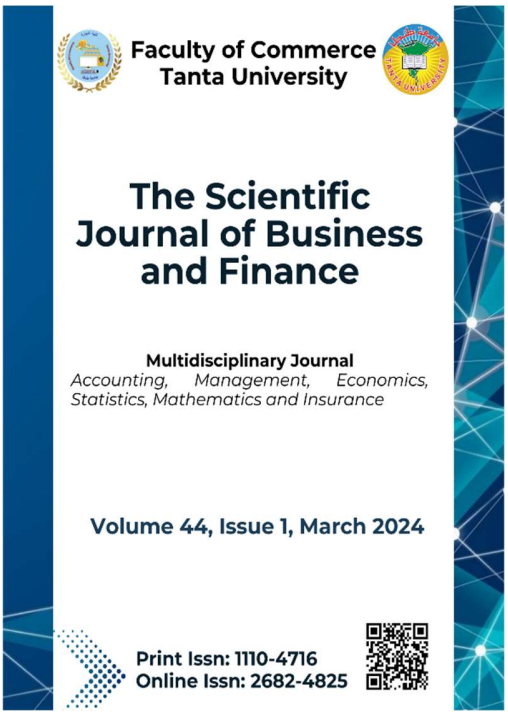

The Scientific Journal of Business and Finance

https://caf.journals.ekb.eg

Statistical Analysis for Credit Scoring based on Logistic regression model

Mona Emad El-Din Mohamed [a] , Mervat El-Gohary[b] , and Ahmed Amin El-Sheikh [c*]

[a] Faculty of commerce- Al-Azhar University (Girls Campus), Egypt

[b] Professor of Statistics Faculty of commerce- Al-Azhar University (Girls Campus), Egypt.

[c] Professor of Applied Statistics and Econometrics- Faculty of Graduate Studies for Statistical Research Cairo University. Egypt

Published online: **March 2024.**

To cite this article: Mohamed, Mona Emad El-Din., El-Gohary, Mervat and El-Sheikh,Ahmed Amin. Statistical Analysis for Credit Scoring based on Logistic regression model , The Scientific Journal of Business and Finance, 44 (1), 345-359. DOI: https://doi.org/10.21608/caf.2024.351751

*Corresponding author: Monaemad527@gmail.com

Statistical Analysis for Credit Scoring based on Logistic regression model

Mona Emad El-Din Mohamed [a]

[a] Faculty of commerce- Al-Azhar University (Girls Campus), Egypt

Mervat El-Gohary[b]

[b] Professor of Statistics Faculty of commerce- Al-Azhar University (Girls Campus), Egypt.

Ahmed Amin El-Sheikh [c]

[c] Professor of Applied Statistics and Econometrics- Faculty of Graduate Studies for Statistical Research Cairo University. Egypt

Article History

Received 3 December 2023, 25 ecember 2023, Available online March 2024

Abstract

A large number of classification techniques for credit scoring can be found in literature. Among These techniques statistical models which mainly comprise logistic regression techniques, linear discriminant analysis, k-nearest neighbor and classification tree. In the study, 614 random loan applications for clients made of a bank branch were examined. In this paper, Logistic Regression Analysis" was conducted to determine the problem and related factors and to predict the credibility according to these factors. In the model, customer age, education status, marital status, gender, profession, income, debt income ratio, credit card debt, other debts and multiplication product are taken as independent variables. As a result of the study, the bank branch will benefit from the statistical model in which it is created, to evaluate according to the customer characteristics in its portfolio, and to give more credit to branch customers.

Keywords: Credit Scoring, logistic regression (LR), loan prediction.

1. Introduction

Many credit scoring techniques have been used to build credit scorecards. Among them, logistic regression model is the most commonly used in the banking**.** There are quite complicated rules and constraints that can be imposed by the bank when the loan issued. Bank branches, which play a direct role in the credit, must accurately determine the customer's credit request to eliminate these difficulties and create an effective payment system according to the customer.

