ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.8, 6027–6034
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.8.71 Special Issue on ICMA-2020

LOGISTIC REGRESSION MODELS FOR PREDICTION LOAN DEFAULTS-QUALTITATIVE DATA ANALYSIS

E. ELAKKIYA, K. RADHAIAH, AND G. MOKESH RAYALU[1]

ABSTRACT. Regression analysis is one of the statistical techniques where we can model the relationship between the dependent variable and the independent variable(s). Before developing the Logistic regression, it will be useful to understand the linear regression first and study why we cannot use that technique instead of logistic regression. In the linear regression, we develop the line of best fit that describes the relationship between the dependent variable and the independent variable(s) and there are few assumptions that needs to be followed in order to get consistent predictions from the line of best fit.

1. INTRODUCTION

In the linear regression case, the dependent variable is a continuous variable but in the logistic regression, the dependent variable is a dichotomous or qualitative variable. So, the dependent variable can take only two values like responder or non-responder. If we fit a linear regression model for this binary data, then we will get the following regression line, [1–3, 9].

$$y_i = X_i^{'}\beta_i + \epsilon_i \tag{1.1}$$

Here, $X_i^{'}$ are independent variables, β_i are regression coefficients and ϵ_i are errors, [5]. Since y_i can take only two value either 1 or 0.

[1]*corresponding author*

2010 *Mathematics Subject Classification.* 05C10.

Key words and phrases. regression, logistic regression.

$\epsilon_i = 1 - X_i'\beta_i$ when $y = 1$, $\epsilon_i = -X_i'\beta_i$ when $y = 0$ (ϵ_i can take only two values.)

In [7], the logistic response function is a nonlinear monotonic increasing (decreasing) S-shaped function and it has the following form.

$$y_i = \frac{e^{(x'\beta)}}{1 + e^{(x'\beta)}} = \frac{1}{1 + e^{-x'\beta}}. \tag{1.2}$$

The other challenge we have here, is the boundaries of the dependent variable y. But the logistic response function will take values between 0 and 1. We can overcome this problem by linearizing the logistic response function. One approach is to transform the basic structure of the logistic response function. We can transform the logistic function as follows,

$$Z = (x'\beta)z = \log(\frac{p}{1-p}) \tag{1.3}$$

where z is the linear predictor. This transformation is known as the logit transformation of the probability p, and the ratio $(\frac{p}{1-p})$ is known as the odds ratio and the equation (1.3) is known as the log of the odds ratio. Now, we can check the bounds of z by taking $p = 0$ and $p = 1$.

If $p = 1$, then $z = \log(\frac{1}{1-1}) = \log(\frac{1}{0}) = \log(1) - \log(0) = 0 - (\infty) = +\infty$ since $\log(0)$ is defined as $-\infty$.

If $p = 0.5$, then $z = \log(\frac{0.5}{1-0.5}) = \log(\frac{0.5}{0.5}) = \log(1) = 0$.

If $p = 1$, then $z = \log(\frac{0}{1-0}) = \log(\frac{0}{1}) = \log(0) - \log(1) = -\infty$ since $\log(0)$ is defined as $-\infty$.

So, the newly transformed log-odds transformation function z will take values between $-\infty$ to $+\infty$ and thus y_i ranges from 0 to 1. After the transformation, we can re-write the equation (1.3) as:

$$y_i = \frac{1}{1 + e^{(-\log(\frac{p}{1-p}))}} = \frac{1}{1 + e^{-z}}. \tag{1.4}$$

The fact that the logistic function y_i ranges between 0 and 1 is the primary reason why the logistic model is so popular. The model is designed to describe a probability, which is always some number between 0 and 1. To obtain the logistic model from the logistic function, we write z as the linear sum:

$$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k, \tag{1.5}$$

where the x_i are independent variables of interest and α is the constant term and $\beta_i^{'}s$ are the regression coefficients.
The logistic model considers the following general epidemiologic study framework: We have observed independent variables, $x_1, x_2, x_3..., x_k$ on a group of test subjects, for whom we have also determined disease status, as either 1 if with disease or 0 if without disease. We wish to use this information to describe the probability that the disease will develop during a defined study period. The probability being modeled can be denoted by the conditional probability statement $P(D = 1|x_1, x_2, x_3, ..., x_k)$. Thus, the logistic model may be written as:

$$P(D = 1|x_1, x_2, x_3, ..., x_k) = \frac{1}{1 + e^{(-\alpha + \sum_1^k \beta_i x_i)}}. \tag{1.6}$$

