Volume 3 Issue 1

ISSN: 2321-1857

International Journal of

Business Analytics & Intelligence

April 2015

A Publication of Publishing India Group

International Journal of Business Analytics & Intelligence

Editorial Message

It's a matter of great joy for the IJBAI fraternity to step into an accomplished third year. With more relevant research papers reaching us, we are witnessing a steady growth in the quality of the journal, true to the vision and philosophy. The papers in the current issue will enlighten our readers to multiple analytics models like application of data envelopment analysis for marketing effectiveness, comparison between support vector machines and logistic regression for propensity based response modelling, analytics used at small and medium enterprises, comparison of logistic regression and artificial neural network on bankruptcy prediction models and the application of sentiment analysis. We are extremely thankful to our erudite editorial board who has meticulously selected the most valuable papers. We are also thankful to the authors for choosing IJBAI in sending their precious research papers to us. Last but not the least, we are extremely thankful and grateful to the number of our loyal readers without whose motivations the journey could never have been so vibrant and colourful. May IJBAI cruise with more accolades through the glorious path ahead!

Tuhin Chattopadhyay, Ph.D.
Editor-in-Chief
International Journal of Business Analytics and Intelligence

International Journal of Business Analytics and Intelligence

Volume 3 Issue 1 April 2015

ISSN: 2321-1857

Journal is available online at www.publishingindia.com

Article can be accessed online at http://www.publishingindia.com

Measuring the Efficiency of Marketing Efforts in the Indian Pharmaceutical Industry using Data Envelopment Analysis

Mihir Dash*, Arunabhas Bose, Samik Shome***, Shamim Mondal***, Dennis J. Rajakumar***, Ramanna Shetty***, Madhumita G. Majumdar****, Debashis Sengupta******

Abstract

Pharmaceutical companies have been spending huge amount of money on marketing and promotions, sales distribution, and travelling done by the sales representatives. However, they find it difficult to directly link the returns with these efforts. This study makes an attempt to examine whether the marketing efforts have significant influence on the sales performance in the industry. It uses the DEA model (Data Envelopment Analysis) to assess the efficiency of marketing efforts by pharmaceutical companies, and uses random effects maximum likelihood panel regression to assess the significance of the impact of marketing efforts.

Keywords: Pharmaceutical Industry, Marketing Efforts, Sales Performance, DEA Model, Random Effects Maximum Likelihood Panel Regression

Introduction

The pharmaceutical industry is a major segment of the Indian healthcare industry. It includes the industrial manufacture, separation, processing, refining and packaging of chemical materials. The Indian pharmaceutical industry meets around 70% of the country's demand for bulk drugs, drug intermediates, pharmaceutical formulations, chemicals, tablets, capsules, orals and injectables. It is ranked 3rd in terms of volume and 14th in terms of value globally. The domestic pharmaceuticals market was worth US$ 19.22 billion in 2012, and is expected to grow to US$ 55 billion in 2020.[1]

The Indian pharmaceutical industry is a highly competitive market, with a growth rate of 16% in 2012. There are almost 20,000 small and big players in the industry who strive hard to capture the market share by differentiating themselves from one another. Both domestic and global pharmaceutical market has become competitive and margins are reducing, so presently the industry is concentrating on manufacturing cost effective drugs in order to make exports possible. Today most companies in the industry have adopted or are in the process of adopting good manufacturing practices so that their products become easily acceptable both by domestic as well as international customers.

1 McKinsey Report on 'India Pharma 2020: Propelling access and acceptance, realizing true potential'

* Professor, Management Science, School of Business, Alliance University, Bengaluru, Karnataka, India. E-mail: mihirda@rediffmail.com
** Professor, Marketing, School of Business, Alliance University, Bengaluru, Karnataka, India. E-mail: arunabhas.bose@alliance.edu.in
*** Professor, Economics, School of Business, Alliance University, Bengaluru, Karnataka, India. E-mail: samik.shome@alliance.edu.in
*** Professor, Economics, School of Business, Alliance University, Bengaluru, Karnataka, India. E-mail: shamim.sm@alliance.edu.in
*** Professor, Economics, School of Business, Alliance University, Bengaluru, Karnataka, India. E-mail: dennis.rajakumar@alliance.edu.in
*** Professor, Economics, School of Business, Alliance University, Bengaluru, Karnataka, India. Email: rshetty@alliance.edu.in
**** Professor, Business Analytics, School of Business, Alliance University, Bengaluru, Karnataka, India. E-mail: madhumita.gm@alliance.edu.in
**** Professor, Business Analytics, School of Business, Alliance University, Bengaluru, Karnataka, India. E-mail: debashis.sengupta@alliance.edu.in

Marketing practitioners and scholars are under tremendous pressure to be more accountable for and to show how marketing expenditure adds to shareholder value. This apparent lack of accountability has undermined marketing's credibility and threatened marketing's standing in the pharmaceutical companies. There are three challenges for justifying marketing investments to measure marketing productivity: firstly, the challenge of marketing activities to long-term effects, and secondly, the separation of individual marketing activities from other actions, and finally, the use of purely financial methods.

There has been a continuous attempt to relate marketing efforts in terms of cost with respect to sales, and this has been a very relevant requirement in the pharmaceutical industry, where marketing activities are quite unique. There are difficulties in directly relating marketing effort to sales in most industries, as many extraneous variables other that marketing inputs affects sales. This often leads to incomplete and erroneous calculation of marketing effectiveness. However in the pharmaceutical industry, the success of products is predominantly dependent on marketing and sales efforts, as other factors have relatively less influence. The usual promotional strategies and modes common in consumer product industries are not so significant here as the sales of the products in this industry largely depend on the efforts of the sales force. As the influence of extraneous variables beyond marketing efforts is less in amount and intensity, it is easier to relate marketing efforts to sales performance and assess its effectiveness in the pharmaceutical industry. In this context, the present study attempts to examine the impact of marketing efforts on sales in the pharmaceutical industry in India.

Literature Review

The literature on marketing efforts is wide, especially for the pharmaceutical industry. One strand of literature focuses on marketing modeling, mathematically relating sales to different marketing efforts. Sinha & Zoltners (2001) emphasized the use of models in measuring sales activities, and suggested why new sales models were required to be looked into to measure performances in pharmaceutical industry. Momaya & Ambastha (2004) emphasized on the usage of mathematical model to understand enhanced competitiveness in pharmaceutical industry. Gagnon & Lexchin (2008) argued that the pharmaceutical industry is marketing-driven, with spending on promotion almost twice as much as spending on research and development. de Boeck *et al.* (2010) argued why the existing sales force marketing model need to be changed and a more relevant outlook is required.

On the other hand, another strand of literature focuses on the effectiveness of marketing efforts. Elling *et al.* (2002) argued that the system of assessing sales cost effectiveness was costly, inefficient, and rife with dissatisfaction, and for these reasons, pharmaceutical companies are considering what can be done to transform their sales model. Momaya & Ambasta (2004) suggested a change in sales force effectiveness measure for Indian pharmaceutical firms to compete in global markets. Gupta & Nair (2009) identified the need for pharmaceutical companies to reduce cost in sales effort and streamline the marketing activities.

Jakovcic (2009) discussed the impact that sales force has on sales, costs and profits, in both the short and the long term. He discussed different situations for sizing a sales force such as expansion into new markets, new product launches and downsizing, and he discussed three different methods that companies use to size their sales force. Agarwal *et al*. (2010) questioned the tradition ROI method to assess sales efforts and emphasizes the need of a new perspective to measure the marketing and sales effectiveness.

Palo & Murphy (2010) focused on the key forces that re-shape the pharmaceutical marketplace, including the growing power of healthcare payers, providers and patients, and the changes required to create a marketing and sales model that is fit for the 21st century. These changes will enable the industry to market and sell its products more cost-effectively, to create new opportunities and to generate greater customer loyalty across the healthcare spectrum.

According to Mariani (2008), companies can monitor the effects of promotional efforts through territorial market dynamics evaluation. Tools are applied in order to isolate the single contribution to information and product prescription. Medical-scientific information, like advertising and promotion, has the goals to improve brand and product notoriety, to improve the perception of product characteristics, and to augment the prescriptive propensity.

Performance measurement also has a wide literature. The balanced scorecard technique was proposed by Kaplan *et al*. (Kaplan & Norton, 1992, 1993, 1996) to understand the correlation between performance and strategies. Performance management received the focus of attention in the last two decades to analyse the multidimensional nature of the firm's performance (Anthony & Govindarajan, 2003; Kaplan & Norton, 1992, 1993, 1996; Zhu, 2000). However, the multidimensional performance measure may not be able to capture the various weights of the parameters explicitly (Ittner *et al*., 2003). Marketing efficiency can be measured in different ways: like (1) the net income generated by a marketing campaign divided by its cost, (2) the value premium attributable to a brand's reputation, (3) the customer lifetime value, which is the net present value of the revenue expected from a customer over the lifetime of the business relationship (Rust *et al*., 2004).

In this backdrop, the present study adopts data envelopment analysis (DEA) to analyse the performance of Indian pharmaceutical industry with respect to the multiple dimensions of their marketing efforts, in terms of tangible resources invested.

Methodology

The objective of the study is to measure the efficiency of marketing efforts of pharmaceutical companies in India. The sample companies selected for the study represented the top eleven pharmaceutical MNCs, with India-wide operations. Two of the companies, FKO and Organ on, were not considered for the analysis, as they represented very specialised segments, so that the final sample for the study included nine pharmaceutical MNCs. The study period was 2002-03 to 2011-12. The data for the study were collected from the Capitaline database.

The study uses data envelopment analysis (DEA) to measure the efficiency of marketing efforts in Indian pharmaceutical companies. DEA was first developed by Farrell (1957), and extended by Charnes *et al*. (1978). It is a non-parametric method that identifies what proportion of a unit's inputs are actually required to produce its given levels of outputs, as compared to other units. Mathematically, it is represented by the model expressed below.

$$\begin{aligned} \min E \quad s.t. \quad & \sum w_j = 1 \\ & \sum w_j I_{ij} \leq E.I_{i*} \\ & \sum w_j O_{ij} \geq O_{i!*} \end{aligned}$$

The inputs used in the study include marketing and promotional expenditure, distribution and selling expenditure, and travel expenditure. Marketing and promotional expenditure includes advertising expenditure, expenditure on sales promotions, and expenditure on marketing materials. Distribution and selling expenditure, for both primary and secondary sales, includes the costs of maintaining inventory through various channel members, logistics costs, and insurance costs. Travel expenditure, which is one the most significant marketing efforts in the pharmaceutical industry, includes travelling costs for marketing calls and other trade-related promotional activities. Sales revenue is taken as the output.

The study also considers a nonlinear form of DEA, taking logarithmic data in place of the input and output variables. The model is expressed as below.

$$\begin{aligned} \min E \quad s.t. \quad & \sum w_j = 1 \\ & \sum w_j \ln(I_{ij}) \leq E.\ln(I_{i*}) \\ & \sum w_j \ln(O_{ij}) \geq \ln(O_{i!*}) \end{aligned}$$

To examine the impact of marketing effort on the efficiency scores, random-effects maximum likelihood panel regression was performed. Panel data allows control for unobservable company specific factors or heterogeneity, or change in variables that vary over time but not over entities (for example, macroeconomic policies). The dependent variable was the efficiency of a particular company, and the independent variables were the proportion of marketing and promotional expenditure, the proportion of distribution and selling expenditure, and the proportion of travel expenditure. It was assumed that the company specific unobserved variables were not correlated with the independent input variables. Thus, a random-effects model was used rather than a fixed-effect model, in which the unobserved company specific heterogeneity is constant and its effect does not change over time.