If people are not enough to obtain the financial means they need, they demand it in various forms. Credit scores are awarded on the basis of different techniques designed by individual lenders. However, irrespective of the varying nature of techniques used, credit scoring is invariably used to answer one key question - what is the probability of default within a fixed period, usually 12 months. Credit scoring can be divided into application scoring and behavior scoring, based on the information used when modeling. Application scoring uses only the information provided in application, while behavior scoring uses both the application information, and (past) behavior information.

A large number of classification techniques for credit scoring can be found in literature. Among These techniques statistical models which mainly comprise logistic regression techniques, linear discriminant analysis, k-nearest neighbor and classification tree.

In the study, 614 random loan applications for clients made of a bank branch were examined. In this paper, Logistic Regression Analysis" was conducted to determine the problem and related factors and to predict the credibility according to these factors. In the model, customer age, education status, marital status, gender, profession, income, debt income ratio, credit card debt, other debts and multiplication product are taken as independent variables. So, the credibility determined based on customer characteristics; A regression model was set up to answer the question of whether or not the loan should approved.

This paper is organized as follows. In Section (2) some literature reviews. In Section (3) Logistic regression model is introduced. The estimation of the parameter is introduced in Section (4). In section (5) some concluding remarks about the results are illustrated.

2. Literature review

Baesens et al., (2003) they studied the performance of various state of the art classification algorithms. They concluded that the simple classifiers such as LR and discriminant analysis perform very well for credit scoring.

Zekic-Susac, et al., (2004) compared the models for small business credit scoring developed by logistic regression, neural networks, and decision trees on a Croatian bank dataset. The most successful neural network model was obtained by the probabilistic algorithm. The best model extracted the most important features for small business credit scoring from the observed data.

Bensic, et al., (2005) purposed extract important features for credit scoring in small-business lending on a dataset with specific transitional economic conditions using a relatively small dataset. The best model extracts a set of important features for small-business credit scoring for the observed sample, emphasizing credit programmed characteristics, as well as entrepreneur's personal and business characteristics as the most important ones.

Dong, et al., (2010) proposed a logistic regression model with random coefficients for building credit scorecards. He concluded that the proposed model needs much more time to estimate parameters.

Tirki et al., (2016) builds a non-parametric credit scoring model based on the Multi-Layer perceptron approach (MLP) and benchmarks its performance against Logistic Regression (LR) techniques.

Khemais, et al., (2016) developed models for foreseeing default risk of small and medium enterprises (SMEs) for one Tunisian commercial bank using two different methodologies (logistic regression and discriminant analysis). The empirical results that we found support the idea that these two scoring techniques have a statistically significant power in predicting default risk of enterprises.

Unver et al., (2018) introduced the" LR Model" which was created to predict creditworthiness according to the identified fugitives. As a result of the study, the bank branch will benefit from the statistical model in which it was created, to evaluate according to the customer characteristics in its portfolio, and to give more credit to branch customers.

Silva et al., (2020), developed a logistic regression model to predict the default credit risk they found that clients in the lowest income tax echelon have more propensity to default. The model was validated in terms of goodness-of-fit, residuals analysis, and lack of influential points.

Febrianti et al., (2021) introduced the maximum likelihood parameter estimation method with Newton Raphson iteration in general to estimate the parameters of the logistic regression model. The modification of the score function can quickly yield values of parameter estimates, especially when the sample sizes are larger.

Gouvêa et al., (2021) studied a sample set of applicants from a large Brazilian financial institution. Results obtained by the logistic regression and neural network models are good and very similar, although the former is slightly better. The genetic algorithm model is also efficient, but somewhat inferior.

3. Logistic Regression Model

The logistic model will now be used where more than one independent variable is available and this is called multivariate logistic regression. Consider a collection of k independent variables denoted by the vector$X = (x_1, x_2, \dots, x_k)$. Denote the conditional probability that the event is observed by: $p(y = 1/X) = \pi(x)$.