2. Logistig Regression Parameter Estimation

We use the maximum likelihood estimation method to estimate the parameters of the logistic regression model, [6, 7]. The general form of the logistic regression model is

$$y_i = E(y_i) + \epsilon_i. \tag{2.1}$$

Here the y_i are the observed independent Bernoulli random variables and has the expected value as

$$E(y_i) = p_i = \frac{e^{(x_i^{'}\beta)}}{1 + e^{(x_i^{'}\beta)}}. \tag{2.2}$$

Since each observation of the model follows a Bernoulli distribution, the probability distribution of each observation can be written as

$$f(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i}, i = 1, 2, 3...n. \tag{2.3}$$

As we know these $y_i^{'}$ s are probabilities, they can take values between 0 and 1 and, they are independently distributed. We can write the likelihood function for these observations as

$$L(y_1, y_2, y_3, ...y_n, \beta) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

If we consider a log on both sides,

$$\log(L(y_1, y_2, y_3, ...y_n, \beta)) = \log(\prod_{i=1}^{n} f_i(y_i)) \tag{2.4}$$

$$= \sum_{i=1}^{n} [y_i \log(\frac{p}{1-p_i})] + \sum_{i=1}^{n} \log(1-p_i) \tag{2.5}$$

$$= \sum_{i=1}^{n} y_i x_i' \beta - \sum_{i=1}^{n} \log[1 + e^{(x_i'\beta)}]$$

Since, $(1-p_i) = [1+e^{x_i'\beta}]^{-1}$ and $z_i = \log\left(\frac{p_i}{1-p_i}\right) = x^{i\beta}$

$$\widehat{p_i} = \frac{e^{x_i\beta}}{1 + e^{-\left(-x_i'\hat{\beta}\right)}} \tag{2.6}$$

We can use equation (2.4) to get the maximum likelihood estimates of the logistic regression model parameters, [4, 8]. The estimated value of the linear predictor is: $\left(\hat{Z}_i\right) = x_i'\hat{\beta}$, and the predicted probabilities are

$$\widehat{p_i} = \frac{e_i\beta}{1 + e^{-(-x'_i\hat{\beta})}} \tag{2.7}$$

TABLE 1. Table 2.1 Percentage of Loans and Loan Amount by Job Grade And Loan Status

Job Grad Loan Status	Loan Amount(In Dollars)		Percentage of Total Loans	
	Default	Fully Paid	Default	Fully Paid
B1	10,243,150	44,510,500	15.22%	84.78%
B2	11,656,350	46,574,750	17.19%	82.81%
B3	9,745,600	34,481,675	18.64%	81.36%
B4	14,236,100	48,071,550	20.80%	79.20%
B5	17,004,950	49,525,600	22.72%	77.28%
C1	17,168,050	47,380,925	23.50%	76.50%
C2	17,260,750	43,515,900	25.93%	74.07%
C3	19,107,250	44,772,000	27.16%	72.84%
C4	19,344,950	43,074,075	28.39%	71.61%
C5	20,456,950	38,654,850	31.88%	68.12%

In table 2.1, we can see the percentage of loans, total loan amount and loan status. If we observe the percentage of default loans, for the job grade C5, we have the default percentage of 31.88 which is accounting for 20 million dollars followed by the job grade C4, which has the default percentage of 28.39 and accounting for around 19 million dollars. The job grade B1 has the lowest default percentage of 15.22, compared to all other job grades and accounting for the default amount of 10 million dollars.