Formally, the model for firm *i* at time *t* can be represented as:

$$E_{it} = \beta_1 MPE_{it} + \beta_2 DSE_{it} + \beta_3 TrE_{it} + u_i + \epsilon_{\text{it}}$$

where MPE represents the proportion of marketing and promotional expenditure, DSE represents the proportion of distribution and selling expenditure, and TrE represents the proportion of travel expenditure. The specification is linear, with the random effect captured in the term u_i, a firm-specific time-invariant random variable. It should be noted that because the proportions add up to unity, only two of these will form a linearly independent system allowing recoverability of parameter estimates, so that the model is specified without a constant term. The maximum likelihood method is used to estimate the parameters, fitting a normal distribution to u_i. The dependent variables are the linear efficiency scores and the nonlinear efficiency scores, in turn.

Findings

The marketing efforts distribution and the overall efficiency scores of the pharmaceutical companies are shown in Table 1.

The company with highest efficiency scores was Abbott (which was 100% efficient in all years except for 2005-06) with an average efficiency score of 99.14%. The company seems to have shifted its marketing efforts, with a decrease in the proportion of distribution and selling expenditure by almost 50%, and an increase in the proportion of marketing and promotional expenditure.

The company with next-highest efficiency was Glaxo SmithKline (which has been 100% efficient from 2006-07 onwards), with an average efficiency score of 97.68%. The company seems to have maintained a consistent marketing effort distribution, with almost equal emphasis on distribution and selling expenditure and travel expenditure, and almost twice as much emphasis on marketing and promotional expenditure.

Pfizer has experienced a different trend in efficiency. The company showed a continuous increase in efficiency until 2007-08, reaching 91.76%, and thereafter dropping to 54.15% in 2011-12. The company seems to have shifted its marketing efforts, with a decrease in the proportion of marketing and promotional expenditure, and an increase in the proportion of distribution and selling expenditure. Interestingly, in 2007-08, at the peak of its efficiency, the company had reached a high proportion of marketing and promotional expenditure and distribution and selling expenditure.

The company with lowest efficiency was Merck, which was 100% efficient in 2002-03, and which dramatically slipped to 36.43% in 2007-08, with some recovery to 60.06% in 2011-12. This could have been affected by the regime change in European Union health industry regulations in 2007, as a significant proportion of Merck's sales are to the European Union. The company seems to have shifted its marketing efforts, with a decrease in the proportion of distribution and selling expenditure and travel expenditure, and an increase in the proportion of marketing and promotional expenditure.

Novartis also experienced consistently low efficiency, with an average efficiency score of 64.03%. The company also seems to have shifted its marketing efforts, with an

Table 1: Marketing Efforts Distribution and Average Efficiency of the Sample Companies

Company	*%age Marketing & Promotional Exp*	*%age Distribution & Selling Exp*	*%age Travel Exp*	*Efficiency*	*Nonlinear Efficiency*
Pfizer	49.83%	24.49%	25.68%	72.36%	86.19%
Abbott	43.89%	22.75%	33.35%	99.14%	98.74%
Astrazeneca	36.66%	17.45%	45.89%	67.19%	72.16%
Novartis	54.81%	25.01%	20.18%	64.03%	79.27%
Sanofi	40.44%	26.81%	32.75%	75.28%	84.94%
Merck	39.45%	24.48%	36.06%	58.64%	73.23%
GSK	48.59%	25.92%	25.49%	97.68%	98.76%
Fulford	17.40%	31.23%	51.37%	84.40%	87.23%
Wyeth	48.98%	22.95%	28.07%	76.08%	75.61%
average	42.23%	24.57%	33.20%	77.20%	84.01%

Table 2: Random-effects Maximum Likelihood regressions

	Dependent variable: linear efficiency score				*Dependent variable: nonlinear efficiency score*			
	Coeff.	Std. Err.	z stat	P > |z|	Coeff.	Std. Err.	z stat	P > |z|
MPE	0.2914	0.1103	2.64	0.008**	0.6202	0.0714	8.69	0.000**
DSE	0.4802	0.1920	2.50	0.012*	0.8273	0.1283	6.45	0.000**
TrE	1.6150	0.1891	8.54	0.000**	1.1399	0.1244	9.17	0.000**
σu	0.1694	0.0457			0.1033	0.0280		
σe	0.1359	0.0108			0.0918	0.0073		
ρ	0.6084	0.1363			0.5589	0.1417		
Wald $\chi 2$	207.69				569.68			
Prob > $\chi 2$	0.000**				0.000**			
Log Likelihood	39.2925				75.5034			

* significant at 5%

** significant at 1%

increase in the proportion of marketing and promotional expenditure and travel expenditure, and a decrease in the proportion of distribution and selling expenditure.

To examine the impact of marketing effort on the efficiency scores, random-effects maximum likelihood panel regression was performed. The results are presented in Table 2.

The random effect models were significant, as indicated by the Wald test. The linear efficiency scores were found to be positively impacted by all three categories of marketing expenditures, and all coefficients were significant at a 5% level. Further, increasing the proportion of travel expenditure is the most effective way to increase efficiency, followed by distribution and selling expenditure and marketing and promotional expenditure. Thus reducing expenditure on marketing and promotional activities and increasing expenditures on travelling as well as distribution and selling would improve efficiency. The firm-specific heterogeneity is estimated to have a standard deviation of 0.1694 and the error term has an estimated standard deviation of 0.1359. The contribution of firm-specific heterogeneity to overall unobserved variability is 60.84%, as captured by the estimate ρ.

The nonlinear efficiency scores showed similar results. This model appears to be a better fit, as it had a higher log-likelihood compared to the linear efficiency scores. All coefficients were significant at 1% level. Here the coefficient estimates indicate the percentage improvement in efficiency if the allocation to a particular marketing expenditure category is increased by 1%. The results are similar to those of linear efficiency scores: reducing expenditure on marketing and promotional activities and increasing expenditures on travelling as well as distribution and selling would improve efficiency.

Discussion

The variables considered in the study were marketing and promotional expenditure, distribution and selling expenditure, and travel expenditure. All the three variables had a significant impact on the efficiency scores, with travel expenditure being the most significant followed by distribution and selling expenditure and marketing and promotional expenditure. The log-likelihood ratio statistics for the nonlinear efficiency score model was considerably higher than that of the linear efficiency score model, suggesting that nonlinear efficiency score may be a more appropriate indicator for efficiency of marketing efforts in the pharmaceutical industry.

The above efficiency characteristics observed seems to align with the practicalities of marketing efforts in the pharmaceutical industry. In the Indian pharmaceutical industry where sales takes place by direct interaction of the sales force to the doctors and medical associations, it is quite pertinent that more travelling would lead to higher exposure and reach in the market, leading to possible enhancement in sales. Distribution also is an important aspect as availability of drugs in the market in due and appropriate time is an essential requirement. Promotional activities are unique in pharmaceutical industry, and

though an important aspect, it is primarily a support function compared to travelling and distribution.

There were some limitations inherent in the study. The study included only nine MNC pharmaceutical firms which contribute to about 55% of prescribed drugs in India. The study can be made more extensive by considering more number of pharmaceutical firms. Domestic pharmaceutical companies in India can be also included to find out whether similar characteristics persist or not. Further, the study can be extended to global perspectives by including pharmaceutical firms which operate in other domestic scenarios in different countries. Also, the variables considered were limited to only major aspects of marketing efforts, and disaggregated data were not available. The study can be extended to include other variables related to marketing efforts, such as sales force size, territorial spread, average sales calls made, and so on.

References

Agarwal, S., Ahlawat, H., & Hopfield, J. (2010). Optimizing spend: Changing the ROI game augmenting reach and cost with a quality assessment to make more informed investment decisions. *Driving Marketing Excellence, Pharmaceutical and Medical Product Practice*, McKinsey Report, 28-35.

Anthony, R., & Govindarajan, V. (2003). *Management Control Systems*, (11th ed.). McGraw-Hill, New York, NY.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring efficiency of decision making units. *European Journal of Operational Research,* 2(6), 429-44.

de Boeck, P., Detlefs, S., & Villumsen, K. (2010). Ten ideas to improve the front line: The future of pharmaceutical sales force. *The eYe of the Storm, Perspectives and Recommendations for European Commercial Pharmaceuticals*, McKinsey Report, 72-78.

Elling, M. E., Fogle, H. J., McKhann, C.S., and Simon, C. (2002), *Making more of pharma's sales force*, McKinsey Quarterly Report.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society*, Series A, CXX, Part 3, 253-290.

Gagnon, M., & Lexchin, J. (2008). The cost of pushing pills: A new estimate of pharmaceutical promotion expenditures in the United States. *PLoS Med,* 5(1).

Gupta, M., & Nair, R. (2009). *Making an Impact: Effective Sales and Marketing With Reduced Costs Leveraging offshore resources to do more with less*, Indegene Report.

Ittner, C., Larcker, D., & Meyer, M. (2003). Subjectivity and the weighting of performance measures: evidence from a balanced scorecard. *The Accounting Review*, 78(3), 725-58.

Jakovcic, K. (2009). Pharmaceutical sales force effectiveness strategies: evaluating evolving sales models & advanced technology for a customer centric approach. *Business Insights Report*.

Kaplan, R. S., & Norton, D. P. (1992). The balanced scorecard: measures that drive performance. *Harvard Business Review,* 70(1), 71-9.

Kaplan, R. S., & Norton, D. P. (1993). Putting the balanced scorecard to work. *Harvard Business Review,* 71(5), 134-43.

Kaplan, R. S., & Norton, D. P. (1996). *The balanced scorecard*. Harvard Business School Press, Boston, MA.

Mariani, P. (2008), *Sales Force Effectiveness in Pharmaceutical Industry: an Application of Shift-Share Technique, Simulated Annealing Theory with Applications*, Sciyo, Croatia.

Momaya, K., & Ambastha, A. (2004). Competitiveness of firms: Review of theory, frameworks & models. *Singapore Management Review,* 26(1), 45-61.

Palo, J. D., & Murphy, J. (2010). *Pharma 2020: Marketing the future. Which path will you take?* PriceWaterhouseCoopers Report.

Rust, R.T., Ambler, T., Carpenter, G. S., Kumar, V., & Srivastava, R. K. (2004). Measuring marketing productivity: Current knowledge and future directions. *Journal of Marketing,* 68, 76-89.

Sinha, P., & Zoltners, A. A. (2001). Sales Force decision Models: Insights from 25 years of Implementation. *Interfaces,* 31(3), S8-S44.

Zhu, J. (2000). Multi-factor performance measure model with an application to Fortune 500 companies. *European Journal of Operational Research,* 123(1), 105-24.

Article can be accessed online at http://www.publishingindia.com

A Comparative Analysis of Support Vector Machines & Logistic Regression for Propensity Based Response Modeling

K. V. N. K. Prasad*, G.V.S.R. Anjaneyulu**

Abstract

Increasing cost of soliciting customers along with amplified efforts to improve the bottom-line amidst intense competition is driving the firms to rely on more cutting edge analytic methods by leveraging the knowledge of customer-base that is allowing the firms to engage better with customers by offering right product/service to right customer. Increased interest of the firms to engage better with their customers has evidently resulted into seeking answers to the key question: Why are customers likely to respond? in contrast to just seek answers for question: Who are likely to respond?This has resulted in developing propensity based response models that have become a center stage of marketing across customer life cycle. Propensity based response models are used to predict the probability of a customer or prospect responding to some offer or solicitation and also explain the drivers – why the customers are likely to respond. The output from these models will be used to segment markets, to design strategies, and to measure marketing performance.

In our present paper we will use support vector machines and Logistic Regression to build propensity based response models and evaluate their performance.

Keywords: Response Modeling, Propensity, Logistic Regression, Support Vector Machines

Introduction

A Propensity Model is a statistical scorecard that is used to predict the behaviour of customers or prospects base. Propensity models are extensively used in marketing arena to build list for solicitation and also act as a robust tool in creating tailored campaigns that are best received by customer. They help in developing analytical infrastructure that helps in identification of prospective opportunities and issues across the customer lifecycle, thus acting as a platform in understanding the dynamics of customer lifecycle.

