not missed payments within the last three months. The model returned a 27.46% chance of this customer defaulting. This is still a fairly high number, clearly better than the previous situation, but again looking at the very high 35% can be extremely hard to pay off in full unless it is properly accounted for ahead of time, or the individual has a much lower total line of credit in the first place. Also in this case, owning a house and a car while also having not missed any recent payments shows a lot more financial responsibility.

## **5. Conclusion**

Looking through these results of the analyses, I would argue that both models could be recommended to use assuming that the sample size is sufficiently large. It would be very easy to run into the issue of bias depending on the sample size of data, but if the data is large enough and properly spread out this data can absolutely be utilized. With both models being significant enough to not be able to reject the null hypothesis as well as each area under the curve value being about as close to perfect as possible, these models are extremely beneficial. These results are great for the company, leading to much easier decisions when approving a customer for a credit card, as well as being able to provide resources to a customer were they be starting to collect some of the sure sign predictors of defaulting on their credit. This can greatly help earn trust within the customer base as well as increase our customer base.