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

Logistic Regression

Catherine Taft

Catherine.taft@snhu.edu

Southern New Hampshire University

**1. Introduction**

I had to swallow my distaste for the field of finance and am now a risk assessor at a credit card company. My parents have expressed some concern about my job hopping, but I press on. I have cats to feed. This data set contains some information about demographics and financial history about a group of people. We’ll analyze the relationships between these factors to determine if it’s likely that certain people would default on their debt, in which case the company would probably not give them a credit card, or would limit certain aspects of their account like the credit ceiling on the card. We’ll be using logistic regression models for this and utilizing some tests like the Hosmer-Lemeshow Goodness of Fit test.

## **2. Data Preparation**

The important variables here are age, sex, education, marriage status, assets owned, missed payments in the last 3 months, credit utilization, and whether or not they’ve defaulted on their debt. There are 600 rows and 8 columns in this dataset.

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

### **Reporting Results**

The general form of a logistic regression model for default potential using credit utilization and missed payments as variables would be:

Where Y is 0 for not defaulting and 1 for defaulting. X1 is credit utilization and X2 is missed payments.

Logistic regression is non-linear so to express this in linear form we’d write it in terms of the natural log of odds:

In terms of someone defaulting on their credit, the terms in this model represent:

* pi

   : the probability of someone defaulting on their credit
* Equation

  A fraction with pi in the numerator, and one minus pi in the denominator : The odds of the event happening (default)

In order to write the model equation we need to run it in R first:

Text, table

Description automatically generated

So now using this we can write the equation:

To rewrite it in terms of the natural log of odds we would have:

The estimated coefficient for credit utilization is 31.209, which means that the change in log odds for default is 0.31209 for each percent increase in credit utilization.

Here’s the confusion matrix for this model:

Graphical user interface, application

Description automatically generated

There are 246 true negatives, 29 false negatives, 295 true positives and 30 false positives.

**Accuracy**

The formula for accuracy is:

=

This evaluates to 0.90167 which is about 90%. That means that generally 90% of the time the model is correct in its classification.

**Precision**

The formula for precision is:

This evaluates to 0.9077 or about 91%. That means that 91% of the time if the model predicts default it’s correct.

**Recall**

The formula for recall is:

=

This evaluates to 0.9105 which is about 91%. This means that 91% of the time if the actual value is default, the model’s prediction is also default.

### **Evaluating Model Significance**

In order to evaluate the model significance we’ll do the Hosmer-Lemeshow goodness of fit test.

The null hypothesis is that the model fits the data and the alternative hypothesis is that it does not fit the data.

Text

Description automatically generated

The test statistic for this test is the X-squared value which here is 49.076. The p-value is 0.4298. That is greater than the significance level of 0.05, so we should not reject the null hypothesis, and can conclude that the model fits the data.

Wald’s Test is sort of like a beta test in that it can determine if the predictor variables are relevant to the model. The null hypothesis is that the coefficient of the variable is 0, meaning that it’s not relevant to the model. The alternative hypothesis is that it is not 0.

For this model we’re looking at credit utilization and missed payments. Based on the above R output, the p-value for credit utilization is 2 \* 10-16 and the p-value for missed payments is 1.16 \* 10-5. Both of those are below the significance level of 0.05 so we can reject the null hypothesis and conclude that they’re both relevant to the model.

The ROC curve and AUC values are here:

Chart

Description automatically generated

Text

Description automatically generated with medium confidence

The ROC describes how well the model distinguishes between the binary values in the classification (Y=0 or Y=1). The AUC value is 0.9746 which means that 97.46% of the area in the graph is under the curve. That is good – the higher the AUC value, the better the model is at classifying.

### **Making Predictions Using Model**

Using the model, we ran the probability of default for someone with a credit utilization of 32% who has missed payments within the last 3 months. That came out to 0.75 which means there’s a 75% chance of default for a person with those specific characteristics.

The odds of default for a person with the same credit utilization percentage who has not missed any payments is 0.4035, or 40.35%.

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

### **Reporting Results**

The general form of a model for default with credit utilization, assets and education as variables would be:

Here X1 is credit utilization, X2, X3 and X4 are the dummy variables for assets (car, house, both) and X5 and X6 are the dummy variables for education (college, grad school).