Propensity models help in identification and generalisation of the "natural inclination or tendency" among the customer base for a given treatment. The identification and generalisation will help in understanding two important aspects – a) who are likely to respond when solicited? b) Why are the solicited customers likely to respond?The outcome of the propensity models will help to a larger extent in designing an optimal marketing strategy "reaching out to right customer with right product at right time through right channel at right price".

In our current paper we will use support vector machines and Logistic Regression to build propensity based response models and evaluate their performance.

Problem Statement

Logistic regression has been the workhorse for developing propensity models in marketing and risk management areas from long time, but for last few years there has been an enormous progress in statistical learning theory and machine learning – providing opportunity to use more robust and less restrictive algorithms to obtain much better results than traditional methods. In the present paper, we will use support vector machines and logistic regression to build propensity based response models and evaluate their performance and also highlight certain positive and negative aspects of the techniques under study.

* Department of Statistics, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.
E-mail:kota.prasad.krishna@gmail.com
** Department of Statistics, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.

Literature Review

Increasing cost of marketing is driving companies to use analytics as corner stone to gain deep understanding of consumer behaviour. Amidst intense competition and dynamic shifts in consumer behaviour, the pressure of improving bottom lines has created enormous emphasis on propensity based response models. Using propensity based response models, one can identify a subset of customers who are likely to respond than others, and also generalise the need for response.

Companies use the knowledge of consumer behaviour to segment, to design marketing strategies, and to measure marketing performance (Schiffman & Kanuk, 1991). The use of SVM is rare in both CRM and customer response model, with exceptions (Viaene *et al*., 2001). Response models have been proven to be highly profitable tool in fine-tuning marketing strategies (Elsner *et al*., 2004). SVMs have great generalisation ability and have strong performance when compared to traditional modeling approaches, but applications of SVMs in marketing are scant (Cui & Curry, 2005). The main purpose of response modeling is to improve future return on investment on marketing (Shin & Cho, 2006). Coussemet & Poel (2006) have used SVM in a newspaper subscription contest, and have proved that SVM have good generalisation ability when compared to logistic regression and random forest. Lately, companies are increasingly deluged by data and sophisticated data mining techniques are available to marketers (Ngai *et al*., 2009).

Support Vector Machines

The Support Vector Machines (SVMs) are supervised learning models with associated learning algorithms that are used for pattern recognition, classification and regression problems. Support Vector Machine (SVM) was introduced by Boser, Guyon & Vapnik in 1992. In 1995, soft margin classifier was introduced by Cortes & Vapnik and the algorithm was extended to problem of regression by Vapnik. Support Vector Machine (SVM) is generalisation of maximal margin classifier.

Maximal Margin Classifier

The maximal margin classifier is defined as the separating hyper plane for which the margin is largest-that is, it is the hyper plane that has the farthest minimum distance to the training observations 1.

Consider class of training observations $x_1 \ldots \ldots, x_n \in R^p$ and the respective associated class labels $y_1 \ldots y_n \in \{-1, 1\}$. The maximal hyperplane is defined as the solution to the optimisation problem:

$$maximize\ \beta\beta_0, \ldots. \beta_{p||\beta|| = 1}$$

Subject to $y_i (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M \forall\ i = 1, \ldots, n$

$$y_i (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M$$

The constraint ensures that each incoming observation will be on the correct side of the hyper plane at least at a distance M from the hyper plane and is called the margin. In above optimisation problem one tries to choose β_o, β_1 ... … … β_p to maximise the distance M.

Soft Margin Classifier

In many real world problems constructing linear separable classifiers is not always possible implying that maximum margin classifiers are no longer valid. In 1995, Corinna, Cortes & Vapnik suggested a modified maximum marginal classifier, in which a new classifier is achieved by relaxing the constraints a little to accommodate small amount of misclassification. The generalisation of the maximal margin classifier to the non-separable case is known as the support vector classifier. The soft margin classifier is defined as the solution to the optimisation problem:

$$maximize\ \beta_{0,\ldots}\beta p, \in 1 \cdots \in n\ ||\beta|| = 1\ M$$

Subject to $y_i (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M (i - \in_i)$ $\forall\ i = 1, \ldots, n$

$$\text{where } \xi i \geq 0 \text{ and } \sum_{i=1}^{n} \xi i < C$$

Where C is a nonnegative tuning parameter, M is the margin and one seeks to maximise the margin as much as possible and ξ_i are slack variables which measures of misclassification of the data x_1. If $\xi_i > 0$ then the i^{th} observation is on the wrong side of the margin, and we say that the i^{th} observation has violated the margin. If $\xi_i > 1$ then we conclude that i^{th} observation is on the wrong side of the hyper plane.

Support Vector Machine (SVM) representation

Given a training set of instance-label pairs $(X_i. y_i)$, i = 1, … l where $X_i \in R^n$ and {1, – 1}, let us assume that patterns with $y_i = 1$ belong to class 1 while with $y_i = -1$ belong to class 2. Then training support vector machines (SVM) require the solution for the following optimisation problem:

$$min_{w,b,\xi} \frac{1}{2} W^T W + C \sum_{i=1}^{l} \xi_i$$

$$\text{Subject to } y_i \left(W^T \phi (X_i) + b \right) \geq 1 - \xi_i$$

$$\xi_i \geq 0, i = 1, \ldots l$$

The above optimisation problem is most general Support Vector Machine (SVM) formulation allowing both non-separable and non-linear cases. The ξ_i are slack variables which measure misclassification of the data and C > 0 is the penalty parameter of the error term. In the above optimisation problem the training vectors X_i are mapped into a higher dimensional space by the function implicitly by employing kernel functions thus, Support Vector Machine(SVM) tries to find a linear separating hype plane with the maximal margin in this higher dimensional space.

Kernel Trick and Kernel Functions

In many real world problems finding a linearly separable hyper plane is not possible, to accommodate non-linearity kernels are used, and the input data are non-linearly mapped into high dimensional space. Consider a vector x in the input space can be represented as $\phi(x)$ in the higher dimensional space H, the mapping of data into higher dimensional space makes it possible to define a similarity measure on the basis of the dot product. If there is a kernel function K such that

$$K(x_1 x_2) = \phi(x_1). \phi(x_2)$$

then mapping is provided by

$$\langle x_1 x_2 \rangle \leftarrow K(x_1 x_2) = \langle \phi(x_1) . \phi(x_2) \rangle$$

Thus, if a kernel function K can be constructed, a classifier can be trained and used in the higher dimensional space without knowing the explicit functional form of mapping. In simple, the kernel trick enables one to find linearly separable hyper plane in feature space for the underlying training data, provided the underlying training data is not linearly separable in input space.A kernel that can be used to construct a SVM must satisfy Mercers condition. The kernel function plays a pivotal role in training SVM and its performance and is based on reproducing Kernel Hilbert Spaces.

Table 1 shows the list of little important kernel function used in practice.

Features Scaling

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalisation. Scaling of input features will help in overcoming the numerical difficulties during training SVMs, since kernel values

Table 1: Important Kernel Function

Linear kernel	It is the simplest kernel function and is equivalent to principal component analysis. It is the inner product of input features plus an optional constant c	$k(x, y) = x^T y + c$
Polynomial kernel	Polynomial kernel are most popular method for non-linear modeling and are well suited where all the training data is normalized	$k(x, y) = (\alpha x^T y + c)d$ α is the slope c is constant and d is polynomial degree
Gaussian kernel	The Gaussian kernel is an example of radial base kernel. The adjustable parameter sigma plays an pivotal role in performance of SVM and should be carefully fine-tuned,	$K(x, y) = exp\left(-\frac{\| x - y \|^2}{2\sigma^2}\right)$
Exponential Kernel	It is also a radial base kernel; the exponential kernel is closely related to the Gaussian kernel, with only the square of the norm left out. The exponential kernel produces a piecewise linear solution and	$(x, y) = exp\left(-\frac{\| x - y \|}{2\sigma^2}\right)$

depend on the dot product of the input feature vectors, large feature values might cause some problems. Thus the input features are scaled between [-1 +1] or [0 1].

Logistic Regression

Consider the following simple linear regression setting with 'r' predictor and binary response variable

$$y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_r x_r + \in_i, i = 1,2, \ldots n$$

Where y_i is the binary response variable, $\in_i \sim N\left(0, \sigma_\in^2\right)$, and are independent.

Let P_i denote the probability that $y_i = 1$ and $x_i = x$

$$P_i = P(Y_i = 1 \mid X_i = X) = \frac{1}{(1 + e^{-z})}$$

Where $Z = \beta_0 + \beta_1 x_i + \ldots + \beta_r x_r$

Or

$$\text{Logit}(p)\ \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \beta_r x_r$$

The above equitation is called logistic regression: a statistical method in which we model the logit (p) in terms of explanatory variables that are available to modeler. It is non-linear in the parameters β_0, β_1,…… β_r. The response probabilities are modeled by logistic distribution and estimating the parameters of the model constitutes fitting a logistic regression.

Performance Evaluation Measures

The following are various methods for assessing the discriminating ability of the trained model and in-time validation dataset.

Confusion matrix

A confusion matrix (also known as an error matrix) is appropriate when predicting a categorical target; confusion matrix helps one to evaluate the quality of the output of the classifier

Figure 1: Confusion Matrix

		Predicted	
		No	Yes
Actual	No	a	b
	Yes	c	d

Accuracy Ratio

It shows the proportion of the total number of predictions that were correctly classified.

$$AR = \frac{(a + d)}{(a + b + c + d)}$$

Precision

It is the proportion of the predicted positive cases that were correctly classified.

$$P = \frac{d}{b + d}$$

Kolmogorov-Smirnov (KS)

This measures the maximum vertical separation (deviation) between the cumulative distributions of goods and bads and is defined as follows

$$KS = MAX\left|F_G^{(s)} - F_B^{(s)}\right|$$

The higher the KS value the better is the models ability for separation.

Lift Curve

Lift is a measure of effectiveness of a predictive model and it is defined as the ratio between the results obtained with and with-out use of the predictive model. The lift curve will help analyse the amount of true responders discriminated in each subset. This is extremely helpful for any marketing team for making optimum decisions.

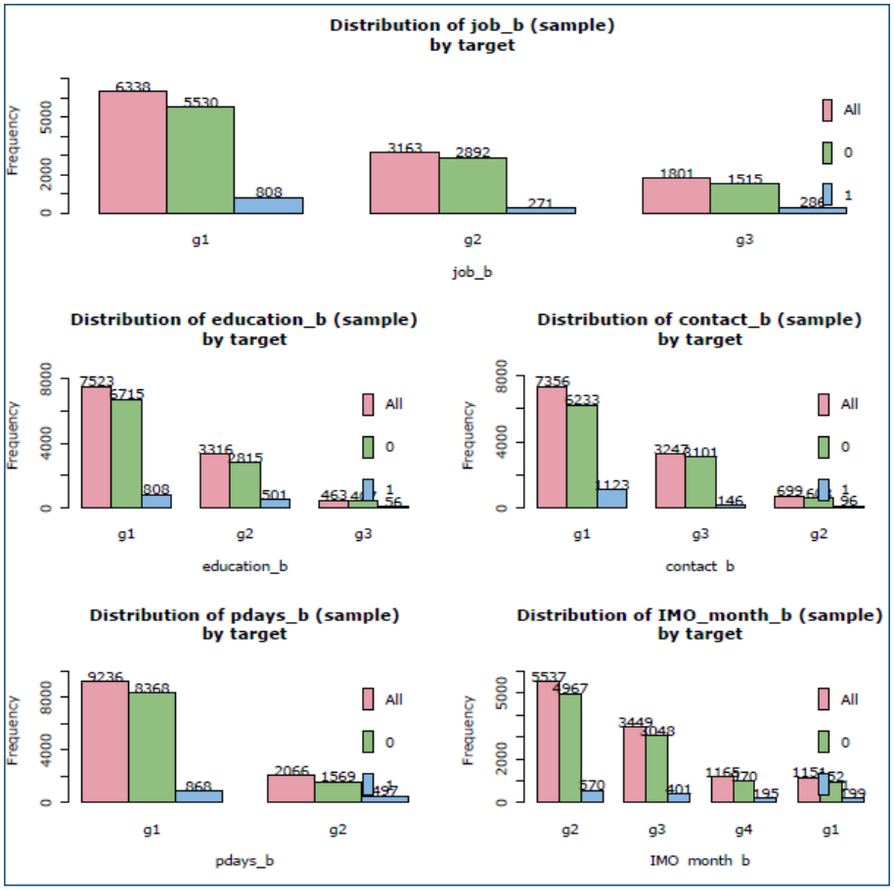
Data Description

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Dataset Preparation

The dataset consist of 45,211 instances. We have used random sampling to construct a dataset of 22,605

Figrure 2: Distribution of Variables



instances.This dataset was used for training and validating logistic regression and support vector machines.