The logit of the multivariate logistic regression (Dong, et al. (2010)) is then given by the equation:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_o + \beta_1 x_1 + \cdots . + \beta_k x_k \quad (1)$$

This means the logistic regression is given by

$$\pi(X) = \frac{e^{g(x)}}{1+e^{g(x)}} \quad (2)$$

Assume that a sample of n independent observations$(x_i, y_i), i = 1, 2, \dots, n$. The estimates of the following vector need to be obtained:

$$\beta' = (\beta_o, \beta_1, \dots, \beta_k)$$

The method of estimation in the multivariate case is also maximum likelihood. The likelihood function will now be:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \quad (3)$$

Where $\pi(x_i)$ defined as:

$$\pi(x_i) = \frac{e^{g(x_i)}}{1+e^{g(x_i)}} \quad (4)$$

The $p + 1$ likelihood equations will be obtained by di§erentiating the log likelihood function with respect to the e $p + 1$ coe¢ cients. As with the univariate case, there is no easy solution for these equations and solving them requires special software packages and numerical methods.

Let $\hat{\beta}$denote the solution to these equations. In the previous chapter, the standard error of the estimate was used. It will now be considered in more detail. The method of estimating the variances and covariances of the estimated coefficients follows from the theory that estimators are obtained from the matrix of second partial derivatives of the log likelihood function. (Febrianti et al., (2021).

Let the $(p+1) \times (p+1)$ matrix containing the negative of these partial derivatives be denoted by $I(\hat{\beta})$ this matrix is called the observed information matrix. The variances and covariance are obtained from the inverse of the matrix, which is denoted by

$$Var(\hat{\beta}) = I^{-1}(\beta)$$

The estimated standard errors of the estimated coefficients will mostly be used, which are:

$$SE(\hat{\beta}_j) = \left(var(\hat{\beta}_j)\right)^{1/2}, \qquad j = 1, 2, \dots, k. \tag{5}$$

A useful formulation of the information matrix is:

$$\hat{I}(\hat{\beta}) = X'VX. \tag{6}$$

Where

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdot\cdot & \cdot\cdot \\ 1 & x_{n1} & x_{n2} & \cdot\cdot & x_{nk} \end{pmatrix} \tag{7}$$

$$V = \begin{pmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & 0 & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & 0 & 0 \\ \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ 0 & 0 & 0 & \hat{\pi}_k(1-\hat{\pi}_k) \end{pmatrix}. \tag{8}$$

Once the multivariate logistic regression model has been fitted, the model assessment begins. The first step is to assess the overall significance of the p independent variables in the model, using the likelihood ratio as in the univariate case. The likelihood of the fitted model is compared to the likelihood of a constant only model.

To test there is no difference between the fitted and full (/intercept only) model we use Wald test.

- **Wald test:**

To assess the significance of the logistic regression coefficients, the Wald statistic is used. (Afifi et al., 2004) and (Bewick et al., 2005). The Wald test is obtained from the following matrix calculation

$$W = \hat{\beta}'(X'VX)^{-1}\hat{\beta}$$

Which will be distributed as chi-square with $p + 1$ degrees-of-freedom under the hypothesis that each of the $p + 1$ coefficients are equal to zero. Tests for just the p slope coefficients are obtained by eliminating $\hat{\beta}_o$ from $\hat{\beta}$ and the relevant row and column from $(X'VX)$. Since evaluation of this

test requires the capability to perform vector-matrix operations and to obtain $\hat{\beta}$, there is no gain over the likelihood ratio test of the significance of the model.

3. 1 Data set and determination of variables

This subsection is introduced to describe the data set; the data set was taken from the online website (http://www.Kaggle.com). Using R program table (1) introduced the case processing summary as follows:

Table 1. Case processing summary

		N	Marginal Percentage
Dependents	0	274	57.1%
	1.00	80	16.7%
	2.00	85	17.7%
	3.00	41	8.5%
Gender	Male	394	82.1%
	Female	86	17.9%
Education	Not Graduate	97	20.2%
	Graduate	383	79.8%
Married	No	169	35.2%
	Yes	311	64.8%
Self Employed	No	414	86.3%
	Yes	66	13.8%
Property Area	Urban	150	31.3%
	Rural	139	29.0%
	Semi Urban	191	39.8%
Loan Status	Y	332	69.2%
	N	148	30.8%
Credit History	0.00	70	14.6%
	1.00	410	85.4%
Valid		480	100.0%
Missing		134	
Total		614	
Subpopulation		480[a]	

a. The dependent variable has only one value observed in 480 (100.0%) subpopulations.

