International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Statistical Methods for Banking Sectors to Detect the Eligible Customers for Home Loan

Sujatha.V, Kalpanapriya.D

Abstract: ***Nowadays people are interested to avail loans in banks for their needs, but providing loans to all people is not possible to banks, so they are using some measures to identify eligible customers. To measure the performance of categorical variables sensitivity and specificity are widely used in Medical and tangentially in econometrics, after using some measures also if banks provide the loans to the wrong customers whom might not able to repay the loans, and not providing to customers who can repay will lead to the type I errors and type II errors, to minimize these errors, this study explains one, how to know sensitivity is large or small and second to study the bench marks on forecasting the model by Fuzzy analysis based on fuzzy based weights and it is compared with the sensitivity analysis.***

Key-Words: Sensitivity Analysis, Specificity Analysis, Triangular Fuzzy Number, True Positive, True Negative

I. INTRODUCTION

Normally in regression models, dependent variable is quantitative, whereas explanatory variables are either qualitative, quantitative, or mixture thereof. In this study dichotomous regression models variables are used as qualitative in nature.

II. THE NATURE OF QUALITATIVE RESPONSE MODELS

In this study of identifying the eligible customers for the loans, is depended on the measures of the banks if the customers satisfies those then the decision is yes or else no. Hence, the response variable or regress and, can take only two values, say, 1 if the person eligible for loan and 0 for not eligible. In other words, the regress is a binary, or dichotomous, variable called as dummy variables. This study suggests that the decision of the bank to provide loan, is a function of the Age, property (own house/ land), professional, income, bank balance, credits etc. In this model Y is qualitative; our objective is to find the probability of identifying a customer for loans. Hence, qualitative response regression models are said to be probability models.The bank data for probabilistic model is given below

Explanatory variables in this study

Manuscript published on November 30, 2019.
* Correspondence Author
Sujatha. V*, Assistant Professor, Department of Mathematics, VIT University, Vellore, India.
Kalpanapriya.D, Assistant Professor, Department of Mathematics, VIT University, Vellore,India

Number of government employees with more service: 63
Number of Pvt. Employees With more service: 15
Number of customers having Business with good Turn over : 15
Number of customers having Business with moderate Turnover: 7
Number of customers Below Age 30 : 0
Number of customers between Ages 30 and 40: 54
Number of customers between Ages between 40 and 50 : 46
Number of customers Above Age 50 : 0
Number of customers having Land in Rural Area: 44
Number of customers having Land in Urban Area : 14
Number of customers having Land in Village : 12
Number of customers having No land : 30
Cibil Score less than 300 : 28
Cibil Score between 300-600 : 5
Cibil Score between 600-700 : 57
Cibil Score more than 700: 10
Eligible for loan Yes :21
Eligible for loan No : 79

Measures followed in this study

Age-30 to 40, 41 to 50 and above 51
salary- below 30000, between 30000 to 40000, above 40000
land Property- well developed city, rural urban
Credit - no loans , loans with good cibil score, loans with worst cibil score
Bank balance- 10000, below 10000, nil
Each attributes has equal weightage of 0.2.

Table 1: Weight for the age

Age		
30-40	41-50	Above 50
0.4	0.6	0

Table 2: Weight for the salary

Below 30000	Between 30000 to 40000	Above 40000
0.2	0.3	0.5

Table 3: Weight for land

Well-developed cite	Urban	Rural
0.5	0.4	0.1

In this study, the scores are given to the attributes based on major and minor criteria.The total scores of the major and minor attributes classified the outcomes. Based on total scores the bankers identify their eligible customers, some times which leads to Type I errors and Type II errors.

III. Materials and Methods

Regression Model is

$$Y = x'\beta + u,$$

where $x'\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, and Y is a binary variable, $E[y|x] = x'\beta$,

Because E [u |x] = 0, because y is a random variable that can have only values 0 or 1, we can define probabilities for y as P(y = 1|x) and P(y = 0|x) = 1 − P(y = 1|x), such that,

E[y |x] = 0 · P(y = 0|x) + 1 · P(y = 1|x) = P(y = 1|x).

Thus, E[y |x] = P(y = 1|x) indicates the success probability.

$P(y = 1|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, the probability of success.