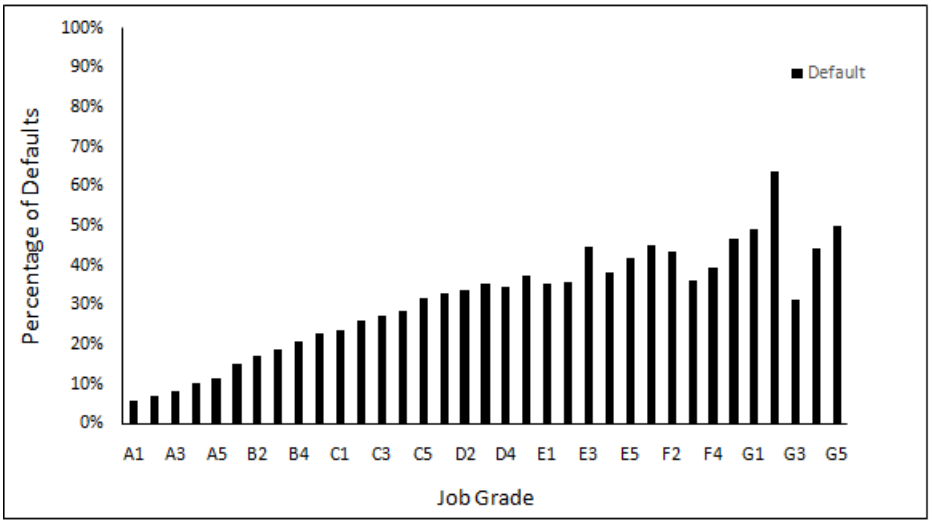

Figure 2.1 Graph showing the percentage of default loans by job grade

In figure 2.1, we have the percentage of defaulted loans by job grade, here we can observe that the defaults percentage is increasing as the job grade is decreasing, which is suggesting that the job grade and the defaults have close relationships and we can use this variable as an independent variable to predict the defaults.

In table 2.2, we can see the total number of loans and total loan amount by loan purpose. Here we can observe that the most number of loans are taken for the purpose of Credit Card Refinancing and Debt Consolidation. The total number of loans given to these two purposes are 56,292 which are accounting for around 903 million dollars and this is 81% of the total loan amount given in the year 2018.

TABLE 2. Table 2.2 Total Number of Loans and Loan Amount by Loan Purpose

Loan purpose	Total Number of Loans	Total Loan Amount(In Dollars)
Business	871	16,947,400
Car financing	967	9,890,700
Credit Card refinancing	15,847	234,597,175
Debt Consolidation	40,449	668, 513,800
Home buying	1,279	22,261,175
Home improvement	6,062	94,563,650
Major Purchase	2,367	36,149,275
Medical expenses	1,357	14,238,100
Moving and relocation	691	6,655,675
Vacation	685	4,774,100

TABLE 3. Table 2.3 Total Number of Loans and Loan Amount by Loan Purpose and Loan Status

Loan Purpose	Percentage of Loans		Total Loan Amount(In Dollars)	
	Default	Fully Paid	Default	Fully Paid
Business	42.1%	57.9%	7,648,950	9,298,450
Car financing	18.1%	81.9%	2,273,400	7,617,300
Credit Card refinancing	20.7%	79.3%	52,761,900	181,835,275
Debt consolidation	24.5%	75.5%	180,020,900	488,492,900
Home buying	24.0%	76.0%	6,237,625	16,023,550
Home improvement	21.0%	79.0%	23,319,175	71,244,475
Major purchase	26.7%	73.3%	12,740,750	23,480,525
Medical expenses	27.6%	72.4%	4,653,875	9,584,225
Moving and relocation	28.5%	71.5%	2,413,425	4,242,250
Vacation	21.6%	78.4%	1,666,825	3,107,275

In table 2.3, we have the percentage of loans and total loan amount by the loan purpose and loan status. Here we can observe that the loan taken for the purpose of Business has higher default percentage of 42.1% but the total loan amount is only 7 million dollars. The loans taken for the purpose of Debt

TABLE 4. Table 2.4 Loan Status by Home Ownership of the customer

Home Ownership	Loan Status	
	Default	Fully Paid
MORTGAGE	22%	78%
OWN	27%	73%
RENT	34%	66%

Consolidation has a default percentage of 24.5% which is accounting for 108 million.