To express it in linear form we’d rewrite this as the natural log of odds:

To get the actual model equation we need to run it in R first:

Table

Description automatically generated

So based on this we can write the equation as:

And then again as natural log of odds:

Here’s the confusion matrix:

Table

Description automatically generated

The results of this are 311 true positives, 266 true negatives, 10 false positives, 13 false negatives.

**Accuracy**

The formula for accuracy is:

This evaluates to roughly 0.9617 or 96.17%. That means that in general the model is accurate in its classification 96.17% of the time which is impressive.

**Precision**

The formula for precision is:

This evaluates to roughly 0.9689 or 96.89%. This means that almost 97% of the time if the model predicts default it’s correct.

**Recall**

The formula for recall is:

This evaluates to roughly 0.9599 or 95.99%. This means that almost 96% of the time if the actual value is default, the model’s prediction is also default.

### **Evaluating Model Significance**

Now to evaluate the model significance we’ll do the Hosmer-Lemeshow GOF test again. The null hypothesis is that the model does fit the data, and the alternative hypothesis is that it does not fit the data.

Text

Description automatically generated

Here we can see the X2 value which is the test statistic is 18.432, and the p-value is 1. That is obviously higher than the significance value of 0.05, so we cannot reject the null hypothesis and can say that the model fits the data.

Now to see which terms are significant we’ll do the Wald’s test for each parameter. The null hypothesis for each test is that the coefficient is 0 and therefore not significant to the model. The alternative hypothesis is that it’s not 0 and therefore is significant.

Based on the R output above, the p-values are as follows: credit utilization: 3.56 \* 10-12, assets1(car): 0.79364, assets2(house): 0.00013, assets3 (car and house): 8.43 \* 10-8, education2(college): 0.00128, education3(grad school): 1.68 \* 10-9. So against a 0.05 significance level, we can see that all of these except for assets1 is below the significance level. In those cases we would reject the null hypothesis and conclude that they are significant to the model. In the case of assets1, we can accept the null hypothesis and conclude that it is not significant to the model.

That’s interesting to me because a car is a huge expense that I could see impacting your ability to pay down your credit card (priorities), which is one reason I do not ever want to own one, which I suppose is why I was fired from that car company a few weeks ago.

Next we’ll get the ROC curve and AUC value:

Chart

Description automatically generated

Text

Description automatically generated with medium confidence

So the AUC value is 0.9936 which means that 99.36% of the area of the graph is under the curve. This being a measure of how good the model is at classifying, the higher the AUC the better, this means that this model is really good at classifying.

### **Making Predictions Using Model**

Now we run a couple of predictions with this model. First we ran the probability of default for someone with a credit utilization of 43%, owns a car and a house and has a high school diploma. That comes out to 0.984, or 98.4%. That is actually kind of astonishing.

The second model was for someone with credit utilization of 43%, owns a car and a house, and has a postgrad degree. The odds of that occurring is only 0.3468. That is honestly shocking. I understand that education level is strongly tied to salary levels in many industries but that’s an enormous difference. I would love to see how this changes factoring student debt into it for those grad students.

## **5. Conclusion**

Based on these analyses I’d recommend the model (because it’s the type of model you’d use for this kind of analysis) but between the two I would definitely choose the second model. The accuracy, precision and recall as well as the AUC value were all much higher for the second model which indicates that it’s just more effective.

The results are unsurprising in general (that if you have a bunch of other bills to pay, a pattern of missing payments, and a high degree of credit utilization, that you’re more likely to default on payments) but the degree to which these results turned out is surprising, like I said. I do sort of question the model itself because I think there are a lot of other factors that can contribute to or indicate someone’s likelihood to keep up on their payments (medical debt, student debt, credit score, cost of living to income ratio, etc). As someone in the discussion board pointed out, the key factor in the enormous discrepancy between the two scenarios in the second model seems to be education level, which strikes me as a bit discriminatory.

The practical importance of this type of model is pretty clear – in general because you need to be able to evaluate the probability of binary outcomes sometimes, and specifically with risk analysis because when you are responsible for huge amounts of money, you want to make sure that you’re lending to appropriate people. That’s assuming good ethics though; the whole subprime situation is a separate conversation.