This is called the linear probability model (LPM). The Probit procedure is used to estimate the effects of one or more independent variables on a dichotomous dependent variable[4].

The slope coefficients indicate the marginal effect of corresponding x-variable on the success probability, i.e., change in the probability as x changes, or

$$\Delta P(y = 1|x) = \beta_j \Delta x_j .$$

In the OLS estimated model, $\hat{Y} = \beta_0 + \hat{\beta}_1 x_1 + \ldots \hat{\beta}_k x_k$

$\hat{Y}$ is the estimated or predicted probability of success. In order to correctly specify the binary variable, it may be useful to name the variable according to the ”success” category (in this study, loan eligibility = 1 and not eligible for loan = 0)

The parameters of the probabilistic models were estimated by OLS methods by using the SPSS.

IV. Results and Discussions

Table 4: Convergence Information

	Number of Iterations	Optimal Solution Found
PROBIT	17	Yes

Table 5: Parameter Estimates

	Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
PROBIT[a]	Occupation	-.011	.041	-.261	.794	-.092	.070
	Property holder	-.008	.040	-.191	.848	-.086	.071
	Cibil score	.007	.032	.221	.825	-.055	.069
	Intercept	-1.67	.144	-11.58	.000	-1.816	-1.527

a.PROBIT model: PROBIT(p) = Intercept + BX

Table 6,7: Chi-Square Test

Co-variances and Correlations of Parameter Estimates

		Occupation	Property holder	Cibil score
P I	Occupation	.002	.222	.003
	Property holder	.000	.002	-.001
	Cibil score	.000	.000	.001

		Chi-Square		Sig.
PROBIT	Pearson Goodness-of-Fit Test	10.888		1.000

a. Statistics based on individual cases differ from statistics based on aggregated cases.

Table 8: ANOVA[a]

Model		Sum of Squares		Mean Square		Sig.
1	Regression	.217		.054		.863[b]
	Residual	15.701		.169		
	Total	15.918				

a. Dependent Variable: Y

b. Predictors: (Constant), Cibil score, Property holder, Occupation, Age

Table 9: Model Summary

	R	R-square	Adj. R-square	Std. Error	Change Statistics					
					Change	Change	Df1	Df2	Change	Durbin-watson
1	.12	.01	-.03	.4	.01	.3	4	93	.8	1.9

Table 10: Coefficients

Model		Unstan.coeff.		Stand.coeff			Collinearity statistics	
		B	Std. Err	Beta	t	Sig.	Tolerance	VIF
	Constant	1.4	0.4		3.7	0.00	.923	1.083
	Occupation	-0.01	0.5	-0.10	-.09	.93	.915	1.093
	Age	0.01	.01	0.11	1.0	.31	.916	1.092
	Property holder	0.1	0.1	0.23	0.2	.83	.971	1.029
	Cibil Score	0.00	0.38	0.00	-0.001	1		

"Qualitatively," the results of the probability model for cibil score are highly statistically significant with estimated R^2 value as 0.014 and with F value 0.321. Turning to the interpretations of the findings, slope of each coefficient gives the rate of change in the conditional probability of the event occurring for a given unit change in the value of the explanatory variable. For instance, the coefficient of 0.009 attached to the variable for "age " means, holding all other factors constant, the probability of Age group is about 0.9 per cent. In this study Durbin Watson value is approximately1.910, which shows that there are some auto correlations among the samples, this lead us to Type I error and Type II errors [1].By chi-square test one may know the measures of the individual covariates are significantly different from the aggregate of the individual values of the covariates.

In Econometrics research the sensitivity of a test is the probability of its giving a 'positive' result when the bank provides the loan to the eligible customers indeed positive and specificity is the probability of getting a negative result when the bank provides the loan to the wrong customers is indeed negative. A theoretical optimal prediction result can achieve 100% sensitive (i.e. predicts all people from a wrong customer population as not eligible customers) and 100% specificity (i.e. not predicts any from the eligible customers population).Using the measures, of the values, banks can either give the loans to the customers, or it may deny the loans to the customers [2].