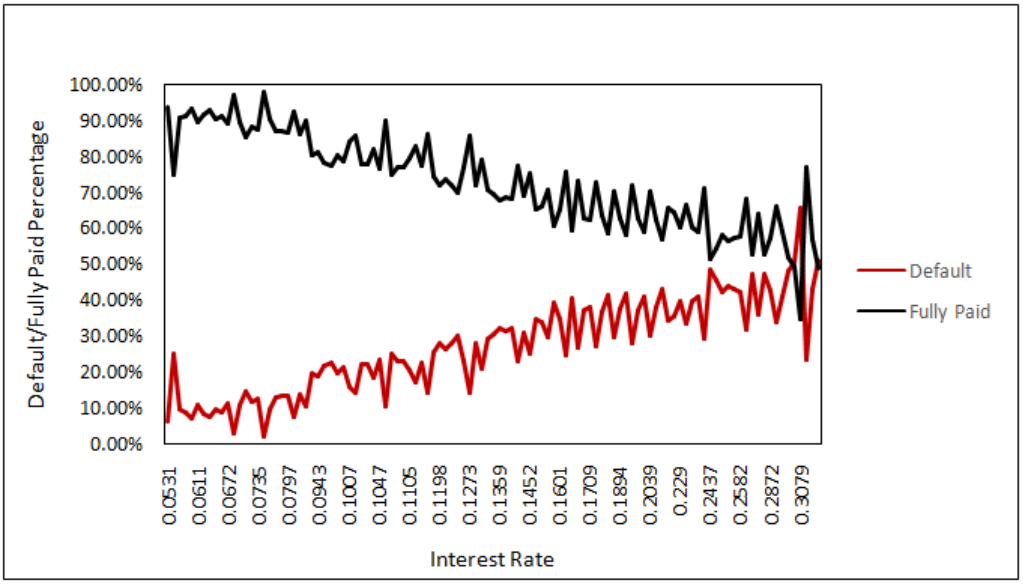

Figure 2.2 Graph Showing percentage of defaults at different interest rates

In the figure 2.2, we have defaults and fully paid percentages at different interest rates. From the figure we can say that as the interest rates are increasing the percentage of defaults are also increasing, so, adding this variable into the independent variable list will help us in predicting the defaults.

In table 2.4, we have loan status by the customers homeownership, here we have three home ownerships namely, OWN, means that the customer staying in his own house; RENT, means that the customer is staying in a rented house and MORTGAGE, means that the customers took some loan on his house. If we observe the table, the percentage of defaults in the RENT section is high. We can include this variable also as an independent variable in the model.

3. Conclusion

In this paper we discussed the linear regression and why we cannot use this method to predict a qualitative dichotomous dependent variable, also discussed about logistic function and how to transform the logistic function to build the logistic regression model. We developed a logistic regression model to predict the Loan Defaulters and discussed various accuracy measures to assess the accuracy of the developed model. To check the accuracy of the model, we used confusion matrix and obtained the accuracy measures as 88.83% accuracy, 91.07% Precision, 58.47% Recall and 71.22% f_1-score.

References

[1] J. Aldrich: *Fisher and Regression*, Journal of Statistical Science, **20** (2005), 401–417.
[2] A. Agresti: *An Introduction to Categorical Data Analysis,* Second Edition, A John Wiley & Sons, Inc. Publications, 2007.
[3] J. Gareth, D. Witten, T. Hastie, R. Tibshirani: An Introduction to Statistical Learning, Springer, 2013.
[4] G. Paliouras, V. Karkaletsis, D. C. Spyropoulos: Machine Learning and Its Applications, Springer, 2001.
[5] P. I . Good, J. W. Hardin: *Common Errors in Statistics (And How to Avoid Them),* Third Edition, John Wiley, 2009.
[6] D. W. Hosmer, S. Lemeshow: *Applied Logistic Regression,* Second Edition, Wiley, 2000.
[7] D. W. Hosmer: *A comparison of goodness-of-fit tests for the logistic regression model,*Journal of Statistics in Medicine, **16** (1997), 965–980.
[8] P. Buhlmann, T. Hothorn: *Boosting Algorithms: Regularization, Prediction and Model Fitting,* Journal of Statistical Science, **22**(4) (2007), 477–505.
[9] R. A. Fisher: *Statistical Methods for Research Workers,* 12th Edition, Oliver and Boyd Publications, 1954.

Department of Mathematics, School of Advanced Sciences, VIT, Vellore, Tamilnadu, India

Analytics Quotient Inc, India

Department of Mathematics, School of Advanced Sciences, VIT, Vellore, Tamilnadu, India
Email address: mokesh.g@vit.ac.in