Credit Risk Analysis Using Logistic Regression Modeling.

Zankhana P. Atodaria
GIDC Rajju Shroff ROFEL Institute of Management Studies Vapi

CREDIT RISK ANALYSIS USING LOGISTIC REGRESSION MODELING

Zankhana Atodaria[1], Shubhangi Pentar[2]

[1]Assistant Professor, GIDC Rajju Shroff Rofel Institute of Management Studies (MBA Programme)

[2]GIDC Rajju Shroff Rofel Institute of Management Studies (MBA Programme), Vapi

Zankhanaatodaria@gmail.com

Abstract

Loan officer at any bank would be interested in detecting the factors which can identify people who are likely to default on loans, consequently good and bad credit risks. Moreover, he will also be interested in designing model which can predict chances of default with reasonable accuracy. This Study attempted to provide a comprehensive analysis of credit risk using logistic regression model. The logistic regression model is run among categorical dependent variable and a set of independent variables. For same, home loan data & information of 487 clients of Bandhan bank(Vapi) have been chosen as sample based on systematic random sampling. The cut off value for prediction has been determined by ROC curve. Classification plots, Hosmer-Lemeshow goodness-of-fit & Pseudo R-Squared Statistics has been analysed. Analysis of study concludes occupation, residential stability & marriageable son/daughter as significant variables in predicting chances of default.

Keywords: Credit Risk Analysis, Logistic Regression Model, ROC Curve, Chances of default.

Introduction

Collecting deposits from customers & channelizing it to through loans to the clients who require it for various purposes is one of the most important functions of commercial banks. Banks as one of the financial institutions face credit risk defined as the probability of a borrower to repay loans and determining the likelihood of the person to fail a payment, in which case the bank incurs a loss. Credit Risk Analysis refers to assessing the possibility of the borrower’s repayment failure and the loss caused to the financer when the borrower does not repay for any reason in the contractual loan obligations. Interest for credit-risk assumption forms the earnings and rewards from such debt-obligations and risk. The cash flow of the financer is impacted when the interest accrued and principal amounts are not paid.

Further, the cost of collections also increases. Though, there is a grey area in guessing who and when will default on borrowings, it is the process of intelligent inquiry of severity of loss of the borrowings and its recovery. The credit risk modelling is very important for banks because it helps them to improve their business and at the same time serve customers better.

Logistic Regression Modeling also known as binary logistic regression studies the association between a categorical dichotomous dependent variable and a set of independent variables which may be categorical or continuous. Many researchers compared different methods in developing credit risk analysis models. logistic regression is one of the most frequently used method. (Sarlija et al.,2009 & Ali Al-Aradi ,2014). Logistic regression competes with discriminant analysis as a method for analysing categorical variables. Many statisticians feel that logistic regression is more versatile and better suited for Modeling most situations than the discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

Review of Literature

Credit default risk carry the greatest hazard to commercial banks in providing and maintaining the financial stability. Consequently, methods of assessment, management as well as prevention of loan risks have to be the priority in development of banking system. Under the conditions of intense competition in the market of retail loan issuing, the designed model allows the bank loan analyst to take justified and grounded self-decisions not only on loan service for customers but also management of the loan portfolio. Though, with the passage of time, any statistical model becomes inaccurate, owing to business cycles, changes of the customer data base of bank, structural shifts in economy, inflation etc. Thus, it requires periodical adjustment to assure a continuous functioning of the credit scoring model (Yurynets et al.)