- Identifying the wrong customers correctly as not eligible customers termed as "True positive"
- Identifying eligible customers as wrongly identified as not eligible customers as "False positive"
- Identifying eligible customers as correctly identified as eligible customers as "True negative"
- Identifying not eligible customers as wrongly identified as eligible customers as "False negative"

From these scenarios, two errors can occur, when an eligible customer is denied for loan leads to Type I error and not eligible customer is provided by the loans as Type II error. Type I error is said to be "α" error, or a false positive", the error of rejecting null hypothesis when it is actually true otherwise called as bankers risk. Type II error is said to be β error or false negative or consumers risk.(they has to bare the difficulties to repay the loan). The error of accepting null hypothesis when it is false. The following table illustrates the condition:

Table11a: Sensitivity and Specificity Analysis

		Actual Condition	
		Present	Absent
Test Result	+ve	Eligible to pay loan + Sanctioned loan = True Positive	Eligible to pay loan + Not sanctioned loan = False Positive (Type I error)
	-ve	Not Eligible to pay loan +Sanctioned loan= False negative (Type II error)	Not Eligible to pay loan +Not sanctioned loan = True negative

The probability of a false positive is estimated by Bayes theorem which is related to conditional and marginal probability of the random variables, false positive and false negative are the functions of the accuracy.

Mathematical Approaches to Hypothetical Testing

A test with low sensitivity has a high type II error rate. A test with low sensitivity refers to more false negatives, and also a test with high specificity has a low type I error rate [5]. The relationship among terms can be illustrated as follows:

Table11b: Sensitivity and Specificity Analysis

		Actual Condition		
		Present	Absent	
Test Result	+ve	True Positive	False Positive (Type I error)	+ve Predicted Value (PPV)
	-ve	Not Eligible to pay loan + Sanctioned loan= False negative (Type II error)	Not Eligible to pay loan + Not Sanctioned loan = True negative	-ve Predicted Value (NPV)
		Sensitivity (Type II error)	Specificity (Type I error)	

	Success	Failures	Total
Success	Eligible to pay loan + Sanctioned loan = True Positive (35)	Eligible to pay loan + Not Sanctioned	61
			loan

		= False Positive(2) (Type I error)	
Failures	Not Eligible to pay loan + Sanctioned loan (18) = False negative (Type II error)	Not Eligible to pay loan + Not Sanctioned loan (19) = True negative	37
Total	53	45	98

Sensitivity = [TP/(TP+FN)]x 100 = 66 %
Specificity = [TN/ (TN+FP)] x100 = 42.22%
False positive rate (α) = [FP / (FP+TN)] x 100 =58 %
False Negative rate (β) = [FN / (FN+TP)] x 100 = 34 %

Fuzzy Approaches to Hypothetical Testing

In this chapter a fuzzy logic model for retail loan evaluation and an application of fuzzy logic to commercial loan analysis is introduced. The fuzzy model consists of five input variables such as "Age", "Cibil score history", "mode of income"," ongoing loans" and "Property" . Here the customer's credit standing is awarded by the triangular fuzzy number [7,8]. Since with using fuzzy logic and triangular fuzzy numbers it is promising to apply qualitative information as linguistic variables and used them for quantitative process of determining the Sensitivity and specificity Analysis.

Table12: Qualitative information

Customer	Age	Property	Cibil score	Mode of Income	Other Loan
1	(33; 34; 35)	(0.5,0.6,0.7)	(0.3;0.4;0.5)	yes	yes
2	(28; 29; 31)	(0.6,0.7,0.8)	(0.5,0.6,0.7)	no	yes
3	(28; 29; 31)	(0.6,0.7,0.8)	(0.5,0.6,0.7)	no	yes
4	(48; 50; 51)	(0.4,0.5,0.6)	(0.3,0.4,0.5)	no	no
5	(26; 27; 28)	(0,0.1,0.2)	(0.5,0.6,0.7)	no	no
6	(47; 48; 49)	(0.4,0.5,0.6)	(0.5,0.6,0.7)	no	yes
7	(55; 56; 57)	(0.6,0.7,0.8)	(0.6,0.7,0.8)	yes	yes
8	(35; 36; 37)	(0.3,0.4,0.5)	(0.5,0.6,0.7)	no	yes
9	(49; 50; 51)	(0.4,0.5,0.6)	(0.5,0.6,0.7)	no	no
10	(33; 34; 35)	(0.6,0.7,0.8)	(0.5,0.6,0.7)	no	Yes

