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

Logistic Regression

Zachary Tankersley

[Zachary.tankersley@snhu.edu](mailto:Zachary.tankersley@snhu.edu)

Southern New Hampshire University

## **1. Introduction**

The data being analyzed is from a historical data set of customer characteristics that will be leveraged to assess the relationships between these characteristics and whether or not a customer is likely to default on their credit. This information can be used to influence decisions made by the company in regard to increasing credit for customers, what their interest rates may be based on the risks, or just to see whether or not a customer is likely to default, among other insights. In this analysis, two logistic regression models will be created and assessed.

## **2. Data Preparation**

The primary variable of concern (the response variable) will be “default,” and the predictor variables will be credit\_utilize (an individual’s percentage of credit used), missed\_payment (whether or not they missed payments in the last 3 months), assets (assets owned by the individual), and education (highest level of education obtained). There are 600 rows of usable data with an additional row containing the header information, and there are 8 columns.

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

### **Reporting Results**

The general form of the equation for the first regression model for defaulting on credit using credit utilization and missed payments as predictors is:

The equation in terms of natural log odds which expressing the beta terms in linear form is:

The *π* shown in the equation above symbolizes the probability of defaulting, where the ratio is the probability of defaulting vs the probability of *not* defaulting. The true model for this equation is shown below after receiving the beta estimates from the summary function in R:

While the linear form is:

The coefficient of credit utilization is 31.2090 which indicates a highly positive correlation with defaulting on credit. The confusion matrix for this model is shown below:

|  |  |  |
| --- | --- | --- |
|  | Prediction: default=0 | Prediction: default=1 |
| Actual: default =0 | 246 (tn) | 30 (fp) |
| Actual: default=1 | 29 (fn) | 295 (tp) |

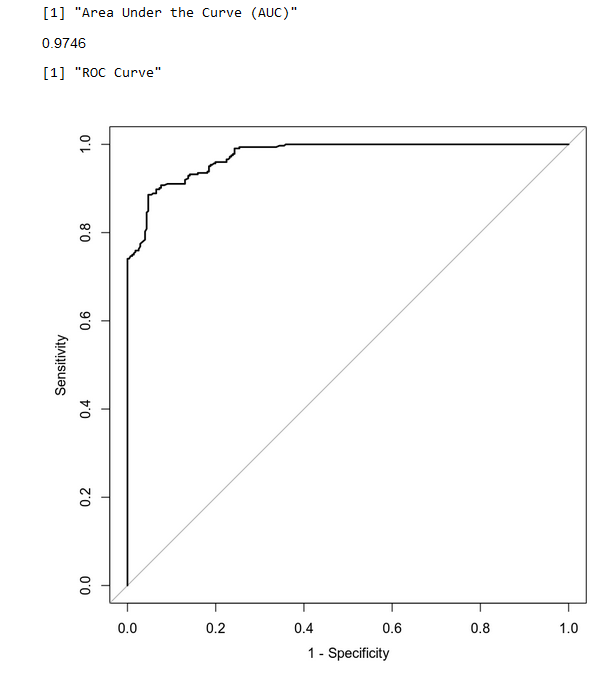
* Accuracy: ~0.9017
* Precision: ~0.9077
* Recall: ~0.9105

### **Evaluating Model Significance**

By performing a Hosmer-Lemeshow goodness of fit test the model can be assessed on its appropriateness for this data. The null and alternative hypotheses are:

The Chi-Square test statistic is 49.076 and the P-value is 0.4298. At a significance level of 5% or 0.05, the null hypothesis *is* met in this case, and the equation *does* fit the data. To identify individual terms significance, a Wald test is performed on both variables at a 5% level of significance. The null and alternative hypotheses for both variables are:

The beta parameters for credit utilization and missed payments are 2e-16 and 1.16e-05 respectively, meaning that in both cases the variables are statistically significantly related to defaulting. The null hypotheses can both be rejected in favor of the alternatives. Below can be seen an ROC curve which shows an incredibly small AUC (area under the curve) of 0.9746 meaning this model has a high accuracy.



### **Making Predictions Using Model**

Using this model, we can predict the probability that an individual with a credit utilization of 32% that has missed a payment in the past three months. The odds of this individual defaulting are 0.75 or 75%, compared to a person with no missed payments in the last three months with odds of defaulting at 0.4035 or 40.35%. Clearly this indicates that the likelihood of defaulting goes up if a payment is missed, even when credit utilization remains the same.

