ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

Logistic Regression: A Novel Approach to Implement in Business Intelligence

D.Kalaivani [1], Dr.T.Arunkumar[2]

Assistant Professor & Head, Dept. Computer Technology, Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamilnadu, India[1]

Professor & Dean, School of Computing Science, VIT University, Vellore, Tamilnadu, India [2]

Abstract: Business Intelligence is defined as a set of mathematical model and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes.[1] Business Intelligence encompasses all aspects of gathering, cleansing, mining, storing and analyzing data as well as disseminating the insights to the right decision makers. Data warehousing and analytic modeling are as much a part of a BI strategy as are visualization tools and digital dashboards [1].Decision tree learning is the most popular and powerful approach in knowledge discovery as well as in data mining. This is used for exploring large and complex bodies of data in order to discover useful patterns. Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. Classification algorithm processes a training set containing a set of attributes. There is a growing popularity of Internet as a medium of information search, communication link and online buying worldwide including India[2].This paper highlights the research opportunities in Business Intelligence [BI]. It also analyses the statistical method Logistic Regression.

Keywords: Business Intelligence, Data Mining, Buyer Behaviour Prediction, Decision Making, Knowledge Management.

I. INTRODUCTION

Gartner states that in 2011, Business Organizations are focusing on three key things: increasing enterprise growth, reducing costs, and attracting new customers. BI is a term that encompasses a broad range of analytical software and solutions for gathering, consolidating, analysing and providing access to information in a way that is supposed to let an enterprise's users make better business decisions.Business Intelligence (BI) are the set of strategies, processes, applications, data, products, technologies and technical architectures which are used to support the collection, analysis, presentation and dissemination of business information [8].

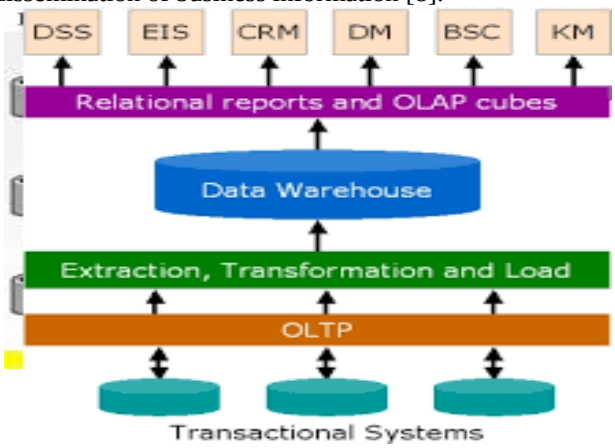


Fig.1. Layers in Business Intelligence

People are switching to On-line buying from traditional Physical shopping (Joines et.al., 2003:and Jeyavardhana 2004).Data Mining indicates the process of exploration and analysis of a dataset, usually of large size in order to find regular patterns [9].·Various Business Intelligence approaches are currently used like spread sheets and databases, Online Analytical Processing (OLAP), Online Transactional Processing (OLTP), Data Mining to assist with strategic planning in online retail (L.VenkataSubramaniam et.al., 2009). Business Intelligence has two basic different meanings related to the use of the term Intelligence[6]. Good Architecture is the key for successful BI implementation.

Small opportunities are often the beginning of great enterprises. (Demosthenes et.al). There are still significant differences between offline and online buyer behaviour that warrant a distinguishing conceptualization. Using Extraction and transformation tools known as Extract, Transform, Load [ETL], the data originated from different sources are stored in the databases which are referred to as Data Warehousing and Data Marts.

