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

Logistic Regression

Tammy Hartline

[tammy.hartline@snhu.edu](mailto:tammy.hartline@snhu.edu)

Week 5

Southern New Hampshire University

## **1. Introduction**

*For this week’s analysis we are exploring the credit\_card\_default dataset. It contains historical data related to credit card customers and their credit behavior. The dataset includes prediction variables, such as education level, assets owned, missed payments in the past three months, and credit utilization. The response variable being used is the default variable. (a binary value indicating whether the customer has defaulted on their credit)*

*The results of this analysis have one primary use as a risk analysis. It is used to assess the creditworthiness of individuals, helping financial institutions make informed decisions about lending money, including who they lend to, and how much the customer qualifies for, with the least chance of the customer defaulting on their line of credit. (Not paying it back) This analysis can also be used for risk assessments, developing credit scoring models, and identifying strategies to mitigate credit default risks.*

*The analysis may also include other statistical procedures, such as model validation techniques to assess the model’s predictive performance and statistical tests to evaluate the significance of predictor variables.*

## **2. Data Preparation**

*In this dataset there are 600 rows and 8 total columns. The important variables are the predictor variables education and credit utilization, as well as the defaulting on credit (default) response variable. The education and default variables both needed converting to factors. The data will now need to be explored to determine the relationships between these important variables. Specifically, how education and credit utilization affect the likelihood of the person defaulting on their credit.*

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

### **Reporting Results**

*The general form of this regression model is:*

*A math equations on a white background

Description automatically generated*

*Where y is 1 for defaulting on credit is true, and 0 for defaulting on credit is false. Variable X1 represents credit\_utilization, and X2 and X3 are dummy variables for education1 and education2.*

*The prediction equation of this model in terms of the natural log of odds to express the beta terms in linear for is expressed:*

*A group of math symbols

Description automatically generated*

*where ln is log, (pi / 1 – pi) represents odds of defaulting = 1 and odds of defaulting = 0. So this term represents the odds of an individual defaulting on their credit. Beta0 represents the intercept term of the logistic regression model which represents the estimated log-odds of defaulting on credit when all predictor variables are set to 0. Also, where beta1, the coefficient for X1, which is credit\_utilization, represents the change in the log-odds of the event (defaulting on credit) for a one-unit change in the corresponding predictor variable, holding all other variables constant. If beta1 is negative, it implies that an increase in credit utilization is associated with a decrease in the log-odds of defaulting on credit. Lastly, where Beta2 and Beta3, coefficients for X2 and X3, which represent education, are similar to beta1, as these coefficients represent the change in the log-odds of the event for a one-unit change in the corresponding predictor variable (education), while holding all other predictor variables constant. It quantifies how education influences the log-odds of defaulting on credit.*

*Pi represents the probability of an individual defaulting on their credit based on the logistic regression model. It is the estimated probability that the response variable, (default, Y), is equal to 1. The Pi/ 1-Pi term represents the ratio of the probability of default (i.e. Pi), to the probability of not defaulting, (i.e. 1- Pi).*

*The values obtained for the prediction model equation based on the coefficients where all values are rounded four decimal places are:*

*Intercept (Beta 0) = -8.8488, the coefficient for credit utilization (beta 1) = 34.3869, the coefficient for education level2 (beta 2) = -1.4976, and lastly the coefficient for education level3 (beta 3) = -4.254. These values give us this equation:*

**

*To determine accuracy (overall correctness of model’s predictions), precision (proportion of true positive predictions among all positive predictions made by the model), and recall (proportion of true positive predictions among all ACTUAL positive cases), we use a confusion matrix, which counts true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).*

*Based on the confusion matrix we get the following values:*

*True positives (TP) = 303, True Negatives (TN) = 254, False Positives (FP): 22, and False Negatives (FN) = 21.*

*For accuracy, we would use this equation:*

*A line with black text

Description automatically generated*

*with the values from the confusion matrix. Therefore accuracy is determined as:*

*Accuracy = ((303 + 254) / (303 + 254 + 22 + 21)) = (557 / 600) = (rounded 4 decimal places) -> 0.9283*

*This indicates the accuracy of the model is approximately 0.9283, which translates to it correctly predicts around 92.83% of the cases.*

*For precision, we would use this equation:*

*A mathematical equation with black text

Description automatically generated*

*with the values from the confusion matrix. Therefore precision is determined as:*

*Precision = ((303) / (303 + 22)) = (303 / 325) = (rounded 4 decimal places) -> 0.9308*

*This indicates that when the model predicts an individual will default on their credit account, (positive prediction), the prediction is correct around approximately 0.9308, which translates to 93.08% of the time.*

*For recall, we would use this equation:*

*A mathematical equation with black text

Description automatically generated*

*with the values from the confusion matrix. Therefore recall (model sensitivity) is determined as:*

*Recall = ((303) / (303 + 21)) = (303 / 324) = (rounded 4 decimal places.) -> 0.9375*

*This indicates the recall of the model is approximately 0.9375, which translates to the model correctly identifying about 93.75% of the actual cases where individuals defaulted on their credit.*