Data Cleaning

As a part of data cleaning exercise, all the prospective variables are subjected to univariate and bivariate analysis; the missing values for numeric variables are imputed with the median and in case of the discreet variables the missing values are imputed by mode.

Variable Transformations

All the numerical variables are scaled between [0, 1] using min-max scaling method, the categorical variables are binned into smaller groups based on the response rates.

Basic Statistics

Table 2 summarizes the basic statistics for the numerical variables (raw andscaled).

Logistic Regression vs Support Vector Machine Performance comparisons

Confusion Matrix

The confusion matrix results for the classifier obtained by both models are shown in Figure 3.

The comparison indicates that SVM are marginally better in development and in validation they perform equivalently well with the logistic model

Rank-ordering

The scores obtained by the classifier (Logistic and SVM) are used to rank-order consumers -how likely they are to respond when solicited. The following table provides the

Table 2: Basic Statistics for the Numerical Variables (raw andscaled)

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
age	22605	40.95067	10.62267	925690	18	95
balance	22605	1343	2935	30347863	-3058	102127
day	22605	15.79889	8.28949	357134	1	31
duration	22605	257.4868	255.6189	5820489	1	4918
campaign	22605	2.74059	3.07033	61951	1	63
pdays	22605	40.47184	100.9995	914866	-1	871
previous	22605	0.56912	1.88134	12865	0	55
rand	22605	0.24761	0.14432	5597	2.74E-05	0.5005
target	22605	0.11878	0.32354	2685	0	1
age_scaled	22605	0.29806	0.13796	6738	0	1
balance_scaled	22605	0.08499	0.02664	1921	0.04504	1
duration_scaled	22605	0.05236	0.05198	1184	0.000203	1
previous_scaled	22605	-0.00207	0.00684	-46.7818	-0.2	0

Figrure 3: Confusion Matrix Results for the Classifier

SVM model

		Predicted		
		Yes	No	
Actual	Yes	9799	138	9937
	No	1013	352	1365
		10812	490	

Accuracy	89.82%
Error	10.18%
Recall	25.79%
precision	71.84%

Logistic model

		Predicted		
		Yes	No	
Actual	Yes	9749	188	9937
	No	1068	297	1365
		10817	485	

Accuracy	88.89%
Error	11.11%
Recall	21.76%
precision	61.24%

SVM model

		Predicted		
		Yes	No	
Actual	Yes	9813	170	9983
	No	1044	276	1320
		10857	446	

Accuracy	89.26%
Error	10.74%
Recall	20.91%
precision	61.88%

Logistic model

		Predicted		
		Yes	No	
Actual	Yes	9781	202	9983
	No	1037	283	1320
		10818	485	

Accuracy	89.04%
Error	10.96%
Recall	21.44%
precision	58.35%

rank-ordering ability of the models in train and test data. In training and test the SVM classifier performs better than the logistic classifier.

The maximum KS for SVM occurs in 2nddecile in train and test , while the maximum KS for logistic regression classifier occurs in 3rd decile in train and test.

ROC Curves

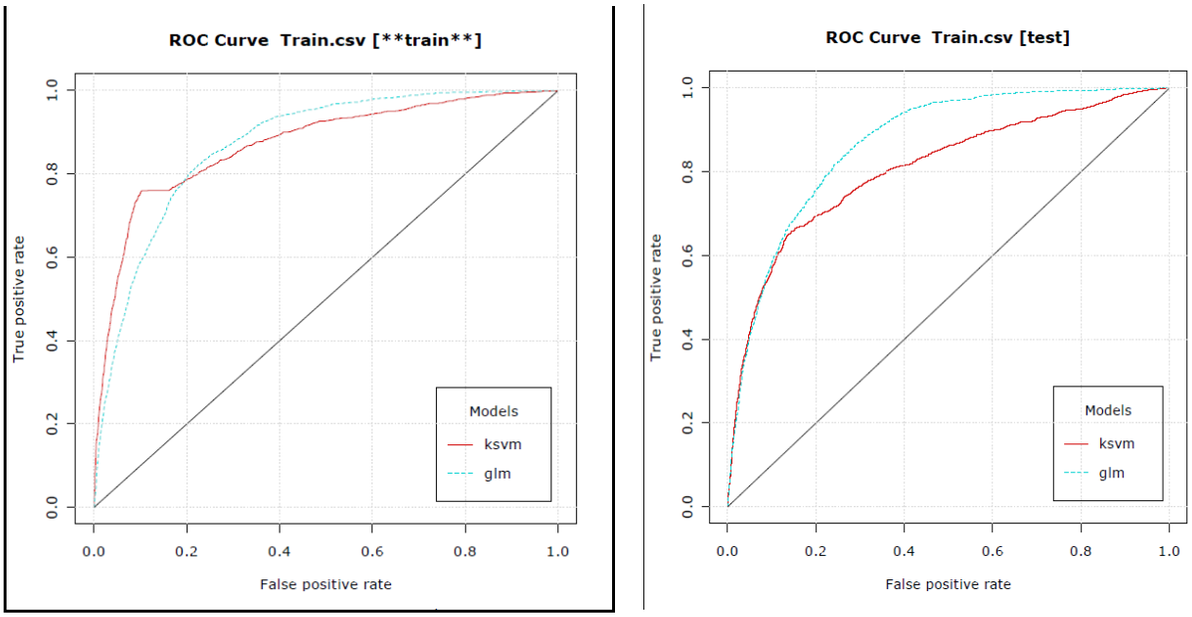
The ROC obtained by the classifier constructed from both the model are populated below, the SVM perform better in marginally better in the training, while in the test it underperform when compared to the logistic regression.

Table 3: Scores Obtained by the Classifier (Logistic and SVM)

	Training							Test						
Decile	Leads	Logistic	SVM	Logistic _ Response Rate	SVM_Response Rate	Logistic_KS _Training	SVM_KS_Training	Leads	Logistic	SVM	Logistic _ Response Rate	SVM_Response Rate	Logistic_KS_Test	SVM_KS_Test
0	**1,130**	575	685	50.88%	60.62%	36.54%	45.70%	**1,130**	565	583	50.00%	51.59%	37.14%	38.69%
1	**1,130**	334	353	29.56%	31.24%	53.00%	63.75%	**1,130**	318	278	28.14%	24.60%	53.10%	51.21%
2	**1,130**	216	63	19.12%	5.58%	59.62%	57.62%	**1,131**	176	85	15.56%	7.52%	56.87%	47.18%
3	**1,131**	101	84	8.93%	7.43%	56.66%	53.24%	**1,130**	129	102	11.42%	9.03%	56.61%	44.60%
4	**1,130**	69	56	6.11%	4.96%	51.04%	46.54%	**1,130**	74	63	6.55%	5.58%	51.64%	38.69%
5	**1,130**	30	35	2.65%	3.10%	42.16%	38.08%	**1,131**	24	58	2.12%	5.13%	42.37%	32.34%
6	**1,131**	20	25	1.77%	2.21%	32.45%	28.78%	**1,130**	19	44	1.68%	3.89%	32.68%	24.79%
7	**1,130**	14	32	1.24%	2.83%	22.24%	20.08%	**1,131**	8	38	0.71%	3.36%	22.04%	16.72%
8	**1,130**	3	24	0.27%	2.12%	11.12%	10.71%	**1,130**	4	43	0.35%	3.81%	11.06%	9.09%
9	**1,130**	3	8	0.27%	0.71%	0.00%	0.00%	**1,130**	3	26	0.27%	2.30%	0.00%	0.00%
Total	**11,302**	1,365	1,365	12.08%	12.08%			**11,303**	1,320	1,320	11.68%	11.68%		

Figrure 4: ROC Curves

ROC Curve		
	SVM	Logistic
Train	87.15%	87.14%
Test	80.76%	86.87%



Lift Charts

The lift obtained by the classifier constructed from both the model are populated below, the SVM perform better in better in the training, while on the test it is performs equivalent when compared to the logistic regression.

Propensity Profile

The output from the SVM will provide no means to profile the prospective lead list, on the other hand theoutput from the logistic model can be use to create customer profiles as shown in Table 4.

Figrure 5: Lift Charts

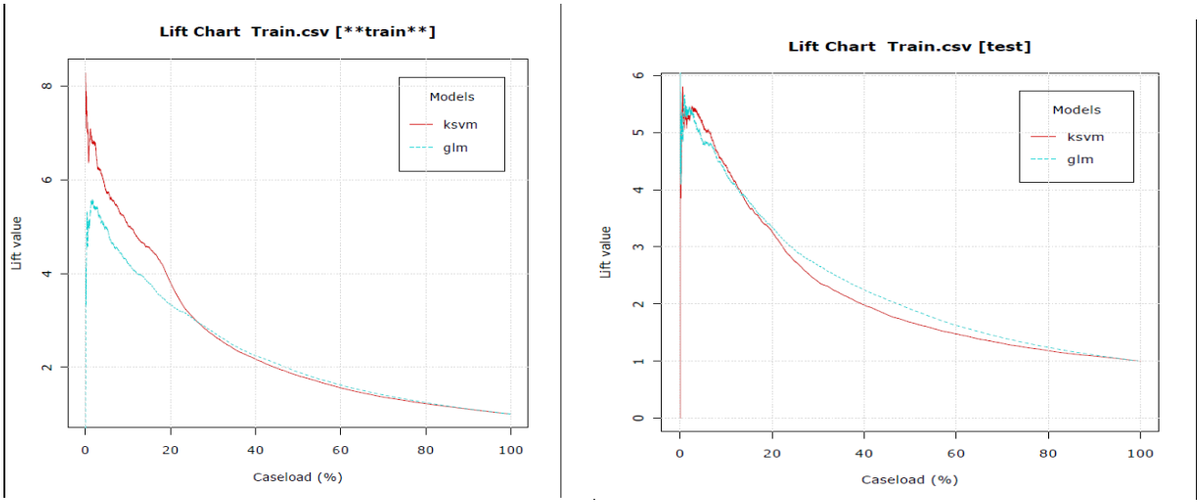


Table 4: Propensity Profile

Bins	Bin's Description	Estimate	Std. Error	z value	Pr(>|z|)	Sig
(Intercept)		-3.47017	0.17144	-20.241	< 2e-16	***
marital	01: Married	-0.2561	0.10625	-2.41	0.015942	*
	02: Single	0.06415	0.11195	0.573	0.566659	
	03: divorced*	0				
default	01: Yes	-0.31495	0.34155	-0.922	0.356471	
	02: No*	0				
loan	01: Yes	-0.62362	0.11023	-5.657	1.54E-08	***
	02: No*	0				
job	01: 'admin.','management','self-employed','technician', 'unemployed','unknown'	0				
	02: 'blue-collar','entrepreneur','hou	-0.34072	0.09203	-3.702	0.000214	***
	03: Other	0.32656	0.09044	3.611	0.000305	***
education	01: 'primary', 'secondary'	0				
	02: 'tertiary'	0.23425	0.0775	3.022	0.002508	**
	03: 'unknown'	0.07789	0.17258	0.451	0.651752	
contact	01:'cellular'	0				
	02:'telephone'	-0.11832	0.1371	-0.863	0.388134	
	03:''unknown'	-1.27635	0.12128	-10.524	< 2e-16	***
pdays	01: Not contacted earlier	0				
	02: Contacted	0.89447	0.09551	9.365	< 2e-16	***
month	01: Jan,Feb, Mar	0				
	02: Apr, May, Jun	-0.10571	0.10622	-0.995	0.319628	
	03: Jul, Aug,Sep	-0.27022	0.11022	-2.452	0.014223	*
	04:Oct, Nov, Dec	-0.01262	0.12531	-0.101	0.919761	
balance_scaled		4.08877	1.09629	3.73	0.000192	***
duration_scaled		20.27482	0.59724	33.947	< 2e-16	***
previous_scaled		-10.41849	4.92502	-2.115	0.034394	*

The prospects who are Single, not defaulted earlier, have no previous loans, having tertiary education, contacted previously on mobiles are more likely to respond to campaigns

SVMs vs Logistic Regression: Pros & Cons

The current study indicates that SVMs perform better than logistic regression on the performance evaluation parameters that we have used to evaluate the classifiers, but still logistic regression continues to be work-horse in response modeling due to the following reasons

Pros

- The SVMs have good generalisation in both in-samples, hold-out and out-of-sample by choosing appropriate parameters.
- The concept of kernels encompasses non-linear transformations, so no prior assumption is made about the functional form of the transformation.
- SVMs are robust to outliers.