The analysis summary offers a detailed glimpse into the dataset, focusing on variables that are crucial for determining credit suitability. Here are the key observations:

1. **Dependents**: The distribution of the number of dependents is as follows:
   - No dependents: 57.1%
   - 1 dependent: 16.7%
   - 2 dependents: 17.7%
   - 3 dependents: 8.5%

This breakdown highlights that a significant portion of individuals have no dependents, with a gradual decline as the number of dependents increases. The number of dependents could potentially indicate the financial responsibilities of an individual, impacting their ability to repay loans.

2. **Gender**: The gender distribution displays a noticeable imbalance:
   - Male: 82.1%
   - Female: 17.9%

The dataset is predominantly male, which might raise questions about gender-based financial disparities and their influence on credit decisions.

3. **Education**: The education distribution shows:
   - Graduate: 79.8%
   - Not Graduate: 20.2%

The majority of individuals are graduates, possibly implying that education level could play a role in securing loans.

4. **Married**: The marital status distribution indicates:
   - Married: 64.8%
   - Not Married: 35.2%

This distribution might reflect differing financial responsibilities and could be a factor in loan approval considerations.

5. **Self Employed**: Most individuals are not self-employed (86.3%), which could be significant in evaluating their income stability and its adequacy for loan repayment.

6. **Property Area**: The dataset is distributed across three property areas:
   - Urban: 31.3%
   - Rural: 29.0%
   - Semi-Urban: 39.8%

The diverse distribution across property areas might capture local economic conditions that could affect repayment capabilities.

7. **Loan Status**: The loan approval status showcases:
   - Approved (Y): 69.2%
   - Not Approved (N): 30.8%

This imbalance suggests that the dataset might be skewed towards approved loans. Addressing this imbalance during modeling is crucial.

8. **Credit History**: Credit history is a pivotal variable with:
   - No credit history (0.00): 14.6%
   - Positive credit history (1.00): 85.4%

A strong credit history is often indicative of a person's creditworthiness.

9. **Missing Data**: There are 134 missing values across the dataset, which need to be handled appropriately before analysis. The method chosen for imputation can influence results.

3.2Model Fitting

This subsection is concerned with clarifying the suitability of the data using some goodness of fit criteria; the results have been presented in table (2).

Table 2. Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1095.953	1108.474	1089.953			
Final	1022.144	1184.922	944.144	145.808	36	0.000

Table 3. Pseudo R-Square

Cox and Snell	0.262
Nagelkerke	0.292
McFadden	0.134

The "Model Fitting Information" section provides important details about the fitting of the logistic regression model and the "Pseudo R-Square" values provide information about the goodness of fit of the logistic regression model in relation to the null model (a model with no predictors). They indicate the proportion of variance explained by the model compared to the total variance that would be explained by a perfect model. Here's an explanation of the key information presented:

Comments on table (2) and (3)

- The "Intercept Only" model is a baseline model with only an intercept term, essentially a simple model with no predictors.
- The "Final" model is the logistic regression model that we've developed using the predictors in our dataset.
- Comparing the "Final" model to the "Intercept Only" model using the likelihood ratio test shows that the "Final" model significantly improves the fit.
- The chi-square value of 145.808 with 36 degrees of freedom is highly significant (p-value < 0.001), indicating that the predictors in the "Final" model collectively contribute to a better fit compared to the intercept-only model.
- The AIC and BIC values of the "Final" model (1022.144 and 1184.922, respectively) are lower than those of the "Intercept Only" model, suggesting a better fit and less complexity in the "Final" model.
- Pseudo R-Square values provide an indication of how well the logistic regression model fits the data compared to a null model.
- While these values give a sense of the model's fit, they should be interpreted cautiously. Pseudo R-Square measures can vary widely and may not have a direct interpretation as in linear regression.
- It's important to consider other aspects of model fit, such as likelihood ratio tests, AIC, and BIC, to assess the overall quality of the logistic regression model.