*Therefore, the logistic regression model achieved an approximate accuracy of 92.83%, precision of 93.08%, and recall of 93.75% based on the confusion matrix.*

### **Evaluating Model Significance**

*The Hosmer-Lemeshow Goodness of Fit Test was performed to assess the suitability of the logistic regression model for the dataset. The null hypothesis (H0) stated that the model fits the data well, while the alternative hypothesis (H1) posited that the model does not fit the data effectively. The test statistic was calculated as X-squared = 31.582 with 47 degrees of freedom, yielding a p-value of 0.9588. At a 5% level of significance, the p-value was compared to alpha (0.05). Since the p-value (0.9588) was greater than alpha, we failed to reject the null hypothesis, indicating that the logistic regression model is appropriate for the dataset. There is no compelling evidence to suggest that the model does not fit the data well.*

*To identify significant terms in the model based on Wald's test, one typically examines the p-values associated with each coefficient in the logistic regression output. Coefficients with p-values less than the chosen significance level (e.g., 0.05) are considered statistically significant predictors. The ROC curve, a graphical representation of the model's performance, was obtained. It illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at various threshold settings. A well-performing model will have an ROC curve that approaches the upper-left corner, indicating high sensitivity and low false positive rates.*

*A graph of a function

Description automatically generated*

*The value of AUC (Area Under the ROC Curve) summarizes the model's ability to distinguish between classes. Higher AUC values (e.g., AUC > 0.7) suggest better discriminatory power, signifying an effective model for classifying outcomes in the given dataset.*

### **Making Predictions Using Model**

*In the context of the logistic regression model, we have made predictions regarding the probability of individuals defaulting on credit based on their education level and credit utilization. For an individual with a high school education and a credit utilization of 35%, the model predicts an extremely high probability of default, approximately 99.27%, and the odds of defaulting are similarly elevated, around 133.3333. This suggests that individuals with these specific characteristics are strongly inclined to default on their credit, according to the model.*

*Conversely, for an individual holding a postgraduate degree and with a credit utilization of 35%, the model predicts a significantly lower probability of defaulting on credit, approximately 25.59%. The odds of defaulting for this group are also considerably lower, estimated at approximately 0.3421. This implies that individuals with postgraduate education and this credit utilization level are less likely to default on their credit, as indicated by the model.*

*These predictions exemplify how the logistic regression model employs education and credit utilization as predictor variables to estimate the likelihood of credit default. Higher probabilities and odds signify a greater propensity for default, while lower values suggest a reduced likelihood. It is important to remember that these model-based predictions are based on the available data and model assumptions and may not encompass all real-world factors influencing credit default. Actual individual circumstances should be thoroughly considered in credit-related decisions.*

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

### **Reporting Results**

*Important variables output value explanation: Education has an output of 1 where highest level of education attained is high school, 2 where highest education level is college, and 3 where highest education level is post-graduate. Assets has an output of 0 where no assets are owned by person (none), 1 if they ONLY own a vehicle, 2 if they ONLY own a home, and 3 if they own both a vehicle and a home. Missed payments has an output of 0 if the individual has missed no payments within the last 3 month, and 1 if they have missed ANY payments within the past 3 months. Credit utilization (credit\_utilize) variable value will vary as it is the average amount of credit being used by the individual out of the total allowed. Lastly, the resource variable, default, has a value of 0 if the person has not defaulted on their credit, and a value of 1 if the person has defaulted on their credit.*

*The general form of this regression model is:*

*A math equations on a white background

Description automatically generated*

*Where y is 1 for defaulting on credit is true, and 0 for defaulting on credit is false. Variable X1 represents credit\_utilization, X2 represents factor assets, and X3 represents factor missed\_payment. E(default) is the probability of defaulting on credit and beta0, beta1, beta2, and beta3 are the coefficients of the model. However, I need to mention that we can add in dummy variables for assets 1, assets 2, and assets 3, and where it is needed if missed\_payment has occurred more than one time in the past three months.*

*The prediction equation of this model in terms of the natural log of odds to express the beta terms in linear form is expressed:*

*A group of math symbols

Description automatically generated*

*Where pi represents the probability of an individual defaulting on their credit based on the logistic regression model. It is the estimated probability that the response variable, (default, Y), is equal to 1. The Pi/ 1-Pi term represents the ratio of the probability of default (i.e. Pi), to the probability of not defaulting, (i.e. 1- Pi).*

*In this specific model, the precision model would be expressed as:*

*The logistic regression model coefficients, along with their confidence intervals, are as follows:*

*The intercept is between -11.6518 and -6.8224. Where credit\_utilize, is between 24.4513 and 40.1140, assets = 1 (car only) is between -1.4624 and 0.4971, assets = 2 (house only) is between -4.2167 and -1.8501, assets = 3 is between -4.5947 and -2.3189, and lastly where missed\_payment = 1 (yes) is between 0.6178 and 2.2373. These coefficients represent the estimated relationships between the predictor variables and the log-odds of defaulting on their credit.*