Credit risk Analysis can be carried out through numerous methods. The logistic regression and neural network models both are alike & worthy. Though the prior one is marginally better. Moreover, the genetic algorithm model is somewhat inferior but efficient. (Gouvea & Goncalves ,2007). logistic regression and multicriteria decision making have been combined to create a proficient strategy to achieve high performance. Logistics regression is used to find probability of default whereas multicriteria decision is used to classify firms for credit scoring based on predefined criteria (Sarlija et al.,2009). Artificial neural network model is better in identifying bad customers in commercial bank than Logistics Regression Model. Increase in interest rate & time delay in repayment leads to higher credit risk. It is more in case of industry sector customers. It reduces with longer history of customer relationship with bank & time period of repayment of loan. (Karimi ,2014). Predictive ability is high yet not similar between logistic regression and multiple discriminant analysis. Return on assets is statistically significant having very high regression coefficients (Memic,2015). Chen et al. (2019) proposed credit risk assessment model through which investor of peer to peer lending can measure the risk with better accuracy & can also predict the amount of profit from a loan.

Awotwi (2011) tried to design a model to predict the probability of default of an applicant. Probability of default of unmarried applicants are 1.24 times higher. Moreover, Lower income earners are more likely to default. Current employment tenure is positively related to loan repayment chances. Probability of default decreases by .998 with marginal increase in the number of months in current employment. Ali Al-Aradi (2014) find that larger checking account holders are less likely to default. However, data analysis also reflected customers with no checking accounts to be more creditworthy. Creditworthiness of single males are 1.636 times higher than divorced/married females. The credit score of customers who intended to buy a used car is 5.497 times higher than the ones whose purpose was education. Probability of someone to default is expected to be at or below 0.25 to receive a credit loan. 31 to 40 % of customers make greatest difference in actual good or bad credit observations (Priestley & Frankel,2016). The results of the study of a microfinance institution based in Accra Ghana conducted reveal that six characteristics were statistically significant in the prediction of loan repayment default (86.67% - predicted default rate). These factors are type of collateral or Security, Marital Status, assessment, duration, dependents & Loan Type. Factors that determine borrowers’ creditworthiness are loan to value ratio, credit history, payment to income ratio & borrower’s type, with average prediction accuracy of 93.4% within the sample (Ergeshidze,2017).

Research Methodology

Objectives:

- To design a loan default model.
- To predict the probability of loan default.
- To identify demographic & behavioural factors causing default in loan payments.

Binomial logistic Regression is used to model the loan default. Secondary data of loans is collected from Bandhan Bank, Vapi Branch. Systematic random sampling is used to select sample of customers in which every 7th element has been chosen out of 3409 account details. Ultimately sample of 487 customers is taken. SPSS software has been used for analysis.

Data Analysis

Iteration		**-2 Log likelihood**	**Coefficients**
			Constant
Step 0	1	336.081	-1.598
	2	318.488	-2.081
	3	317.948	-2.186
	4	317.947	-2.190
	5	317.947	-2.190

Iteration History [a, b, c]

Table 1: Building Block 0: Null Model

a. Constant is included in the model.
b. Initial -2 Log Likelihood: 317.947.
c. Estimation terminated at Iteration no.5 because parameter estimate changed by less than 0.001.

Table 1 (block zero) shows the very basic log likelihood value which should decrease in the subsequent blocks reflecting improvement in model & model predictive accuracy.

Step	**-2Log likelihood**	**Cox & Snell R Square**	**Nagelkerke R Square**
a. 1	283.770[a]	0.068	0.141

Table 2: Model Summary

a. Estimation terminated at iteration number 20 because maximum iterations have been reached. Final Solution cannot be found.

Log likelihood value could be seen decreased in table 2 from 317.947 to 283.770 which shows improvement in model. Nagelkerke R^2 is a measure which exhibit the explained variance which is 14.1%.

Step 1	**Chi-square**	**Df**	**Sig.**
Step	34.177	14	0.002
Block	34.177	14	0.002

Model	34.177	14	0.002

Table 3: Omnibus Tests of Model Coefficients

Omnibus test covered in table 3 test whether the explained variance is significantly more than the unexplained variance. Higher χ^2 & lower p value signifies the same.