4. Parameter Estimation

In this section, the estimation to the real data set using the logistic regression model is introduced. The results are illustrated in table (4) as follows:

Table 4. Parameter Estimates

Dependents[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
0	Intercept	2.921	1.591	3.370	1	0.066			
	Applicant Income	.000	.000	.957	1	0.328	1.000	1.000	1.000
	Co-applicant Income	.000	.000	2.433	1	0.119	1.000	1.000	1.000
	Loan Amount	-.005	.003	3.248	1	0.072	.995	.990	1.000
	Loan Amount Term	.004	.003	1.863	1	0.172	1.004	.998	1.009
	[Gender=1.00]	-1.805	1.056	2.920	1	0.087	.164	.021	1.304
	[Gender=2.00]	0[b]	.	.	0	.	.	.	.
	[Education=1.00]	-.386	.440	.768	1	0.381	.680	.287	1.612
	[Education=2.00]	0[b]	.	.	0	.	.	.	.
	[Married=.00]	2.046	.562	13.224	1	0.000	7.733	2.568	23.289
	[Married=1.00]	0[b]	.	.	0	.	.	.	.
	[Self Employed=.00]	-.106	.552	.037	1	0.848	.899	.305	2.651
	[Self Employed=1.00]	0[b]	.	.	0	.	.	.	.
	[Property Area=1.00]	.365	.486	.565	1	0.452	1.441	.556	3.735
	[Property Area=2.00]	-.278	.418	.443	1	0.506	.757	.334	1.718
	[Property Area=3.00]	0[b]	.	.	0	.	.	.	.
	[Loan Status=1.00]	-.258	.499	.268	1	0.605	.772	.291	2.053
	[Loan Status=2.00]	0[b]	.	.	0	.	.	.	.
	[Credit History=.00]	-.666	.578	1.329	1	0.249	.514	.165	1.595
	[Credit History=1.00]	0[b]	.	.	0	.	.	.	.
1.00	Intercept	4.072	1.652	6.074	1	0.014			
	Applicant Income	.000	.000	.049	1	0.825	1.000	1.000	1.000
	Co-applicant Income	.000	.000	.550	1	0.458	1.000	1.000	1.000
	Loan Amount	-.001	.003	.230	1	0.631	.999	.994	1.004
	Loan Amount Term	-.001	.003	.034	1	0.854	.999	.994	1.005
	[Gender=1.00]	-2.094	1.080	3.757	1	0.053	.123	.015	1.024
	[Gender=2.00]	0[b]	.	.	0	.	.	.	.
	[Education=1.00]	-.345	.501	.475	1	0.491	.708	.265	1.890
	[Education=2.00]	0[b]	.	.	0	.	.	.	.
	[Married=.00]	.395	.634	.389	1	0.533	1.485	.429	5.139
	[Married=1.00]	0[b]	.	.	0	.	.	.	.
	[Self Employed=.00]	-.699	.576	1.472	1	0.225	.497	.161	1.538
	[Self Employed=1.00]	0[b]	.	.	0	.	.	.	.
	[Property Area=1.00]	.825	.516	2.554	1	0.110	2.281	.830	6.272
	[Property Area=2.00]	-.797	.494	2.600	1	0.107	.451	.171	1.187
	[Property Area=3.00]	0[b]	.	.	0	.	.	.	.
	[Loan Status=1.00]	-.649	.538	1.456	1	0.228	.523	.182	1.500
	[Loan Status=2.00]	0[b]	.	.	0	.	.	.	.
	[Credit History=.00]	-.861	.638	1.823	1	0.177	.423	.121	1.475
	[Credit History=1.00]	0[b]	.	.	0	.	.	.	.
2.00	Intercept	1.284	1.758	.534	1	0.465			
	Applicant Income	.000	.000	1.038	1	0.308	1.000	1.000	1.000

