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

Logistic Regression

Nolan Pratt

nolan.pratt@snhu.edu

Southern New Hampshire University

## **1. Introduction**

The data set being used for reference within this report is the “credit\_card\_default.csv” file that houses various credit card statistics that will be used to extract valuable data about financial trends. These results can be used to measure the interactions between socio-economic status and financial credibility, gain insight to the effects relative life scenarios have on an individual’s financial discretion, or extract financial trends based on current consumer choices that are led by modern social circumstances. In this instance, the data will be used to discover defaulting trends based on individual circumstances.

The data will be utilized to extract valuable statistics regarding the socio-economic differences of discrete individuals that are based on various characteristics derived from the party’s financial discretion and life circumstances. The statistics will be derived using various statistical analysis methods, including the following: logistic regression models, confidence intervals, Hosmer-Lemeshow Goodness of Fit tests, confusion matrices, and ROC curves.

## **2. Data Preparation**

The data has been housed within a CSV file, which contains 600 credit card records that are discerned by 8 data fields. The 8 data fields are denoted as the following variables:

Graphical user interface, text, application, email

Description automatically generated

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

### **Reporting Results**

The general, logistic regression equation is denoted as the following:

The logistic regression model needed for the analysis of defaulting on credit against credit utilization and education, is denoted as the following:

where the education2 predictor represents the choice of college, the education3 predictor represents the choice of post-graduate, the above odds denoted by the following ratio:

and the probability of the occurrence represented as:

The following regression model was generated from this scenario:

Table

Description automatically generated

Using the logistic regression model, the calculated beta estimates can be utilized to extract a prediction equation:

Following the extraction of the model’s prediction equation, the estimated coefficient of credit utilization of 34.3869 can be utilized to calculate the odds of defaulting for each unit increase within the predictor, when all other variables are held constant, using the following approach:

This formula represents the percentage increase in the odds of defaulting on credit for each percentage increase within a static model. This means that each unit increase of credit usage causes the odds of defaulting on credit to increase by 41.04%.

The confusion matrix can then be generated to confirm the Accuracy, Precision, and Recall, of the regression model. The following confusion matrix is calculated from the regression model:

Graphical user interface

Description automatically generated with medium confidence

The Accuracy, derived from the formula ((TP+TN)/(TP+TN+FP+FN)), Precision, derived from the formula (TP/(TP+FP)), and Recall, derived from the formula (TP/(TP+FN)), of the regression model is then calculated as the following:

A screenshot of a computer

Description automatically generated with low confidence

These calculated values represent the regression model’s performance of correct predictions (Accuracy), accuracy of the predicted positives (Precision), and ability to correctly identify positives (Recall), have been calculated to roughly 93% for each term. These values indicate that the logistic regression model correctly predicts outcomes roughly 93% of the time.

### **Evaluating Model Significance**

The model’s significance can be evaluated using the Hosmer-Lemeshow Goodness of Fit test, where the null hypothesis is the case that the logistic regression model fits the data well and the alternative hypothesis is the case that the logistic regression model does not fit the data well to a significant degree when compared to the significance level. The calculated p-value returned as 0.9676 from the test statistic of 31.582, which is significantly greater than the significance level of 0.05, therefore the null hypothesis should not be rejected. This concludes that the model is a good fit for the dataset. The following results were returned from the test:

Text, letter

Description automatically generated

The Wald Confidence intervals for slope parameters can test the fitness for each predictor’s beta estimates within the model to determine which parameters significantly impact the outcome of the response variable. Each predictor’s significance within the model is determined within this test by determining whether the predictor’s calculated confidence interval passes over zero. The model returned showcases intervals that do not include zero for each predictor within the model, therefore each term is significant. Using a 0.05 level of significance, the Wald test returned with the following predictor intervals:

Table

Description automatically generated

Following the prior tests for model fitness, the Receiver Operating Characteristic curve can be displayed as a graphical representation of the logistics regression model’s performance in distinguishing between positive and negative cases. The AUC is a simplified term for the Area Underneath the Curve within the graph. In this scenario, a differential solution was utilized to estimate the AUC to a value of 0.9859. This AUC value represents the percentage of accuracy within the model based on its generated prediction values, so the ROC curve showcases that nearly 99% of the predicted values were correctly distinguished by the regression model. The following depiction was generated for visualizing this trend:

Chart

Description automatically generated

### **Making Predictions Using Model**

Using the extracted prediction equation, the probability of an individual defaulting on credit who has a credit usage of 35% and a high school education can be calculated as follows:

Using the extracted prediction equation, the probability of an individual defaulting on credit who has a credit usage of 35% and a post-graduate education can be calculated as follows:

Comparing these two prediction values, two individuals who have the same credit usage, only differing in education, showcases a significantly varied probability and odds for defaulting. The individual with the high school education has a probability to default at nearly 100%, which was calculated from the returned odds of 24.2064 for defaulting, whereas the individual with the post-graduate education has a significantly lower probability to default at roughly 25%, which was calculated from the returned odds of 0.3439 for defaulting.

