A Pivot Study of an Innovative Approach to Identify Potential Buyers: A Case of Logistic Regression

Dr Abhijeet Kaiwade[1], Dr Atik Shaikh[2], Prof Amit Kaiwade[3]

[1] *Director, Abhinav Education Society's Institute of Management and Research, Pune*
[2] *Professor, Allana Institute of Management Science, Pune*
[3] *Lecturer, MMCC, Pune and PhD Scholler, Abeda Inamdar Sr College, Pune*

Abstract:- Purpose of this study is to demonstrate application of logistic regression as predictive tool when outcome variable is categorical. In real life, many times we come across with categorical outcome variable. These categorical variables are either binomial or multinomial. In such situations, we cannot use linear regression, as it requires metric criterion variable. Logistic regression is extension of linear regression. A case of antique shop is discussed in this paper. The shop owner was interested to identify prospects who would buy antiques. Case study approach has used in this paper. Paper presents important statistics and model fit criteria for logistic regression. Paper concludes with developing significant logistic model which predict potential buyer. This paper is useful to students, research scholars, academicians and business managers to understand application of logistic regression as predictive tool.

Keywords*: Buying Behavior, Categorical variables, Logit Analysis, Logistic regression, Multivariate Data Analysis.*

1. Introduction

Today competition is fierce and business are struggling to sustain and grow. One of the keys for successful business is effective marketing strategy. Customers are at core of strategy formulation.

"So I think instead of focusing on the competition, focus on the customer"

- Scott Cook, Chairman of Intuit, Director of eBay and Procter & Gamble

As aptly said by Scott Cook, businesses need to focus on the customer. For any business knowledge of its customers is very important. The objective of this paper is to demonstrate how logistic regression can help managers to identify potential customers.

Linear regression is popular tool for prediction. It requires criterion variable measured on continuous scale. Continuous variables are considered as superior and more informative as compare to categorical variables. However, in reality we come across with many variables, which are categorical in nature. e.g. when you grow what would you become - a doctor, an engineer, a pilot etc., students either attend the class or not, students' grade in final exam may be either O or A or B etc., people who visit retail store either purchase or not.

Parents are interested to know what their child become when s/he grow. A teacher is interested to know the students who likely fail in exam. Store manager is interested to know the customers who would likely buy. What factors cause buying decision or non-buying decisions? Answers to these questions help managers to understand the reasons for buying and non-buying decisions. He can address the issues and attain stores' selling objective.

Purpose of this paper is to demonstrate application of logistic regression to classify customers in two groups-buyer and non-buyer. A case of antique store is discussed in this paper. This paper is useful to students, research scholars, academicians and business managers.

Literature Review:-

Logistic regression is a predictive technique to answer above-mentioned questions. Multiple regression is popular tool used for prediction. However, one of the serious limitation of multiple regression is it requires criterion variable strictly measured on either interval or ratio scale (Johnson & Wichern, 2007). In case of logistic regression, criterion variable is categorical (binomial or multinomial). Logistic regression predicts categorical outcome. Depending on number of categories for criterion variable, there are two types of logistic regression. When criterion variable have two categories, it is called as binary logistic regression. When categories are more than two it is called as multinomial (polynomial) logistic regression (Hosmer & Lemeshow, 2000). Logistic regression is extension of multiple regression (Darren & Mallery, 2011). Logistic regression is sometimes called as logit analysis (Hair, Black, Babin, & Anderson, 2009).

Logistic regression is used to predict group membership based on information available on predictor variables. e. g. based on information about income level, logistic regression predict buyer and non buyer of a car.

Logistic regression is used to understand relationship and strength among predictor and criterion variables (Burns & Burns, 2013).

2. Objectives

A popular antique shop was in business of selling antiques for more than 10 years in Pune. Owner of the shop observed that number of footfall to his shop has increased however; he wondered why sale has not increased proportionately. He thought of sending promotional mailers to his prospects. He was not sure who could be the potential buyers. This research was undertaken to help the owner to identify the probable buyers.

Two hypotheses are proposed for this study –

H1-1: Income level is significant predictor of buying decision.

H1-2: Interest in antiques is significant predictor of buying decision.