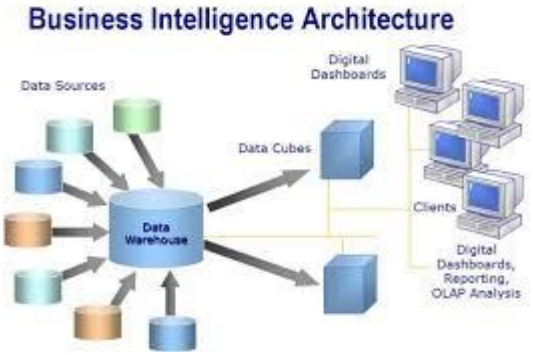


Fig.2. Business Intelligence Architecture

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

1.1 Components of Business Intelligence

- Multidimensional aggregation and allocation
- Denormalization, tagging and standardization
- Realtime reporting with analytical alert
- A method of interfacing with unstructured data sources
- Group consolidation, budgeting and rolling forecasts
- Statistical inference and probabilistic simulation
- Key performance indicators optimization
- Version control and process management
- Open item management

1.2 Data Exploration

The main purpose of Exploratory Data Analysis is to highlight the relevant features of each attribute contained in the Dataset using graphical methods and calculating summary statistics and to identify the relationships among the attributes [11].

The three main phases of Exploratory Data Analysis are:

- Univariate Analysis:-The properties of each single attribute of a Dataset are investigated.
- Bivariate Analysis:-Pairs of attributes are considered to measure the intensity of relationship existing between them.
- Multivariate Analysis:-The relationship holding within a subset of attributes are investigated.

According to Russell and Petersen (2000), market basket analysis focuses on the decision process by which a consumer selects items from a given set of product categories on the same shopping trip.

II. Related Work

Harold M.Campbell created a BI model through Knowledge management in his paper, "The role of Organizational Knowledge Mangement strategies in the quest for Business Intelligence."

Hai Wang proposed a BI model of Knowledge Development through Data Mining in the research paper, "A knowledge management Approach to Data Mining process for Business Intelligence".

L.R.Vijayasarathy (2001) integrated the web-specific factors influencing online shopping into theory of reasoned Action [TRA] to better explain the buyer online shopping behavior.

Li Niu,Jie Lu, et al., (2007) proposed a Cognitive Business Intelligence System(CBIS) in their research on "An Exploratory Cognitive Business Intelligence System".

Venkatadri.M [2010] presented the paper titled as A Novel Business Intelligence Framework that states Business Intelligence [BI] system plays a vital role in effective decision making.

Gartner(1996) defined BI as the application of a set of methodologies and technologies such as J2EE, DOTNET, Web Services, XML, Data Warehouse, OLAP, Data Mining, Representation Technologies etc.,.

Brad Quinn Post presented the paper titled "Building the Business Case for Group Support Technology" (1992) (Adelman et.al, 2002).

(Malhotra, 2000) describes BI that facilitates the connections in the new-form organization, bringing real-time information to centralized repositories and support analytics that can be exploited at every horizontal and vertical level within and outside the firm. BI describes the result of in-depth analysis of detailed business data.

III.PROPOSED METHODOLOGY

Logistic Regression Algorithm to predict the Customer Buying Behaviour: The 2008 survey of 300 companies by the Aberdeen Group-2008 [13] shows that the recent economic downturn has lengthened traditional sales cycles. As businesses have been forced to reduce spending, sales representatives have been challenged to meet quotas. Top performing companies have implemented sales intelligence programs to improve the quality and quantity of sales leads. SI contextualizes opportunities by providing relevant industry, corporate and personal information. Frequently SI's fact-based information is integrated or includes customer relationship management (CRM). In statistics, logistic regression, or logit regression, or logit model [14] is a regression model where the dependent variable (DV) is categorical. This article covers the case of binary dependent variables—that is, where it can take only two values, such as pass/fail, win/lose, alive/dead or healthy/sick. Cases with more than two categories are referred to as multinomial logistic regression, or, if the multiple categories are ordered, as ordinal logistic regression[14].