The crisp real number $\mu_{\tilde{a}}$ corresponding to the triangular fuzzy number, $\tilde{a} = (a,b,c)$ is obtained from the following relation

$$\mu_{\tilde{a}} = \frac{a+b+c}{3}$$

Table 13: Membership grades

Customer	Age	Property	Cibil score	Mode of Incom	Other Loan
1	0.6	0.6	0.4	0.4	0.3
2	0.1	0.7	0.6	0.1	0.3
3	0.1	0.7	0.6	0.2	0.3
4	0	0.5	0.4	0.3	0.7
5	0.1	0.1	0.6	0.4	0.7
6	0.3	0.5	0.6	0.4	0.3
7	0	0.7	0.7	0.3	0.3
8	0.6	0.4	0.6	0.1	0.3
9	0	0.5	0.6	0.2	0.7
10	0.6	0.7	0.6	0.3	0.3

Table 14: Estimation of Sensitivity and Specificity

	Success	Failures	Total
Success	Eligible to pay loan + Sanctioned	Eligible to pay loan + Not	6
Failures	Not Eligible to pay loan + Sanctioned	Not Eligible to pay loan +	4
Total	2	8	10

Sensitivity = [TP/(TP+FN)]x 100 = [1/2] x 100 = 50 %
Specificity = [TN/ (TN+FP)] x100 = [3/8] x100 = 37.5%
False positive rate (α) = [FP / (FP+TN)] x 100 =62.5 %
False Negative rate (β) = [FN / (FN+TP)] x 100 = 50 %
Power=50%

An approach of Machine learning Metrics to Hypothetical Testing

The metrics precision , Recall and F-measure are also identified as

Precision= [TP/(TP+FP)]x 100 = [1/6] x 100 = 16.7 %

Specificity = [TP/ (TP+FN)] x100 = [1/2] x100 = 50%

F-Measure= $2 \times \frac{\Pr ecision \times \operatorname{Re} call}{precision + \operatorname{Re} call} = 0.249$

Table 15: Comparison table

	M.L. metrics	Statistical approach	Fuzzy approach
Precision	16.7%	66%	50%
Recall	50%	42.2%	37.5%
F-Measure	0.249	.51	0.42

VI CONCLUSIONS

Probability model is most appropriate to estimate the effects of one or more independent variables on a binomial dependent variable In probabilistic model, taken age as a factor variable, and the remaining variables were taken as covariates, so that the results implies that there are autocorrelations among the variables, which leads to the Type I and Type II errors. About 58% of the customers even though they are eligible to pay loan due to some other factors they were not qualified to get their loans, and also 34% of the customers even though they are not eligible to pay loan, they were provided by bank loans, so there exists risks to re-collect the amount for banks.

REFERENCES

1. Altman,D.G and Bland,J.M.(1994).Statistics notes: Diagnostic tests :sensitivity and specificity,Br.Med.Jour.,308,p.1152.
2. Bennet, M.B. (1972) On Comparisons of Sensitivity, Specificity and Predictive Value of a Number of Diagnostic Procedures.Biometrics,28,793-800.
3. Sharma, D., Yadav, U.B. and Sharma, P. (2009) The Concept of Sensitivity and Specificity in Relation to Two Types of Errors and Its Application in Medical Research. Journal of Reliability and Statistical Studies, 2, 53-58.
4. Finney, D. J. 1971. Probit analysis . Cambridge: Cambridge University Press.
5. Gaddis, G.M. and Gaddis, M.L. (1990) An Introduction to Biostatistics Part 3: Sensitivity, Specificity, Predictive Value and Hypothesis Testing. Annals of Emergency Medicine, 19, 591-597.
6. Buckley. J.J(2005), Fuzzy Statistics, NY: Springer-Verlag.
7. Renkuan Guo, Yanhong and Cui Danni Guo (2012), Uncertainty Statistics, Journal of Uncertain Systems, Vol.6, No.3, pp.163-185.
8. Zadeh. L.A (1975), The concept of linguistic variable and its application to approximate reasoning I, II and III, Information Sciences ,8 199–249.