Cons

- Unlike logistic regression - SVMs is the lack of transparency of results.
- The choice of kernel is another shortcoming
- Unlike logistic regression - SVMs has high algorithmic complexity and requires extensive memory requirements.

Conclusion

The current study indicates that SVMs perform better than logistic regression on the performance evaluation parameters that we have used to evaluate the classifiers. But lack of transparency of results, extensive memory requirements, issues in implementation in production system for regular scoring, no comparable standards to monitor SVMs on an ongoing basis to track performance make's logistic regression models, the preeminent choice due to extensive theory around regression framework, ease of understanding and implementation, sensible results along with actionable insights could be used to identify generic and niche segments that enable the marketing teams to develop more tailored campaigns.

References

Athanassopoulos, A. D. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. *Journal of Business Research,* 47, 191–207.

Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* 2, 121–167.

Bennett, K. P., & Campbell, C. (2000). Support vector machines: hype or hallelujah? SIGKDD Explor. Newsl. 2, 1–13

Cui, D., & Curry, D. (2005). Prediction in marketing using the support vector machine. *Marketing Science,* 24, 595–615.

Cui, G., Wong, M. L., & Lui, H. K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science,* 52(4), 597–612.

Chang, C. C., & Lin, C. J. (2001). LIBSVM: A library for support vector machines, Software Retrieved from http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Elsner, R., Krafft, M., & Huchzermeier, A. (2004). Optimizing Rhenania's Direct Marketing Business through dynamic Multilevel Modeling (DMLM) in a Multicatalog-Brand Environment. *Marketing Science*, 23(2) 192-206.

Hosmer, D.W., & Lemeshow, S. (1989). *Applied logistic regression,* New York: John Wiley & Sons, Inc

Moro, S., Laureano, R., & Cortez, P. (2011).Using data mining for bank direct marketing: An application of the CRISP-DM methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

Ngai, E., Xiu, L., & Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Syst Appl, 36, 2592–2602. doi:10.1016/j.eswa.2008.02.021

Kim, D., Lee, H., & Cho, S. (2008). Response modelling with support vector regression. *Expert Systems with Applications,* 34, 1102–1108.

Hsu, C., Chang, C., & Lin, C. (2010). *A practical guide to support vector classification*. Department of Computer Science and Information Engineering, National Taiwan University.

Viaene, S., Baesens, B., Van Gestel, T., Suykens, J., Van den Poel, D., Vanthienen, J., De Moor, B., & Dedene, G. (2001). Knowledge discovery in a direct marketing case using least squares support vector machines. *International Journal of Intelligent System*, 16, 1023–1036

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*, Springer, New York.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). The Elements of Statistical Learning. Springer, 2001.

Schhiffman J. B., & Kanuk, L. L. (1997). Consumer Behavior published by Prentice Hall Sixth edition, 446.

Shin, H. J., & Cho, S. (2006). Response modeling with support vector machine. *Expert Systems with Application,* 30(4), 746-760.

Williams, G. J. (2011). Data mining with Rattle and R: The art of excavating data for knowledge discovery, Use R!, Springer.

Article can be accessed online at http://www.publishingindia.com

A Study on the Analytics Tool Used in Decision Making at Small and Medium Enterprise

Vivek N. Bhatt*

Abstract

The article focuses on the study of prevailing decision making styles of Small Scale Industrial (SSI) Units. It presents data collected from 200 SSI units from Bhavnagar – a coastal city of Gujarat, India. The objective of writing the article is to depict heuristic decision patterns of small and medium enterprises, and the rare use of analytical or statistical business intelligence tools in decision making processes. It would be interesting to study the design of decision taken on routine basis in small units, poorly equipped with technology and technical know-how. The paper is descriptive in terms, and lays a lucid picture of present decision making processes.

Keywords: Decision Making, Smallscale Units, Analytical Tools, Statistical Tools, Business Intelligence

Introduction

In a very simple and primary endeavour to study decision making process adopted by small business units in Bhavnagar town in South Gujarat, India, data were collected for basic business decision areas, use of software or analytical tools was done for these decisions, and the efficiency of decisions was considered.

According to the Ministry of Micro, Small and Medium Enterprises, recent ceilings on investment for enterprises to be classified as micro, small and medium enterprises are as shown in Table 1.

For manufacturing units, investment limit is for plant and machinery, and for service unit it is for equipment.

The data for the research are collected from 200 different small and medium scale entrepreneurs or decision makers. It was proposed to be a convenient sample of the entrepreneurs, but in the due course of data collection process, it was realised that to collect data from a large mass of busy entrepreneurs, who are not using computers to a noticeable extent, and who are especially not looking up to any customized computerised solution, is a challenging task.

Table 1 : SME Definition

Scale	*Manufacturing*	*Service*
Micro	Rs. 25 Lakh	Rs. 10 Lakh
Small	Rs. 5 Crore	Rs. 2 Crore
Medium	Rs. 10 Crore	Rs. 5 Crore

Literature Review

In a study (Paul Foyce, 2003), it is emphasised strategic decisions, keeping in pace with rapid changes and growth of the unit in mind structure decision processes are high in demand. Speed of change and demand of competency enhancement areso resilient that formalised decision making seems indispensable which further induces the need for software applications of business intelligence systems. Since globalisation and liberalisation in India, requirement for formalised decision models and decision making systems has surfaced.

In another study (Simon Mosey, 2002), the focus was on growth and innovation. Decision for growth includes innovation in product design. For the decisions about innovation in any aspect, enterprise requires market and competitor analysis in strategic planning.

In an article by Hedgebeth (2007), describes origin of business intelligence (BI), BI applications, and BI value in

* Associate Faculty with Entrepreneurship Development Institute, Ahmadabad, Gujarat, India.
E-mail: shree.vivek@gmail.com

decision making. It clearly concludes informed decisions greatly depend upon accuracy of BI applications. Efficiency of decisions leads to minimising cost, market forecast and analysis, which further is an outcome of proper implementation and use of BI applications and analytical methods.

One more study (Martin Aruldoss, 2014), which is more comprehensive in nature, discusses composition and development of BI applications, and points out the fact that most of the researches about BI tools talks, in majority, about development of BI applications, than about the usage of these applications in real business decision-making processes.

Methodology

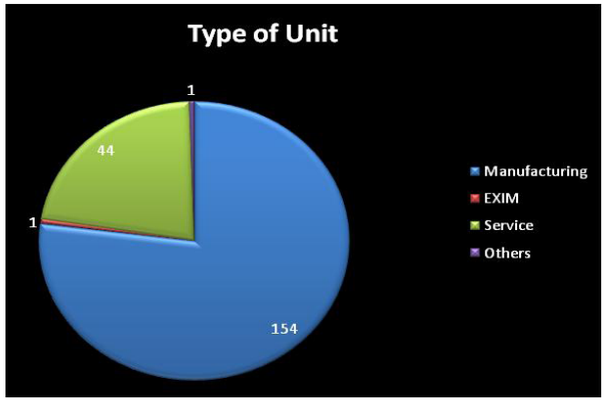
As mentioned above, the study is more descriptive in nature, with a view to display a picture of present decision making pattern of small enterprises. Since the objective was to collect data about some basic decision areas and their impact in a small business, study proceeded with data collected from 200 entrepreneurs.

In the method adopted to collect data and to regularise it, a questionnaire was designed, but the details were collected through informal talks, indirect approach,online form filling. However, a structured format of personal interview was not followed. The details collected were arranged in the strict format of questionnaire.

Data Collection and Analysis

The classification of the details collected is as follows.

Figure 1: Type of Unit



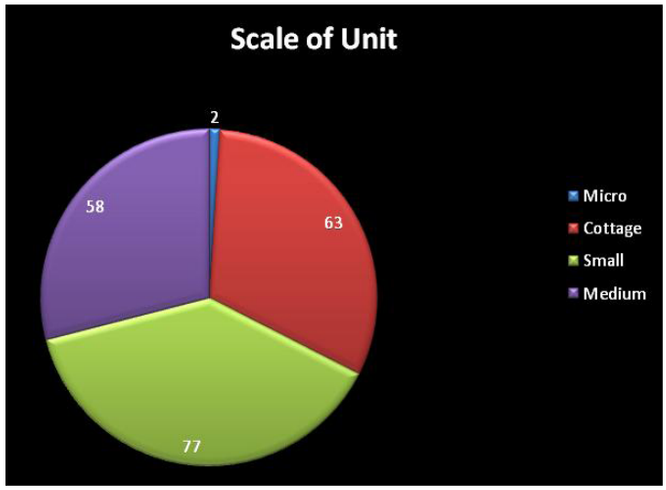
Types of Units

The first criterion selected is the type of unit, since decision pattern changes a lot with the change of type. As displayed in Figure 1, out of 200 units 44 units were manufacturing, 155 units were service providing, 1 unit is EXIM unit.

Scale of Units

The second criterion is the scale of units, because financial and technical feasibility largely depend upon the scale and financial capacity of the unit. As it can be seen in Figure 2, out of 200 units interacted with, 38% are small and 29% units are medium scale units.

Figure 2: Scale of Unit



Market Share

Past decisions can be assessed with different parameters, one of which is market share. For 200 units contacted, the scenario about market share is displayed in Figure 3. 61% units claim to hold 20% to 40% market share in their respective area.

Utilisation of Established Capacity

The best decisions majorly lead to optimum allocation of resources, and eventually to maximum utilisation of established capacity of a firm. As seen in Figure 4, 160 units (80% units) utilise50% to 70% established capacity

of their firm, which is a signal of poor resource allocation and so of weak decision making.