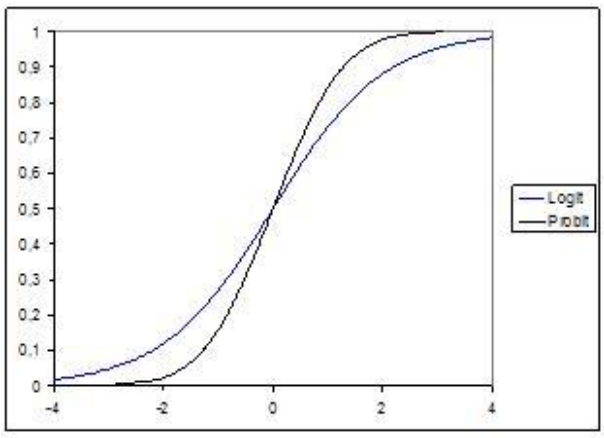


Fig.3 Logistic Regression

$$\textbf{logit}(p) = \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 x_{i2} + \beta_2 \cdot x_{i2} + \_ + \beta_p \cdot x_{in}$$

for $i = 1 \ldots n$ .., When selecting the model for the logistic regression analysis another important consideration is the model fit. Adding independent variables to a logistic regression model will always increase its statistical validity, because it will always explain a bit more variance of the log odds (typically expressed as R^2). However,

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

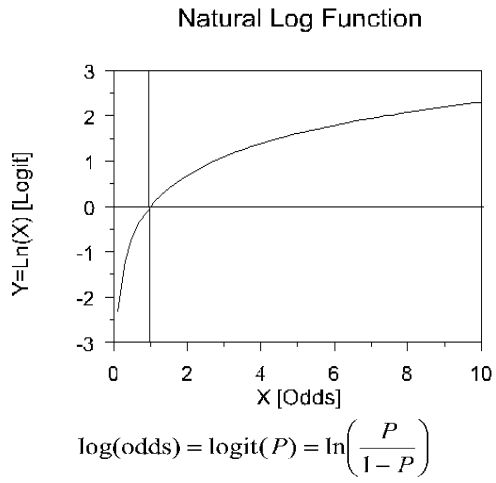
adding more and more variables to the model makes it inefficient and over fitting occurs.

The Logistic Curve: The logistic curve relates the independent variable, X, to the rolling mean of the DV, P ($\bar{Y}$). The formula to do so may be written either

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

$$P = \frac{1}{1+e^{-(a+bX)}}$$

Where, P is the probability of a 1 (the proportion of 1s, the mean of Y), e is the base of the natural logarithm (about 2.718) and a and b are the parameters of the model. The value of a yields P when X is zero, and b adjusts how quickly the probability changes with changing X a single unit (we can have standardized and unstandardized b weights in logistic regression, just as in ordinary linear regression).Probability of Buying Behaviour of the customer is predicted. A full model could have included terms for the continuous variable, the categorical variable and their interaction (3 terms). Restricted models could delete the interaction or one or more main effects (e.g., we could have a model with only the categorical variable). A nested model cannot have as a single IV, some other categorical or continuous variable not contained in the full model. If it does, then it is no longer nested, and we cannot compare the two values of -2LogL to get a chi-square value.



$$\log(\text{odds}) = \text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

The chi-square is used to statistically test whether including a variable reduces badness-of-fit measure. This is analogous to producing an increment in R-square in hierarchical regression. If chi-square is significant, the variable is considered to be a significant predictor in the equation, analogous to the significance of the b weight in simultaneous regression. Because the relation between X and P is nonlinear, b does not have a straightforward interpretation in this model as it does in ordinary linear regression.

Note that the natural log is zero when X is 1. When X is larger than one, the log curves up slowly. When X is less than one, the natural log is less than zero, and decreases rapidly as X approaches zero. When P = .50, the odds are .50/.50 or 1, and ln(1) =0. If P is greater than .50, ln(P/(1-P) is positive; if P is less than .50, ln(odds) is negative. The customer increases online-buying by 50%.A number taken to a negative power is one divided by that number, e.g. e-10 = 1/e10. A logarithm is an exponent from a given base, for example ln(e10) = 10.