3. Methods

A case study approach has used in this research. The shop has maintained data of its visitors. These data contains name, contact details and purchase details. Shop visitors data of previous one year was selected as sampling frame for the study. Sampling frame was divided in two groups- buyers and non-buyers. Buyers were those who had purchased atleast once from the shop in last one year. A questionnaire was prepared and mailed to 500 visitors (300 buyers and 200 non-buyers). More number of buyers were selected, as shop owner was interested to identify the prospects. Samples were selected using systematic random sampling approach from both groups. Out of 500 targeted visitors 233 participated in survey. This has yielded 46.6% response rate. Sample of 200 or more is considered as large sample (Field, 2009). The participants' number 233 for this study exceed threshold of 200 for large sample size. Data was entered in SPSS (Version 18) and analysed using Logistic regression.

Instrument Development:-

Previous studies indicated that income, gender, age, hobby or interest, convenience are some of the important factors of buying behaviour (Loudon, L. & Bitta, Della, J., 2002; Schiffman, Kanuk, & Kumar, 2010). Scale for interest was developed for this study. The scale comprises of five items, measured on five point Likert scale (Strongly agree through strongly disagree). Open ended questions were asked to collect responses on income, gender and age.

Validity and Reliability of Scale:-

Face and Content Validity of the instrument was established by a panel of exports. Reliability measure Cronbach's alpha was used to establish reliability. Alpha value should be 0.7 or more (Nunnaly, 1978) for a scale to be reliable. Table No. 1 presents alpha value for the scale interest. Alpha value is greater than threshold value of 0.7. Thus, reliability of scale is established.

Reliability Statistics for scale Interest

Cronbach’s Alpha	N of Items
0.887	5

4. Results

Data was collected using Google form. Google form collects data in a excel sheet. The data in excel sheet cannot be used as it is in SPSS. All responses were assigned unique codes. Responses in excel sheet were replaced with these codes. These coded data then imported in SPSS. After cleaning and validating the data logistic regression was run.

Output of the logistic procedure was presented in Table No. 2 through Table No. 9. For sake of simplicity, only important tables are presented here. (Note: To ensure confidentiality, data analysis with only two predictors is presented here)

Table No. 1 Classification Tablea,b for Block 0 (Beginning Block)

			Predicted		
			Buying decision		Percentage Correct
	Observed		No	Yes	
Step 0	Buying decision	No	0	51	.0
		Yes	0	182	100.0
	Overall Percentage				78.1
a. Constant is included in the model. b. The cut value is .500					

Table No. 2 is a classification table for Block 0 (Beginning Block). It shows prediction percentage when predictors are not included in the model. Without including any predictors, model predicts 78.1% buying decisions accurately. This is compared with prediction percentage of model where in predictors are included. For a better model, we expect latter model should have greater prediction percentage than the former one.

Table No. 2 Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	1.272	.158	64.474	1	.000	3.569

Table No. 3 shows variable in the equation at step 0. At step 0, only constant is included in the model. In this case constant is significant (p<.001)

Table No. 3 Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Income	36.012	1	.000
		Interest	37.280	1	.000
	Overall Statistics		71.797	2	.000

Table No. 4 presents details of variables not included in the equation at step 0. Both variables Income and Interest are significant as p for both are less than .001. Next part of output (Table No. 5 through Table No. 9) presents results of model with predictors included.

There are various methods of inclusion of predictor variable in the model available. In this case, Enter method is used. Enter method enters all predictor variables at once. Another method includes stepwise approach. They either enter or remove predictor variable one at a time. Discussion on these methods are beyond the scope of this study.

Table No. 4 Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	77.732	2	.000
	Block	77.732	2	.000
	Model	77.732	2	.000

Table No. 5 presents overall fit of model. Chi-square value 77.732 with 2 degree of freedom is significant (p<.001). This indicate model has poor fit if only constant is included and predictors have significant effect (Burns & Burns, 2013).

Table No. 5 Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	167.147a	.284	.436
a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.			

In Table No. 6, -2 log likelihood is difference between best fitting model and base line model (Burns & Burns, 2013). Cox & Snell R Square and Nagelkerke R Square are pseudo R square. They indicate how much variance in criterion variable is explained by model. Nagelkerke is mostly reported (Field, 2009). In our case 43.6% variation in criterion variable is explained by the model.