Figure 3: Market Share

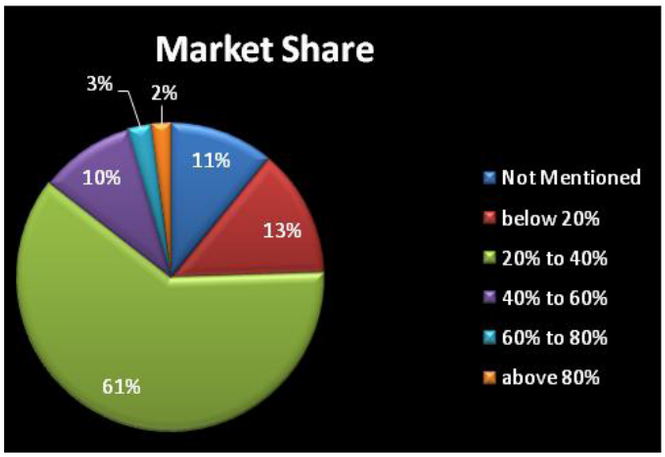
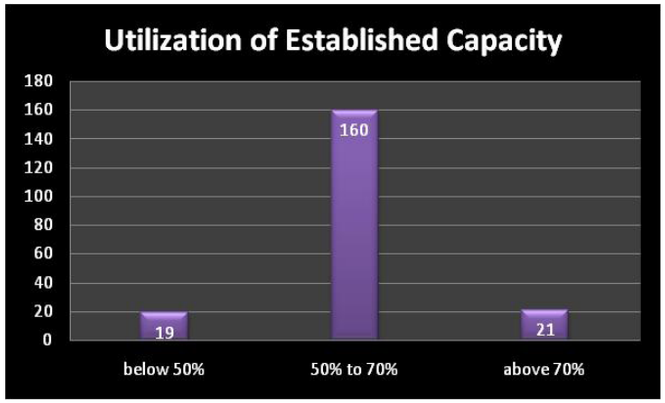


Figure 4: Utilisation of Established Capacity



Failure of Production Schedule

Production scheduling is a routine decision. Such decisions are normally taken by gross experience. This results in failure of production schedule in 10% to 30% cases, for 133 units (66.5% units). More shocking figure is that of 7 units having failure rate from 30% to 50%. Although the number is not too big in a sample of 200 units, it might count remarkable in the complete assessment if conducted.

Demand Forecasting Method Adopted

Planning and scheduling greatly depends on demand forecasting efficiency. Demand forecasting can be done with different methods. We have collected data, and found the units using methods like agency services, intuition/ experience, statistical methods, software or other methods. It is noticeable here that 168 units (84% units) forecast demand only using Intuition / Experience, which is quite disappointing.

Figure 5: Failure of Production Schedule

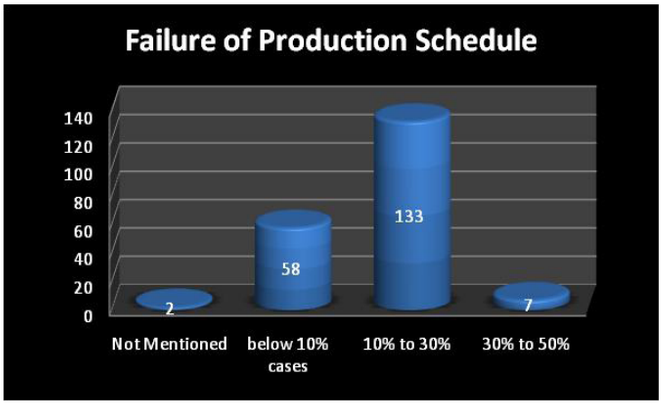
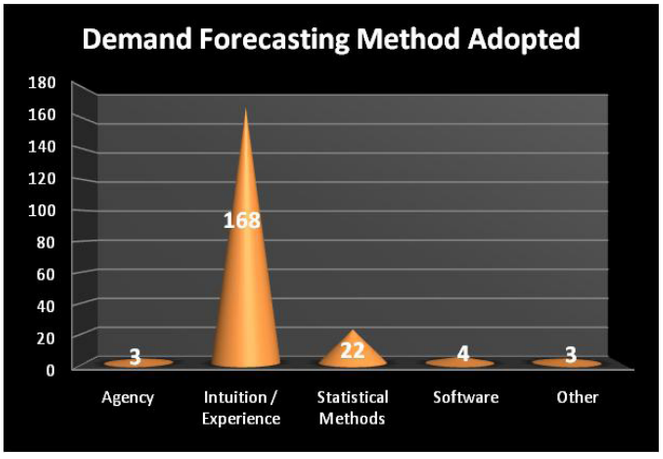


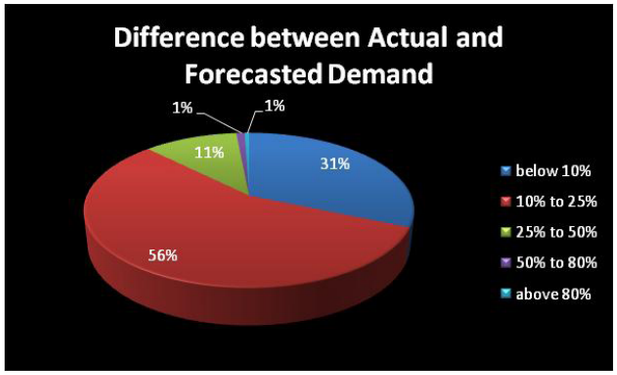
Figure 6: Demand Forecasting Method Adopted



Difference between Actual and Forecasted Demand

Efficiency of the method employed for demand forecasting is assessed with the help of deviation observed of actual demand from estimated/ forecasted demand. Apparently, there are 56% units (112 units) having deviation rate of 10% to 25%.

Figure 7: Difference between Actual and Forecasted Demand



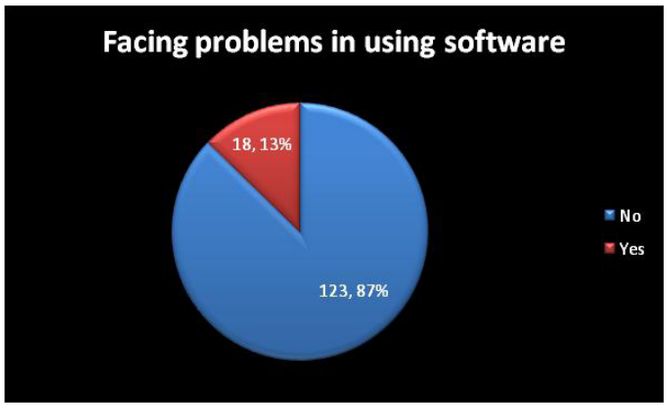
Software/ Application Used

Around 29.5% units (59 units) do not use any software. Units using computer software comprise of the units using Microsoft Word (62 units), Microsoft Excel (39 units), Microsoft Access (2 units), Tally Accounting software (15 units), Word and Excel (12 units), customised software (3 units – these units also have got the applications designed mainly for inventory and accounting), and Lotus (8 units). Now this discomforting number of 59 units explains the whole case. When in a sample of 200 units 59 units are not using any software, for the whole population this number may elevate to a greater extent.

Facing Problems in using Software

Out of 141 units using some software for their routine activities, detail storage or analysis, 123 are happy with the basic applications they use, 18 decision-makers face problems with the applications they utilise.

Figure 9: If Facing Problem in using Software



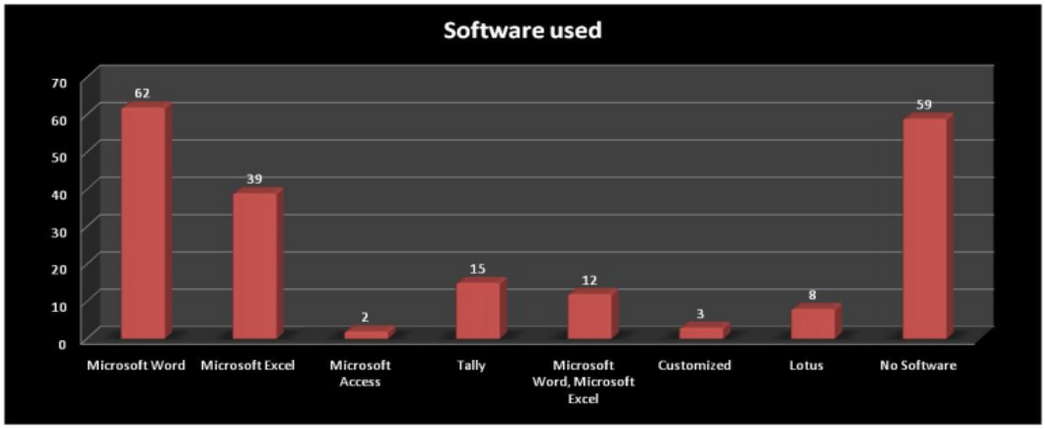
Problems in using Software

Looking into this point, it amazes us with a fact that, majority of problems comes as technical infeasibility to use a computerised tool rather than financial infeasibility or non-affordability. Although, here software application in question are very primary applications like Word, Excel, Access and Tally and one or two customised applications, it does not lose significance that prime concern is technical comfort over financial one.

Do You Opt for Customised Software?

This is a question which leads to the future scope for Business Intelligence tools to play role in decision

Figure 8: Software Used



making of small businesses, the way they contribute in large business processes. At juncture too subdued outcome is 144 units are quite indifferent towards the need of a customised application and so for the need of computerisation in the day-to-day business process for decision making. We can consider lack of skilled employees or the technical infeasibility in the root of this pessimistic retort. Here it exhibits a clear vacuum in the area of Business Intelligence tool/ application designed principally for SMEs.

Figure 10: Problems in using Software

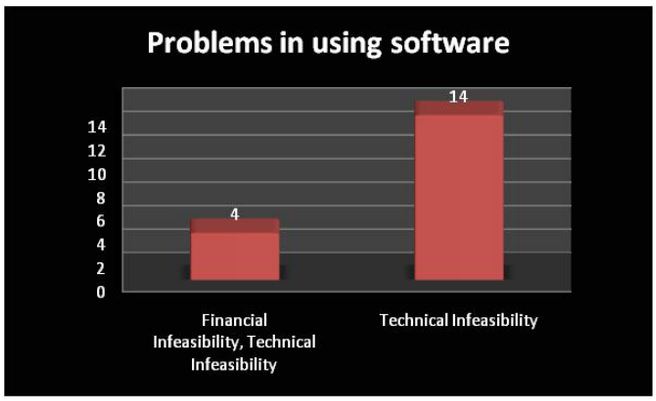
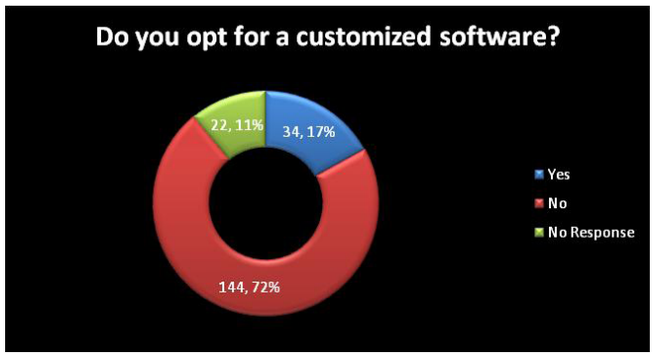


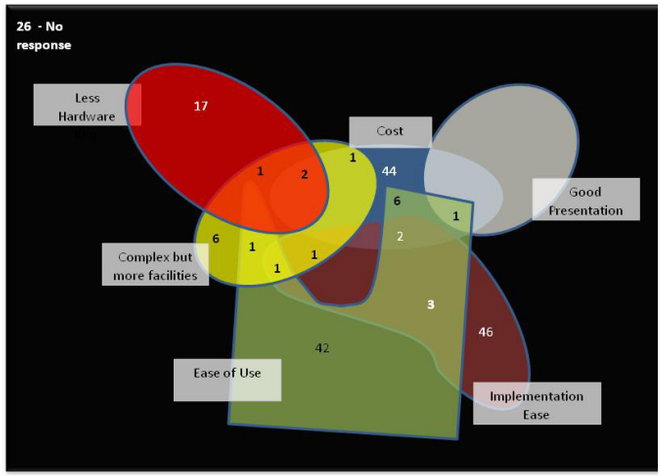
Figure 11: Do You Opt for Customised Software?