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

### **Reporting Results**

The general form of the second logistic regression model using credit utilization, assets, and education as predictor variables is:

Where x1 is credit utilization, x2, x3, and x4 are dummy variables for assets (car, house, car and house), and x5 and x6 are dummy variables for education levels. The linear form of the equation is:

The true form of this equation can be written using the parameter estimates from the summary function in R and written as:

Again, using the summary function in R, the parameter estimates can be used to build the true model which is:

The linear form of the equation, or the equation in terms of natural log of odds is:

The confusion matrix for this model is:

|  |  |  |
| --- | --- | --- |
|  | Prediction: default=0 | Prediction: default=1 |
| Actual: default =0 | 266 (tn) | 10 (fp) |
| Actual: default=1 | 13 (fn) | 311 (tp) |

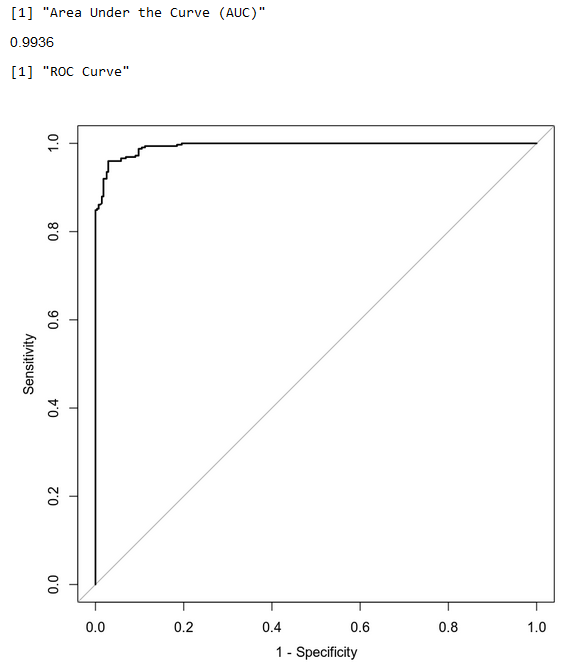
* Accuracy: ~0.9617
* Precision: ~0.9688
* Recall: ~0.9599

### **Evaluating Model Significance**

By performing a Hosmer-Lemeshow goodness of fit test the model can be assessed on its appropriateness for this data. The null and alternative hypotheses are:

The Chi-Square test statistic is 18.423 and the P-value is 1 on 48 degrees of freedom. At a significance level of 5% or 0.05, and a P-value *much greater,* the null hypothesis *is* met in this case, and the equation *does* fit the data. To identify individual terms significance, a Wald test is performed on all variables/parameters at a 5% level of significance. The null and alternative hypotheses for *all* the variables are:

Referencing the P-values in the summary analysis function in R for the second model, all P-values are much lower than the 0.05 significance level indicated, save for “assets1” which corresponds to the dummy variable for owning a car. , or “assets1” has a P-value of 0.79365 meaning that it is *not* significant in this model. The AUC for this model on an ROC curve is 0.9936 indicating that this model is more accurate than the first as its plot contains more area under its curve. The ROC curve is included below:



### **Making Predictions Using Model**

Using this model, the probability of an individual defaulting on their line of credit with a credit utilization rate of 43% who also owns a car and a house, and who has also attained a high school diploma is 98.4%, while the same individual having received post-graduate levels of education has only a 34.68% of defaulting on their credit line. The most immediate conclusion would be that a higher level of education, given a constant rate of credit utilization and similar assets, greatly reduces the probability of defaulting on a credit line by a consumer.

## **5. Conclusion**

Based on both analyses performed and assuming a sufficient sample size, I would have to recommend the second model, if any. The AUC on the ROC curve for the second model would indicate a *slightly* more accurate model than the first. With that said, both sets of predictions indicate an *extremely high* predicted probability of default. Leveraging real world data would suggest this model is not *entirely* accurate and different predictors may need to be used, or in different combinations. FICO says that the average credit utilization rate among Americans is slightly higher than 31% (Rossman, 2022), while census bureau data indicates that only 37.5% of Americans had a college degree (Statista, 2022). The second, more “accurate” model would indicate that the average American has about a 43.99% probability of defaulting on their credit, while CNBC states that the average *delinquency* rate is only about 5.32% (White, 2022). While CNBC’s reported delinquency rate seems rather low, the models predicted default probability also seems rather high, and while the model doesn’t predict the national average, only the individual, the model does *not* seem to be perfect. Further investigation is recommended.

The model may be used to gain a vague understanding of the influences certain predictors may have on credit default probability; however, the results of the model should be used with a modicum of caution.

## **6. Citations**

Rossman, T. (2022, September 9). *Average Credit Score Remains At Record High, But Warning Signs Abound*. Bankrate. Retrieved October 3, 2022, from https://www.bankrate.com/finance/credit-cards/average-credit-score-remains-at-record-high/

Statista. (2022, June 10). *Educational attainment in the U.S. 1960-2021*. Retrieved October 3, 2022, from https://www.statista.com/statistics/184260/educational-attainment-in-the-us/