	Co-applicant Income	.000	.000	.693	1	0.405	1.000	1.000	1.000
	Loan Amount	-.001	.003	.118	1	0.731	.999	.993	1.005
	Loan Amount Term	.003	.003	1.004	1	0.316	1.003	.997	1.009
	[Gender=1.00]	-.773	1.160	.444	1	0.505	.462	.048	4.484
	[Gender=2.00]	0[b]	.	.	0	.	.	.	.
	[Education=1.00]	-.219	.483	.205	1	0.651	.804	.312	2.072
	[Education=2.00]	0[b]	.	.	0	.	.	.	.
	[Married=.00]	-.193	.680	.080	1	0.777	.825	.218	3.125
	[Married=1.00]	0[b]	.	.	0	.	.	.	.
	[Self Employed=.00]	-.570	.580	.964	1	0.326	.566	.181	1.764
	[Self Employed=1.00]	0[b]	.	.	0	.	.	.	.
	[Property Area=1.00]	.857	.515	2.767	1	0.096	2.356	.858	6.468
	[Property Area=2.00]	-.248	.468	.281	1	0.596	.781	.312	1.952
	[Property Area=3.00]	0[b]	.	.	0	.	.	.	.
	[Loan Status=1.00]	.003	.554	.000	1	0.996	1.003	.339	2.969
	[Loan Status=2.00]	0[b]	.	.	0	.	.	.	.
	[Credit History=.00]	-.627	.658	.907	1	0.341	.534	.147	1.941
	[Credit History=1.00]	0[b]	.	.	0	.	.	.	.
a. The reference category is: 3.00.									
b. This parameter is set to zero because it is redundant.									

- **According to the results:**

The parameter estimates table we've provided is a crucial part of the logistic regression analysis. It provides insights into how each predictor variable influences the log-odds of the outcome variable while accounting for other variables in the model. Below is an interpretation of the table based on the given information:

1. Intercept: The intercept represents the log-odds of the outcome when all predictor variables are at their reference levels (categorical variables) or zero (continuous variables). For the reference category of Dependents (3.00), the intercept is 2.921.

2. Applicant Income, Co-applicant Income, Loan Amount, Loan Amount Term: These continuous variables have coefficients close to zero. This suggests that small changes in these variables have minimal impact on the log-odds of the outcome.

3. Gender: The coefficients for Gender indicate how different genders affect the log-odds compared to the reference category (Dependents=3.00). The coefficient for Gender=1.00 is -1.805, suggesting that being male (Gender=1.00) decreases the log-odds by 1.805 compared to the reference. The coefficient for Gender=2.00 is not provided (set to 0), indicating that it's redundant or collinear with other variables.

4. Education: Similar to Gender, Education coefficients show the impact of different education levels compared to the reference. However, both coefficients for Education levels (1.00 and 2.00)

are not statistically significant (p > 0.05), implying that education level might not be a strong predictor in this model.

5. Married: Being married (Married=1.00) increases the log-odds by 2.046 compared to not being married (Married=3.00).

6. Self Employed: Self-employment (Self Employed=1.00) doesn't significantly impact the log-odds. The coefficient is -0.106, and the variable is not statistically significant (p > 0.05).

7. Property Area: The coefficients for different property areas (1.00 and 2.00) indicate how they affect the log-odds compared to the reference category (Property Area=3.00). However, the variable is not statistically significant (p > 0.05) as the p-values are greater than the common significance level of 0.05.

8. Loan Status: The coefficients for Loan Status levels (1.00 and 2.00) are not statistically significant, suggesting that the variable might not be a strong predictor in this model.

9. Credit History: Having a credit history (Credit History=1.00) decreases the log-odds by -0.666 compared to not having a credit history (Credit History=3.00). However, the variable is not statistically significant (p > 0.05).

Keep in mind that the significance of the coefficients is determined by their p-values. A p-value less than the chosen significance level (e.g., 0.05) indicates that the predictor variable is statistically significant in predicting the outcome. If a coefficient is not statistically significant, it suggests that changes in that variable are not associated with changes in the log-odds of the outcome, at least in the context of the current model.

Keep in mind that some variables, such as Gender=2.00, Education=2.00, Property Area=3.00, and so on, have coefficients set to zero (indicated by "0b"). This means they are not included in the equation due to redundancy or collinearity with other variables.