Features Expected in Customised Solution

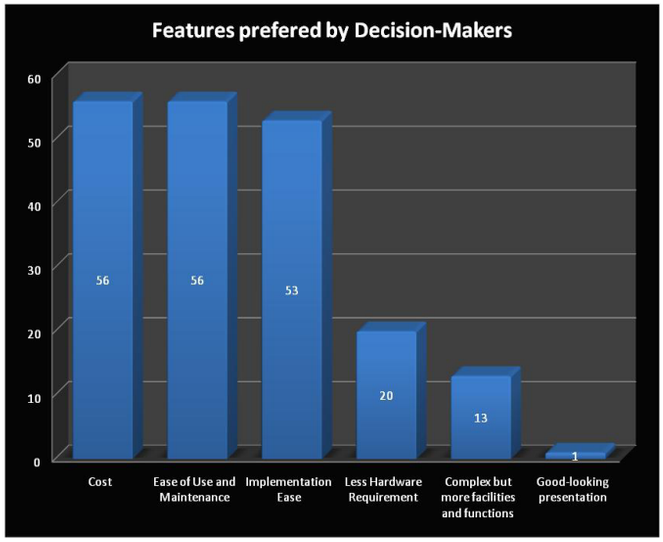
When asked about the features decision-makers seek in customised software solution, their responses were woven around six main factors i.e. Cost, Good Presentation, Less Hardware Requirement, Complex but more facilities and functions, Ease of use and maintenance, and Implementation Ease. The combinations they suggested are displayed in Figure 12.

Figure 12: Features Expected in Customised Solution



Independently treating the features, we have the preferences as shown in Figure 13.

Figure 13: Features Preferred



Conclusion

Preliminary qualitative analysis was preferred. The study leads to two major ideas; first there is a need for systematic data storage, analysis and use in decision making – looking, mainly, at low capacity utilisation and noticeable failure rate of demand estimates; and second, (29% of the units do not use any tool or software application) as low as three units out of 200,i.e.one and a half percent of the total 200 units studied, understand

the need of a customised software and use for decision making. It could be,therefore, concluded that cost and ease of use and maintenance are the characteristics most preferred in a customised software solution. Means financial and technical plus technological feasibility is the main concern for the entrepreneurs. Eventually it narrows down to the path leading to a tailor-made software solution which satisfies the primary requirements and helps an individual to arrive at a decision following an easy alternative selection process.

References

Hedgebeth, D. (2007). Data-driven decision making for the enterprise: an overview of business intelligence applications. *The Journal of Information And Knowledge Management Systems*, 414-420.

J. A. Cannon-Bowers, E. S. (1996). Establishing the boundaries of a paradigm for decision-making research. *Human Factors*, 38, 193-205.

Martin Aruldoss, M. L. (2014). A survey on recent research in business intelligence. *Journal of Enterprise Information Management*, 831-866.

Paul Foyce, A. W. (2003). Managing for growth: decision making, planning, and making changes. *Journal of Small Business and Enterprise Development*, 144-151.

Simon Mosey, J. N. (2002). Innovation decision making in British manufacturing SMEs. *Integrated Manufacturing Systems*, 176-183.

Comparison of Logistic Regression and Artificial Neural Network based Bankruptcy Prediction Models

Easwaran Iyer *, Vinod Kumar Murti**

Abstract

Logistic Regression is one of the popular techniques used for bankruptcy prediction and its popularity is attributed due to its robust nature in terms of data characteristics. Recent developments have explored Artificial Neural Networks for bankruptcy prediction. In this study, a paired sample of 174 cases of Indian listed manufacturing companies have been used for building bankruptcy prediction models based on Logistic Regression and Artificial Neural Networks. The time period of study was year 2000 through year 2009. The classification accuracies have been compared for built models and for hold-out sample of 44 paired cases. In analysis and hold-out samples, both the models have shown appreciable classification results, three years prior to bankruptcy. Thus, both the models can be used (by banks, SEBI etc.) for bankruptcy prediction in Indian Context; however, Artificial Neural Network has shown marginal supremacy over Logistic Regression.

Keywords: Bankruptcy, Logistic Regression, Artificial Neural Network, Classification Accuracy.

Introduction

Many researchers have built bankruptcy prediction models and tested in different countries. Among them, the most popular has been the model developed by Edward Altman (USA) in 1968 in which Multiple Discriminant Analysis was used. Next to Multiple Discriminant Analysis, Logistic Regression has been used extensively. Several other techniques like Probit Regression, Data Envelopment Technique, Time Series CUSUM Methodology, Cox Regression, Decision Tree Analysis, Simple Hazard Model, Black-Scholes Option-Pricing Model, Simple Fuzzy Logic, Artificial Neural Networks and Genetic Programmed Decision Trees were also used for exploring better discriminating models for bankruptcy prediction. We have found that very few researchers have conducted researches with Indian data. Moreover, most of the researches have been around Altman's model (1968) with Multiple Discriminant Analysis. There is a distinct gap between the researches done abroad and researches done in India with regard to application of discriminating techniques.

In this study, we compared one of the most popular techniques used for bankruptcy prediction, that is *Logistic Regression,* with a comparatively newer one that is *Artificial Neural Network* with Indian data. The independent variables were considered the same as those considered by Ohlson (1980) due to their worldwide acceptability. To the best of our knowledge, this is the first of its kind, which compares both techniques with Indian Data for bankruptcy prediction. We present a quick relevant *Literature Review* under section 2 followed by *Data and Methodology* under section 3 followed by *Analysis and Results* under section 4 followed by *Conclusions* under section 5 and *References* are mentioned at last under section 6.

Literature Review

In the paper published by Ohlson (1980), he had mentioned two unpublished papers by White and Turnbull (1975a; 1975b) and by Santomero and Vinso (1977) which were the first studies that had logically and systematically developed probabilistic estimate of failure. Ohlson had also used the methodology of maximum likelihood so called logit model or logistic regression. He had used 58 bankrupt and 2058 non-bankrupt companies in his sample

* Director and Dean – Commerce and Management, Jain University, Bangalore, Karnataka, India. E-mail: easwaran.iyer@jainuniversity.ac.in

** Academic Head - i Nurture Education Solutions Private Limited, Bangalore & Faculty at Commerce and Management, Jain University, Bangalore, Karnataka, India. E-mail: vinod@inurture.co.in

pertaining to the time period of 7 years from 1970 to 1976. He believed that if financial statements are disclosed after the declaration or filling of bankruptcy, there are very high chances that the firm might back-cast the results. Under these circumstances, the financial results were bound to reflect bias and did not reflect true position. For the sake of not including such companies in his study, Ohlson had referred to Accountant's Reports. Ohlson had considered only three reporting periods (years) prior to bankruptcy. He had further mentioned that many important problems pertaining to the development of data for bankrupt firms had not been addressed in the literature. Ohlson had mentioned the strict assumptions of Multiple Discriminant Analysis which demands mainly i) equality of variance-covariance matrices of the predictors among failed and non-failed groups and ii) normally distributed predictors. He had further mentioned that under many circumstances, researcher is interested towards more traditional econometric analysis and test variables for statistical significance where the above-cited assumptions create limitations. Furthermore, he had mentioned that use of dummy predictors displays departure from these assumptions. Ohlson formed the opinion that discriminant scores developed by Multiple Discriminant Analysis have little intuitive interpretation. Ohlson had further stated that the matching of sample on the basis of size and industry is somewhat arbitrary. Logit analysis, on the other hand, is free from assumptions regarding prior probabilities and distributional properties of predictors. He had cited this as a major advantage of using logit analysis. With regard to statistical significance, Ohlson had stated that this could be obtained through asymptotic (large sample) theory. He had used the following predictors: *i) Size: log (total assets/GNP price index). The index assumed a base value of 100 for 1968. ii) Total Liabilities/Total Assets. iii) Working Capital/Total Assets. iv) Current Liabilities/Current Assets. v) OENEG; One if total liabilities exceed total assets, otherwise zero. vi) Net Income/Total Assets. vii) Funds from Operations/Total Liabilities. viii) INTWO; One if net income was negative for the last two years, otherwise zero. And ix) CHIN;* $(NI_t - NI_{t-1})/(Mode\ NI_t - Mode\ NI_{t-1})$. Ohlson had built three models for one year prior to bankruptcy, two years prior to bankruptcy and three years prior to bankruptcy and found 96.12, 95.55 and 92.84 percent correctly classified. Finally, Ohlson had concluded that i) the timing issue with regard to declaration of bankruptcy and disclosure of financial statements were important and should not be ignored and ii) additional predictors, particularly market related predictors should be explored for improving the predictions.

Luther (1994) had compared ANN with Logistic Regression with the help of 104 sample size of US companies. The study period was 1984 through 1989. The neural network was trained using the Genetic Algorithm technique, which iterates towards the optimum solution by looking only at the value of the objective function and not getting trapped in the local minima. Thirteen predictors were selected for model building by ANN and LR. The study concluded that ANN model had significantly higher prediction accuracy than the Logit Model in both the training samples and the hold-out samples at almost all cut-off points. Luther mentioned that that the prediction accuracy was less sensitive to changes in the cut-off point in the model, thus making ANN more robust technique than Logit.

Zhang, Hu, Patuwo, and Indro (1999) had studied matched sample of 220 US firms 1980 through 1991 and explained the link between ANN and traditional Bayesian classification theory. They found that ANN models were significantly better than Logistic Regression models in prediction as well as classification rate estimation. They reported that ANN was robust to sampling variations in overall classification performance.

Javanmard and Saleh- (2009) had used a sample of 80 companies and compared Multiple Discriminant Analysis and Artificial Neural Network. They mentioned that the ANN has been used to solve many financial problems including forecasting financial distress and many researchers using ANN to forecast financial distress have come to the conclusion that the accuracy of ANN is much more effective than the traditional statistical methods. They quoted Cerano-Sinka's work on comparison of MDA & ANN where Cerano-Sinka got forecasting accuracy as 86% and 94% respectively. Javanmard and Saleh had also reported superiority of ANN over MDA in their study.

Lin (2009) had compared MDA, Logit, Probit and ANN models for bankruptcy prediction in Taiwan. He had studied Taiwan public industrial firms for the period 1998-2005. Final models were validated through out-of-the sample data. Lin found Probit models as the best among all in terms of classification accuracies and stability. Lin had mentioned that if the data does not satisfy the assumptions of statistical approach, then the ANN approach would

demonstrate its advantage and achieve higher prediction accuracy.

Wang and Campbell (2010) had studied the application of Ohlson's model on Chinese publicly traded companies during the period 1998-2008. They had mentioned that the Chinese economy was significantly affected by the 2008-09 Global Financial Crisis due to the export oriented nature of the economy. They had reported from the reference of Thurston (2009) that 28 percent increase in overall bankruptcy filling and 29 percent increase in commercial bankruptcy filling in February, 2009 were witnessed vis-à-vis last year. Chinese stock exchange started in 1990; but, the first delisting took place in the year 2001 with the delisting of Shanghai Narcissus Electric Appliances Co., Ltd. In 2006, Chinese GAAP for reporting financial statements was changed to IFRS. Wang and Campbell had collected a sample of 1336 companies from manufacturing and non-manufacturing industry sectors. The Ohlson model gave overall prediction of 95 percent; but, prediction of non-failed companies was very poor.

Data and Methodology

We have explained design of hypothesis, data preparation, brief outline of discriminating techniques used in this study viz., Logistic Regression and Artificial Neural Network and tools and techniques used for analysing classification results in the following section.

Design of Hypothesis

For comparing and judging the best displayed classification accuracies by the models based on Artificial Neural Network and Logistic Regression, the following hypothesis was designed and tested later on:

H_0: *There is* ***no*** *statistically significant difference between the Classification Accuracies for Bankruptcy Predictions displayed by the models based on Artificial Neural Network and Logistic Regression.*

H_a: There is *statistically significant difference between the Classification Accuracies for Bankruptcy Predictions displayed by the models based on Artificial Neural Network and Logistic Regression.*

The Proposed Level of significance was 5%.

Data Preparation

This section explains the sample preparation, validation sample, and independent and dependent variables.