*Next, I populated the confusion matrix, which provided counts for the model’s predictions compared to the actual outcomes. This resulted in the following values:*

*A screenshot of a test

Description automatically generated*

*True positives (TP) = 303, true negatives (TN) = 262, false positives (FP) = 14, and false negatives (FN) = 21. Using those counts, we can calculate the following performance metrics:*

*A line with black text

Description automatically generated*

*which would be accuracy = (303 + 262) / (303 + 262 + 14 + 21) = 0.9458 or 94.58% accuracy. This indicates the accuracy of the model is approximately 0.9458, which translates to it correctly predicts around 94.58% of the cases.*

*A mathematical equation with black text

Description automatically generated*

*which would be precision = 303 / (303 + 14) = 0.9558 or 95.58% precision. This indicates that when the model predicts an individual will default on their credit account, (positive prediction), the prediction is correct around approximately 0.9558, which translates to 95.58% of the time. Lastly,*

*A mathematical equation with black text

Description automatically generated*

*which would be recall = 303 / (303 + 21) = 0.9356 or 93.56%. This indicates the recall of the model is approximately 0.9356, which translates to the model correctly identifying about 93.56% of the actual cases where individuals defaulted on their credit.*

*These metrics provide an assessment of the model’s performance in predicting credit defaults. It appears to have high accuracy, precision, and recall, indicating that it is performing well in classifying default and non-default cases.*

### **Evaluating Model Significance**

*After performing a Hosmer-Lemeshow Goodness of Fit Test, these are our calculations:*

*A close-up of a computer code

Description automatically generated*

*Test statistic (X-squared) = 26.733, degrees of freedom (DF) = 47, and p-value = 0.9924.*

*This goodness test evaluates how well the logistic regression model fits the data. The null hypothesis (HO) is that the model fits the data well, while the alternative hypothesis (Ha), is that the model does not fit the data well.*

*With a p-value of 0.9924, which is significantly higher than the 5% level of significance, we accept the null hypothesis, and reject the alternative hypothesis. This suggests that the logistic regression model fits the data well, which indicates that the model is indeed appropriate for the dataset.*

*Therefore, this model’s performance, when evaluating the null hypothesis, appears to be quite good, with a high number of true positives, (correctly predicted defaults) and true negatives, (correctly predicted non-defaults). This was further verified and made more comprehensive upon evaluation of the accuracy, precision, and recall values.*

*This is then further confirmed using the AUC-ROC, where the AUC value = 0.9874, which is very close to 1. This indicates the logistic regression model has excellent discriminatory power and is highly effective at distinguishing between individuals who default on their credit (case = 1) and those do not (control = 0). Therefore the AUC value indicates that the model’s predictions align very well with the actual outcomes.*

*The ROC curve, along with the AUC value, further confirm the model’s strong performance, by being visually appealing with a steep rise in the true rate (sensitivity) and a low false positive rate (1 – specificity) as the threshold changes.*

*A graph of a function

Description automatically generated*

*Overall, based on the AUC value and the ROC curve, it can be concluded that the logistic regression model we have built is highly effective for predicting credit defaults, and it provides a strong separation between positive and negative cases. This is indicative of a robust and accurate predictive model.*

### **Making Predictions Using Model**

*In scenario one, we calculate the probability of an individual who has a credit utilization of 35%, owns only a vehicle, and has missed a payment in the last three months, defaulting on their credit. The probability in this scenario is approximately 0.9529, indicating there is a high probability of this individual defaulting on their credit when they have these characteristics. The odds of this event occurring, is approximately 20.2210, suggesting for every 21 individuals in this scenario, an estimated 20 of those are likely to default on credit.*

*In scenario two, we calculate the probability of an individual who has a credit utilization of 35%, owns a home and a vehicle, and has not missed any payments in the past three months, defaulting on their credit. The probability in this scenario is approximately 0.1986, indicating there is a low probability of this individual defaulting on their credit when they have these characteristics. The odds of this event occurring is approximately 0.2478, suggesting that the odds of this are less than 1 that individuals in this scenario, meaning they are very unlikely to default on their credit.*

*A white paper with black text

Description automatically generated*

*These two probabilities and odds provide insights into the risk factors associated with credit default and can be valuable for decision making and risk assessment in lending or credit-related contexts.*

## **5. Conclusion**

*Based on the analysis, assuming a sufficiently large sample size, it is recommended to use the logistic regression model for assessing credit default risk. The model demonstrates high accuracy, precision, and recall. It fits the data well according to the Hosmer-Lemeshow test, and the AUC value suggests excellent discriminatory power.*

*In practical terms, these results mean that the model effectively predicts credit defaults based on education, credit utilization, assets, and missed payments. It is important for financial institutions in making informed lending decisions and risk assessments.*

*The practical importance of these analyses lies in their ability to aid financial institutions in reducing credit default risk, enhancing credit scoring models, and making more informed lending decisions. Additionally, it contributes to improved risk management and customer-specific credit limit determinations.*