Please note that the equation uses log-odds, and to obtain predicted probabilities, you'll need to apply the logistic function (inverse logit) to the right-hand side of the equation:

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}}$$

This equation allows you to estimate the probability of credit suitability based on the predictor variables in our logistic regression model.

The natural logarithm of the odds ratio of the probability of credit suitability (p) to the probability of non-suitability ($1-p$) is equal to a linear combination of the coefficients and predictor variables.

Based on the parameter estimates we've provided, the logistic regression equation for our model could be as follows:

$$\log\frac{p}{1-p} = 2.921 + 0.000\ \text{ApplicantIncome} + 0.000\ \text{CoapplicantIncome} - 0.005\ \text{LoanAmount} + 0.004\ \text{LoanAmountTerm} + -1.805\ \text{Gender} + 2.046\ \text{Married} + -0.106\ \text{SelfEmployed} + 0.365\ \text{PropertyArea} + -0.258\ \text{LoanStatus} - 0.666\ \text{CreditHistory}$$

5. Concluding Remarks

The assessment of credit risk during the decision to grant credit remains the main concern of microfinance institutions that have set considerable effort by trying to determine the most effective ways to make it a task easier to manage, requiring a minimum of time. Credit scoring using a nonparametric statistical technique with the microfinance industry is a relatively recent application. In this paper we propose a logistic regression model with random coefficients for building credit scorecards. The empirical results indicate the proposed model can improve prediction accuracy of the logistic regression with fixed coefficients without sacrificing its desirable features. The parameters of the model are also estimated.

References

Afifi, A., Clark, V. A., & May, S., (2004). Computer- Aided Multivariate Analysis. Fourth Edition, Champman and Hall/CRC.

Baesens, Bart, et al. ”Benchmarking state-of-the-art classification algorithms for credit scoring.” Journal of the operational research society 54.6 (2003): 627-635.

Bensic, Mirta, Natasa Sarlija, and Marijana Zekic-Susac. ”Modelling small-business credit scoring by using logistic regression, neural networks and decision trees.” Intelligent Systems in Accounting, Finance and Management 13.3 (2005): 133-150.

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. Critical Care (London, England), 9(1), 112-118. http://dx.doi.org/ 10.1186/cc3045.

Dong, Gang, Kin Keung Lai, and Jerome Yen. ”Credit scorecard based on logistic regression with random coefficients.” Procedia Computer Science 1.1 (2010): 2463-2468

Eliana Costa e Silva, Isabel Cristina Lopes, Aldina Correia & Susana Faria (2020) A logistic regression model for consumer default risk, Journal of Applied Statistics, 47:13-15, 2879-2894, DOI: 10.1080/02664763.2020.1759030

Febrianti, R., Widyaningsih, Y., & Soemartojo, S. (2021). The parameter estimation of logistic regression with maximum likelihood method and score function modification. In *Journal of Physics: Conference Series* (Vol. 1725, No. 1, p. 012014). IOP Publishing.

Gonçalves, E. B., & Gouvêa, M. A. (2021). Credit Risk Analysis Applying Logistic Regression, Neural Networks and Genetic Algorithms Models. *International Journal of Advanced Engineering Research and Science*, *8*(9), 198-209.

Khemais, Zaghdoudi, Djebali Nesrine, and Mezni Mohamed. ”Credit Scoring and Default Risk Prediction: A Comparative Study between Discriminant Analysis & Logistic Regression.” International Journal of Economics and Finance 8.4 (2016): 39

Triki, I. (2016). Credit scoring models for a Tunisian microfinance institution: comparison between artificial neural network and logistic regression. *Review of Economics & Finance*, *6*, 61-78.

Unver, M., Sahin, B., & Ersoz, F. (2018). An application of logistics regression model to determining the credit suitability and impacting factors in a special bank branch. Communication in Mathematical Modeling and Applications, 3(1), 1-12.

Zekic-Susac, Marijana, Natasa Sarlija, and Mirta Bensic. ”Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models.” Information Technology Interfaces, 2004. 26th International Conference on. IEEE, 2004.