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

### **Reporting Results**

The general, logistic regression equation is denoted as the following:

The logistic regression model needed for the analysis of defaulting on credit against credit utilization, assets, and missed payment, is denoted as the following:

where the assets1 predictor represents the possession of only a car, the assets2 predictor represents the possession of only a house, the assets3 predictor represents the possession of both a car and a house, the missedpayment1 predictor represents whether the individual missed a payment within the past 3 months, the above odds denoted by the following ratio:

and the probability of the occurrence represented as:

The following regression model was generated from this scenario:

**Table

Description automatically generated**

Using the logistic regression model, the calculated beta estimates can be utilized to extract a prediction equation:

The confusion matrix for the second model can then be generated to confirm the Accuracy, Precision, and Recall, of the regression model. The following confusion matrix is calculated from the regression model:

Table

Description automatically generated with low confidence

The Accuracy, derived from the formula ((TP+TN)/(TP+TN+FP+FN)), Precision, derived from the formula (TP/(TP+FP)), and Recall, derived from the formula (TP/(TP+FN)), of the regression model is then calculated as the following:

A picture containing text, orange

Description automatically generated

### **Evaluating Model Significance**

The model’s significance can be evaluated using the Hosmer-Lemeshow Goodness of Fit test, where the null hypothesis is the case that the logistic regression model fits the data well and the alternative hypothesis is the case that the logistic regression model does not fit the data well to a significant degree when compared to the significance level. The calculated p-value returned as 0.9945 from the test statistic of 26.733, which is significantly greater than the significance level of 0.05, therefore the null hypothesis should not be rejected. This concludes that the model is a good fit for the dataset. The following results were returned from the test:

Text

Description automatically generated

The Wald Confidence intervals for slope parameters can test the fitness for each predictor’s beta estimates within the model to determine which parameters significantly impact the outcome of the response variable. Each predictor’s significance within the model is determined within this test by determining whether the predictor’s calculated confidence interval passes over zero. The model returned showcases intervals that do not include zero for each predictor within the model, except for the assets1 predictor, which passes over zero within its interval, therefore each term is significant, aside from the assets1 predictor. Using a 0.05 level of significance, the Wald test returned with the following predictor intervals:

Text, table

Description automatically generated

Following the prior tests for model fitness, the Receiver Operating Characteristic curve can be displayed as a graphical representation of the logistics regression model’s performance in distinguishing between positive and negative cases. The AUC is a simplified term for the Area Underneath the Curve within the graph. In this scenario, a differential solution was utilized to estimate the AUC to a value of 0.9874. This AUC value represents the percentage of accuracy within the model based on its generated prediction values, so the ROC curve showcases that nearly 99% of the predicted values were correctly distinguished by the regression model. The following depiction was generated for visualizing this trend:

Chart

Description automatically generated

### **Making Predictions Using Model**

Using the extracted prediction equation, the probability of an individual defaulting on credit who has a credit usage of 35%, owns only a car, and has missed payments within the past 3 months, can be calculated as follows:

Using the extracted prediction equation, the probability of an individual defaulting on credit who has a credit usage of 35%, owns both a car and a house, and has not missed payments within the past 3 months, can be calculated as follows:

Comparing these two prediction values, two individuals who have the same credit usage, while differing in assets and missed payments, showcases a significantly varied probability and odds for defaulting. The individual with only a car who missed payments within the past 3 months has a probability to default at 100%, which was calculated from the returned odds of 22.2208 for defaulting, whereas the individual with both a car and a house whom did not miss payments within the past 3 months has a significantly lower probability to default at roughly 56%, which was calculated from the returned odds of 0.2478 for defaulting.

## **5. Conclusion**

When comparing the two models for utility within real-world scenarios, I would recommend both, though my recommendation depends on the scenario where the model is being applied. Both models offer valuable insights into the defaulting trends that present through credit card usage, accompanied by other external factors, however some applications would only be suitable for specific scenarios. For instance, the first model predicts defaulting trends for individuals who differ only in relative education and credit usage, which would mainly apply to scenarios where the study is concerned with only students, and the second model would only be applicable in studies where adult asset ownership and responsible payment practices are concerned. Both models showcase valuable statistics, however they would be best implemented in separate studies, rather than together. I believe a more complete model would be a combination of the two to include both predictor groups into a single cluster to better identify defaulting trends across the whole of the population, rather than only within specific subsets since this dataset is involved with socio-economic characteristics – variability in one aspect can greatly influence another since these are non-interchangeable and non-negotiable under normal living standards.

The analyses performed within this report showcase defaulting trends regarding the current credit state and can offer various insights within multiple scenarios. For instance, we know from the first model that as the education of an individual increases, they become less likely to default on their credit. This can be useful in determining which subset of students would harbor the least risk factors for credit companies. The second model showcases the behavior and interaction of asset ownership in combination with missing credit payments. The implementation of this model within credit companies would allow for better predictions when lending credit lines to adults, who have varying socio-economic circumstances. These are the two main applications for the models within this report, however I am sure that more than these can be realized. Though as I mentioned previously, I believe the best model to encompass these aspects, when concerned with calculating the risk-factor involved in lending credit to new customers, should also include the other predictors within the data since these ulterior factors also play a role within credit defaulting, and although these extra predictors are omitted from the two previously-constructed models, that does not mean that they did not also play a role within the defaulting trends present within the dataset.