IV. RESULTS AND DISCUSSION

In logistic regression, the dependent variable is a logit, which is the natural log of the odds. So a logit is a log of odds and odds are a function of P, the probability of a 1. In logistic regression, we find **logit(P) = a + bX,** Prediction is carried out according to the results of LOGIT(P).Which is assumed to be linear, that is, the log odds (logit) is assumed to be linearly related to X, our IV. So there's an ordinary regression hidden in there. We could in theory do ordinary regression with logits as our DV, but of course, we don't have logits in there, we have 1s and 0s. Then, too, people have a hard time understanding logits. We could talk about odds instead. Of course, people like to talk about probabilities more than odds. To get there (from logits to probabilities), we first have to take the log out of both sides of the equation. Then we have to convert odds to a simple probability: Logistic Regression Is a Novel Approach for Predicting Buying Behaviour of Customers.

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

The simple probability is this ugly equation that you saw earlier. If log odds are linearly related to X, then the relation between X and P is nonlinear, and has the form of the S-shaped curve you saw in the graph and the function form (equation) shown immediately above.

V. CONCLUSION

The Analysis of Datasets to predict the Customer Buying Behaviour gives its gain in rich analysis of combining all-round view of Customer. There are some technical challenges like selection of datasets, analyses, associating, linking, pattern recognition, classification, pruningetc. Customer communications include e-mails, text messaging, chat transcripts, agent notes etc., the data are collected and implemented in the model and tested

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

appropriately. The various algorithms which are used for the classification of data are decision trees, linear programming, neural network and statistics. Among these algorithms Decision trees is one of the most popular and powerful approaches in data mining. The model acts as a crucial business developer which knowledge-centered approach. This paper demonstrates the various processes involved in successful prediction of the Customer Buying Behaviour.

ACKNOWLEDGMENT

I would like to thank my family members and colleagues for their support and encouragement extended. I am always grateful to my Research Supervisor Dr.T.Arunkumar for his valuable guidance and mentorship.

REFERENCES

[1] Carlo Vercellis, "Business Intelligence" – Data Mining and Optimization for Decision Making"-Wiley Student Edition-2009
[2] ArchanaShrivastava et al," Behavioural Business Intelligence Framework for Decision Support in Online Retailing in Indian Context."- International Journal of Scientific and Research Publications, Vol.2, issue May 2012 - ISSN-2250-3153.www.ijsrp.org.
[3] Richard Herschel," International Conference on Information Technology Interfaces (ITI 2011) in Cavtat, Croatia sponsored by the University of Zagreb.
[4] The Wall Street Journal (March 28, 2011)"U.S. Products Help Block Mideast Web."
[5] State of Business Intelligence: Results from Survey of BI Professionals(2012)
[6] JayanthiRanjan, " Business Intelligence: Concepts , Components, Techniques and Benefits" , - Journal of Theoretical and Applied Information Technology,2009.,Vol.9.,No.1.,(pp.60-70)
[7] JiaweiHan Michelin Kamber (2011), "Data Mining-Concepts and Techniques", Morgan Kaufmann Publishers.
[8] Dedić N. &Stanier C. (2016). Measuring the Success of Changes to Existing Business Intelligence Solutions to Improve Business Intelligence Reporting. Lecture Notes in Business Information Processing. Springer International Publishing. Volume 268, pp. 225-236.
[9] Lawrence r., Almasi G., Kotlyar V., Viveros M., Duri.S , Personalization of supermarket product recommendations.Data Mining and Knowledge Discovery, Vol.5, 11-32. (2001)
[10] Kudyba S., Hoptroff R. (2001), Data Mining and Business Intelligence: A Guide to productivity., Idea Group.
[11] Agrawal.J. (2001) Data Mining for Association rules and sequential patterns:sequential and parallel algorithms.,Springer.
[12] GorryG.,Scott Morton M., (2003) , A Framework for management information systems,Sloan Management Review ,Vol.13., pp.,55-70
[13] Sales Intelligence, Aberdeen Group Study – 2008
[14] David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press. p. 128.

BIOGRAPHY

D.Kalaivani has received B.Sc., Computer Science and Master of Computer Applications from Bharathiar University. She has got M.Phil. Degree from Bharathidasan University. She is working as Assistant Professor & Head, Department of Computer Technology in Dr.SNS Rajalakshmi College of Arts and Science.