Step	**Chi-square**	**df**	**Sig.**
1	10.205	8	0.251

Table 4: Hosmer and Lemeshow Test

Lower χ^2 & higher p value of Hosmer and Lemeshow Test covered in table 4 signifies the predictive accuracy in addition to proving data fit for model.

Observed		**Predicted**		
		Default Status		**Percentage Correct**
		No	Yes	
Step 1	No	438	0	100.0
Default Status	Yes	49	0	.0
Overall Percentage				89.9

a. The cut off value is .500

Table 5: Classification Table [a]

Initial analysis was run with 0.5 as cut off value to create a classification table depicting 89.9% overall accuracy. Though overall percentage accuracy is high, model failed to predict default cases completely. Consequently, better cut off value has been found through ROC curve analysis.

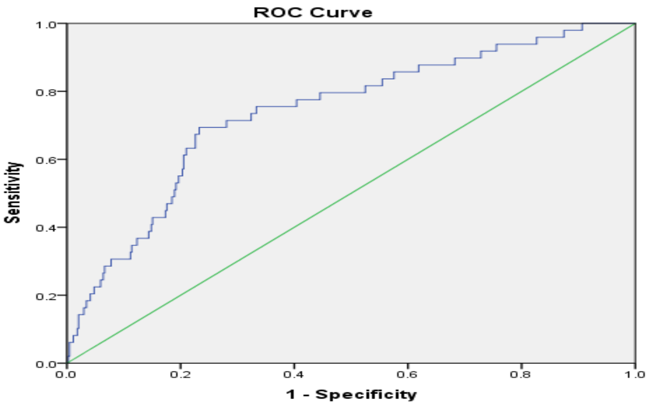

Chart 1: ROC curve

Chart 1 covers two parameters i.e. Sensitivity & Specificity. Sensitivity score reflects the ability of a model to correctly classify defaulters whereas specificity score denoted ability of model to correctly classify non-defaulters. Researcher has to trade-off between these two parameters to decide classification cut-off. The farthest coordinate on chart 1 suggest cut off of 0.25 for optimum classification. After running the model once again based on 0.25 cut off, the following classification table resulted.

Observed		**Predicted**		
		Default Status		**Percentage Correct**
		No	Yes	
Step 1	No	425	13	97.0
Default Status	Yes	39	10	20.4
Overall Percentage				89.3

a. The cut off value is .250

Table 6: Classification Table [a]

It is evident from table 6 that overall accuracy has not reduced much. Instead, sensitivity percentage has been improved from 0 to 20.4 %.

Step 1 [a]	B	S.E.	Wald	Df	Sig.	Exp(B)
gender(1)	0.164	0.652	0.064	1	0.801	1.179
Age	-0.009	0.022	0.159	1	0.69	0.991
Income	0	0	0.45	1	0.502	1
Education			0.032	2	0.984	
Education(1)	18.677	9733.69	0	1	0.998	129220153
Education(2)	18.741	9733.69	0	1	0.998	137759991
OS	0.025	0.025	1.016	1	0.313	1.026
Occupation(1)	1.619	0.415	15.243	1	0	5.047
Re.Status	-0.13	0.068	3.66	1	0.056	0.878
No of Children	-0.397	0.298	1.781	1	0.182	0.672
MSD(1)	1.126	0.65	3.004	1	0.083	3.084
Dependency	0.249	0.193	1.653	1	0.199	1.282
Loan Amt	0	0	1.276	1	0.259	1
OS Amt	0	0	1.012	1	0.314	1
Debt-Income R	1.629	2.503	0.424	1	0.515	5.101
Constant	-22.943	9733.69	0	1	0.998	0

a. Variable(s) entered on step 1: gender ,Age, Income, Education, OS, Occupation, Resta, No of Child ,MCD ,T Dependency ,loan Amount ,OS Amount, Debt Income R.