Table No. 6 Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7.847	8	.504

Hosmer and Lemeshow test tests difference between observed and model predicted values. For acceptable model this test should be non-significant (Hosmer & Lemeshow, 2000). Table No. 7 shows p > .05, which indicate model has good fit.

Table No. 7 Classification Tablea

			Predicted		
			Buying decision		
	Observed		No	Yes	Percentage Correct
Step 1	Buying decision	No	28	23	54.9
		Yes	9	173	95.1
	Overall Percentage				86.3
a. The cut value is .500					

Table No. 8 presents observed and predicated cases using best fitting model. After including predictors, model predicts 86.3% buying decisions, which is greater than prediction percentage of base line model i.e. 78.1% (Table No. 2). This indicate that model with predictor variable included is better model than model without predictor variable.

Table No. 8 Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1a	Income	1.139	.207	30.322	1	.000	3.125	2.083	4.687
	Interest	1.242	.232	28.580	1	.000	3.464	2.197	5.462
	Constant	1.803	.236	58.509	1	.000	6.069		
a. Variable(s) entered on step 1: Income, Interest.									

Table No. 9 provides important information about significance of each predictor variable and its respective odd ratio. B values are beta coefficients of predictors. Sig. column presents significance value for respective Wald statistics. Wald statistics informs whether beta coefficient of predictors significantly different that 0 (Field, 2009). In our case, both predictors income and interest are significant (p<.001). Last column in Table No. 9 is odds ratio. Odds for income is 3.125. It means, if level of income increases by one unit, odds of buying will increase by 3.125 times. Odds for interest is 3.464. It means, if level of interest increases by one unit, odds of buying will increase by 3.464 times.

Using beta coefficient values from Table No. 9, model is presented as bellow-

Equation 1

$$P_{(Buyer)} = \frac{1}{1 + e^{-(1.803 - 1.139\times\text{Income} + 1.242\times\text{Interest})}}$$

Equation 2

$$P_{(Non-Buyer)} = 1 - P_{(Buyer)}$$

By replacing values of income and interest in equation 1, shop owner can easily predict who would buy antiques from his shop. To predict non-buyer equation 2 has to be used.

5. Discussion

Logistic Regression successfully predicted buyer and non-buyer group membership for antique shop. Income and Interest are found to be significant and positive factors to determine whether visitor would buy antiques or not. The outcome has helped shop owner to identify prospects based on their income level and interest level in antiques.

Limitations and Future Scope for the study:-

Emphasis of this paper is on binary logistic regression. However, this paper does not demonstrate multinomial logistic regression. Further, this paper does not discuss various methods used for entering predictor variables in the model. This study has not used categorical predictor variable.

This is a case study of a single antique shop, therefore, findings of this study may not be applicable to other stores.

A more extensive study is required to overcome the limitations of this study. Further study may include more number of shops, more number of predictor variables (including categorical variables), and multinomial criterion variable.

Refrences

[1] Burns, R. A., & Burns, R. B. (2013). Business Research Methods and Statistics Using SPSS. SAGE..

[2] Darren, G., & Mallery, P. (2011). SPSS for Windows Step by Step: A Simple Guide and Reference (11th ed.). Pearson Education Inc.

[3] Field, A. (2009). Discovering Statistics Using SPSS (3rd ed.). SAGE Publications Ltd..

[4] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). Multivariate Data Analysis.

[5] Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression (2nd ed.). New York: John Wiley and Sons.

[6] Johnson, R. a., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Pearson (Vol. 6).

[7] Loudon, L., D., & Bitta, Della, J., A. (2002). Consumer Behavior (4th ed.). New Delhi: Tata McGraw-Hill.

[8] Nunnaly, J. C. (1978). Psychometric Theory. McGraw-Hill (2nd ed.). New York: McGraw-Hill.

[9] Schiffman, L. G., Kanuk, L. L., & Kumar, S. R. (2010). Consumer Behavior (10th ed.). New Delhi: Pearson Education Inc.

[10] B. Bhamangol, A. Kaiwade, B. Pant, A. Rana, A. Kaiwade and A. Shaikh, "An Artificial Intelligence based Design and Implementation for classifying the Missing Data in IoT Applications," 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, 2022, pp. 1376-1382, doi: 10.1109/IC3I56241.2022.10072634.