Table 7: Variables in the Equation

It is clear from table 7 that Occupation, residential stability & marriageable son/daughter are the independent variables which have a significant effect on repayment of loan at the

Bandhan bank. Out of total 50 defaulters, 41 are doing business of which majority are male. Moreover, 29 clients have residential stability less than 5 years (Table 8, Table 9).

Occupation	No. of Customer
Business	41
Job	09
Total	50

Table 8:

Residential Stability	No of customer
1-4	29
5-9	17
10-14	03
More than 14	01
Total	50

Table 9:

Conclusion

Credit risk models are beneficial to evaluate the risk of default in repayment of consumer loans. Greater precision of a prediction model will provide financial returns to the institution. The main objective of this study is to model credit default using logistics regression technique. The Hosmer and Lemeshow test support the model. The explained variance of model is 14.1%. Gender, age, income, education, occupation, residential stability, Number of children, marriageable son/daughter, total dependency, loan amount, outstanding amount & debt income ratio pertaining to home loan were used to model credit default. Among these, three characteristics are found to be significant in predicting default probability, these are occupation, residential stability & marriageable son/daughter. Out of 50 defaulters, 41 runs business & 29 have 1-4 years' stability. Predictive accuracy has been increased to 20.4% after applying 0.25 cut off instead of 0.50.

Bibliography

1. Agbemava, E., Nyarko, I. K., Adade, T. C., & Bediako, A. K. (2016). Logistic Regression Analysis Of Predictors Of Loan Defaults By Customers Of Non-Traditional Banks In Ghana. *European Scientific Journal January*, *12*(1), 175–189. https://doi.org/10.19044/esj.2016.v12n1p175
2. Al-aradi, A. (2014). Credit Scoring via Logistic Regression. In *Methods of Applied Statistics* (pp. 1–12).
3. Awotwi, E. K. (2011). *Estimation of The Probability of Default of Consumer Credit in Ghana : Case Study of an International Bank.*
4. Chen, S., Charkaborty, G., Li, L., & Lin, C. (2019). Credit Risk Assessment Using Regression Model on P2P Lending. *International Journal of Applied Science and Engineering*, *16*(2), 149–157. https://doi.org/10.6703/IJASE.201909
5. Ergeshidze, A. (2017). Credit Risk Model : Assessing Default Probability of Mortgage Loan Borrower. *RSEP International Conferences on Social Issues and Economic Studies*, 5–7. https://doi.org/10.19275/RSEPCONFERENCES055
6. Gouvea, M. A. (2007). Credit Risk Analysis Applying Logistic Regression, Neural Networks and Genetic Algorithms Models. *POMS 18th Annual Conference*, 1–49.
7. Jerić, Silvija & vSarlija, Natavsa & vSori'c, Kristina & Rosenzweig, Vivsnja. (2009). Logistic Regression and Multicriteria Decision Making in Credit Scoring, Proceedings of the 10th International Symposium on Operational Research SOR '09.
8. Karimi, A. (2014). Credit Risk Modeling for Commercial Banks. *International Journal of Academic Research in Accounting Finance and Management Sciences*, *4*(3), 187–192. https://doi.org/10.6007/IJARAFMS/v4-i3/1181
9. Leopard, S., Song, J., Priestley, J., Frankel, M., Leopard, S., Song, J., Priestley, J., & Frankel, M. (2016). Using Logistic Regression to Predict Credit Default Using Logistic Regression to Predict Credit Default. *Analytics Experience*, 4–9.
10. Memic, D., & Competition, B. (2015). Assessing Credit default using Logistics Regression & Multiple Discriminant Analysis: Empirical Evidence from Bosnia & Herzegovina. *Interdisciplinary Description of Complex Systems*, *13*(1), 128–153. https://doi.org/10.7906/indecs.13.1.13
11. Yurynets, R., & Yurynets, Zoryna , Dmytro Dosyn, Y. K. (n.d.). *Risk Assessment Technology of Crediting with the Use of Logistic Regression Model*.