Sample Preparation

For building models for bankruptcy prediction, pairs of bankrupt and non-bankrupt companies were needed. The names of bankrupt companies were taken from the official web site of BIFR (Board of Industrial and Financial Reconstruction). A ten year period was studied in this study. In this ten-year time period, total 2678 companies had filed for bankruptcy. Availability of financial data was one of the major constraints. Out of 2678, only 1152 companies had their presence in **Capitaline** data source. Further, the study was pertaining to only manufacturing and listed companies only, we could find only 827 bankrupt manufacturing companies available in **Capitaline** data source. Out of 827, only 245 bankrupt companies had their financial data available for the past five years prior to bankruptcy. Further, 50 bankrupt companies were found repetitive in the list available on site. These 50 companies were deleted from 245 bankrupt companies. As a practice followed by previous researchers to not select smaller companies in their studies, which seems logical as small companies are more prone to financial distress due to variety of reasons, we also deleted 58 small companies whose Total Assets for the third year prior to bankruptcy was less than INR 30 Crores. We made third year as the reference year prior to bankruptcy for pairing purpose and for the purpose of comparison of Total Assets. This is because that the year bankruptcy was filed is bound to show lowest Total Assets. Third year prior to bankruptcy is supposed to show comparable financial health (of bankrupt and non-bankrupt companies) in terms of Total Assets. On the same note comparatively too big should also not be included in the sample. For the same reason, 2 bankrupt companies were also deleted. Finally, we left with 135 bankrupt companies belonging to only manufacturing and listed category and filed for bankruptcy. These 135 bankrupt companies were attempted for pairing. For meaningful pairing, the following considerations were taken into account. The prospective pair of a bankrupt company should be i) belonging to manufacturing industry ii) listed in any of the stock exchange so that Market Capitalisation can be computed iii) belonging to the same or nearly

same industry classification (of bankrupt company), iv) of almost same size with a plus minus variation of 30% v) free from bankruptcy filing vi) financially healthy and vii) having financial data of last five years prior to bankruptcy. As a result of the above mentioned criteria, we could match or pair only 109 bankrupt companies with non-bankrupt companies. Thus, the sample size became of 218 cases. Six data sets were prepared pertaining to year of bankruptcy, t0 through fifth year prior to bankruptcy, t5.

Selection of Validation Sample

We randomly selected 20% of 218 cases which resulted into 22 paired cases totalling 44 cases for validation purpose. These 44 cases (hold-out sample) were not used for building the model. The build models were validated by this holdout sample. The model building data sets were having 174 cases (87 cases for bankrupt nomenclature as group_1 and 87 cases for non-bankrupt cases nomenclature as group_0 in this study). Six data sets were prepared for validation pertaining to year of bankruptcy, t0 through fifth year prior to bankruptcy, t5.

Independent and Dependent Variables

We have mentioned in the introduction (section 1) that we considered the *independent variables* selected by Ohlson (1980) due to their worldwide acceptability. These are mentioned at the beginning of section 2 where we have mentioned the notable paper of Ohlson (1980). Dependent variables were 0 and 1 for non-bankrupt and bankrupt outcome.

Software Used

We used SPSS Version 20 for model building, validation, building Receiver Operating Characteristic Curves and applying One Sample Kolmogorov Smirnov and Paired Sample t-tests.

Classification Techniques

Two classification techniques were applied on 6 data sets and classification results were compared for judging the efficacy of discriminant techniques. Both the techniques are explained in short as below.

Logistic Regression

Logistic Regression is a specialized form of regression that is formulated to predict and explain (two-group) categorical variable rather than a metric dependent measure. The form of the logistic regression variate is similar to the variate in multiple regressions. The variate represents a single multivariate relationship with regression-like coefficients indicating the relative impact of each predictor variable. Logistic Regression differs from multiple regression, however, in being specifically designed to predict the probability of an event occurring that is the probability of an observation being in the group coded 1. Because the binary dependent variable has only the values 0 and 1, the predicted value (probability) must be bounded to fall within the same range. To define a relationship bounded by 0 and 1, logistic regression uses the *logistic curve* to represent the relationship between independent and dependent variables.

This *logistic curve* ensures i) the predicted values are always between 0 and 1 and ii) the predicted values correspond to the probability of Y (Dependent variable) being 1, in our case, bankruptcy. The logistic regression is first performed with a transformed value of Y, called the *logit function* as shown below:

$$\text{Logit}\,(Y) = \text{In(odds)} = a + k_1x_2 + k_2x_2 + \cdots + k_nx_n \quad (2.1)$$

Where *odds* refer to the odds of Y being equal to 1.

$$\text{Odds} = \frac{\text{Probability}}{1 - \text{Probability}} \quad (2.2)$$

Odds are defined mathematically as:

$$\text{Probability} = \frac{\text{Odds}}{1 + \text{Odds}} \quad (2.3)$$

Odds can be converted into probabilities by the following expression:

The right hand side of Equation (2.3) does not guarantee values between 0 and 1. Hence, exponent of each side is taken as shown below:

$$e^{\text{In(odds)}} = \text{odds} = e^{(a + k1x1 + k2x2 + \cdots + knxn)} \quad (2.4)$$

Now dividing both sides of Equation (2.4) by (1+odds) results into:

$$\frac{\text{Odds}}{(1 + \text{odds})} = \frac{e^{(a + k_1x_1 + k_2x_2 + \cdots + k_nx_n)}}{(1 + e^{(a + k_1x_1 + k_2x_2 + \cdots + k_nx_n)})} \quad (2.5)$$

The right hand side of Equation (2.3) is equitable to right hand side of Equation (2.5). Hence, the equation looks like:

$$\text{Probability} = \frac{e^{(a + k_1x_1 + k_2x_2 + \cdots + k_nx_n)}}{(1 + e^{(a + k_1x_1 + k_2x_2 + \cdots + k_nx_n)})} \qquad (2.6)$$

The Equation (2.6) yields p, probability of belonging to a group (Y=1, bankrupt) rather than the log of the *odds* of the same. SPSS version 20 has the capability of computing the coefficients k_n for the regression shown in Equation (2.1). Thus, the computation of probabilities of belonging to Group_1 is done. The Equation (2.6) can only yield values that are between 0 and 1.

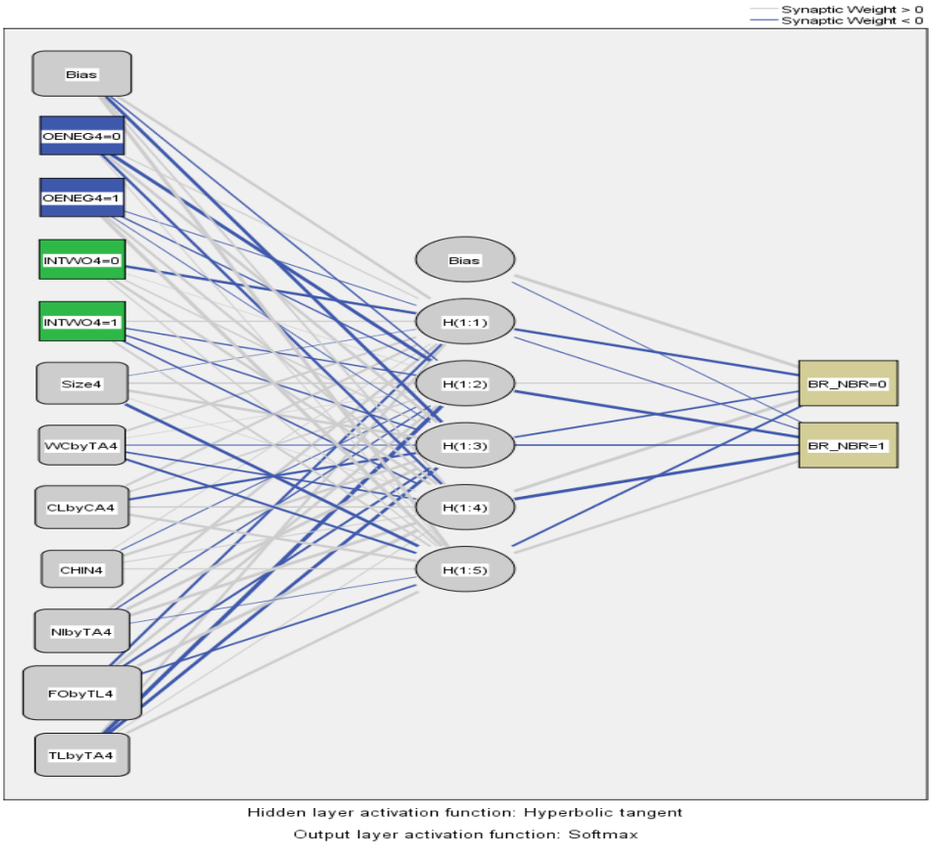
Artificial Neural Network

A neural network's ability to perform computation is based on the hope that we can produce some of the flexibility and power of the human brain by artificial means. Network computation is performed by a dense mesh of computing nodes and connections. They operate collectively and simultaneously on most or all data and inputs. The basic processing elements of neural networks are called artificial neurons, or simply neurons. Neurons perform as summing and nonlinear mapping junctions. They are often organized in layers, and feedback connections both within the layer and toward adjacent layers are allowed. Each connection strength is expressed by a numerical value called a *weight*, which can be modified. A typical Neural Network diagram (used for data set for year t2 as an example) is shown in Figure 1.

As shown in the Figure 1, there is one input layer (left most), one hidden layer (middle one) and one output layer (right most). Within input layer, there are 5 nodes equal to numbers of predictors. Output layer has 2 nodes as levels of dependent variable (bankrupt 1 and non-bankrupt 0). The numbers of nodes, 2 in hidden layer can be adjusted.

The network specifications followed in building the models were: (i) 70:30 ratio was set for training and testing network (ii) Hyperbolic Tangent function was used as activation function for hidden layer (iii) Softmax

Figure 1. Artificial Neural Network Diagram



Source: SPSS Output.

function was used as activation function for output layer (iv) Range of nodes in hidden layers was set as 1 to 50 (v) Batch Training was used for training network (vi) Scaled Conjugate Method was used as Optimization algorithm (vii) Initial Lambda was set as 0.0000005 (viii) Initial Sigma was set as 0.00005 (ix) Interval centre was set as 0.00 (x) Interval offset was set as ± 0.50 (xi) Minimum Relative change in Training Error was set as 0.0001 (xii) Minimum Relative change in Training Error Ratio was set as 0.001(xiii) Maximum Training Time was set as 15 minutes and (xiv) Maximum steps without a decrease in error was set as 1.

Hyperbolic Tangent function has the following form:

$$Y(c) = \tan h\,(c) = \frac{(e^{c} - e^{-c})}{(e^{c} + e^{-c})} \qquad (2.7)$$

Where, c is the input from previous nodes. Y(c) takes real-value arguments and transforms them to the range (– 1, +1).

Softmax function has the following form:

$$Y(c) = \frac{1}{1 + e^{-c}} \qquad (2.8)$$

Y(c) takes real-value arguments and transforms them to the range (0, +1).

Tools for Analysing Classification Results

Empirical analysis of classification results which was vertical (across the years) and horizontal (across the discriminating schemes) was the preliminary analysis tool. Comparison of classification results produced by models was done with the help of *ROC curves*. *Paired Sample t-test* was used for hypothesis testing.

Analysis and Results

The 6 data sets pertaining to year of bankruptcy (t0), one year prior to bankruptcy (t1), two years prior to bankruptcy (t2), three years prior to bankruptcy (t3), four years prior to bankruptcy (t4) and five years prior to bankruptcy (t5) were run through Logistic Regression (LR) and Artificial Neural Network (ANN) by SPSS. The classification accuracies of models and validated results have been discussed in the following section.

Models Overall Classification Accuracies

The following Table 1 shows overall classification accuracies in percentages displayed by Artificial Neural Network (ANN) and Logistic Regression (LR). The last column shows the difference between overall classification accuracies displayed by each models.

Table 1: Overall Classifications through ANN and LR Models

Years	*Overall ANN Model*	*Overall LR Model*	*Difference: ANN-LR*
T0	91.10	89.70	+1.40
T1	84.40	83.90	+0.50
T2	81.50	77.59	+3.91
T3	75.20	73.56	+1.64
T4	75.00	70.12	+4.88
T5	70.20	69.54	+0.66

Source: *SPSS Output.*

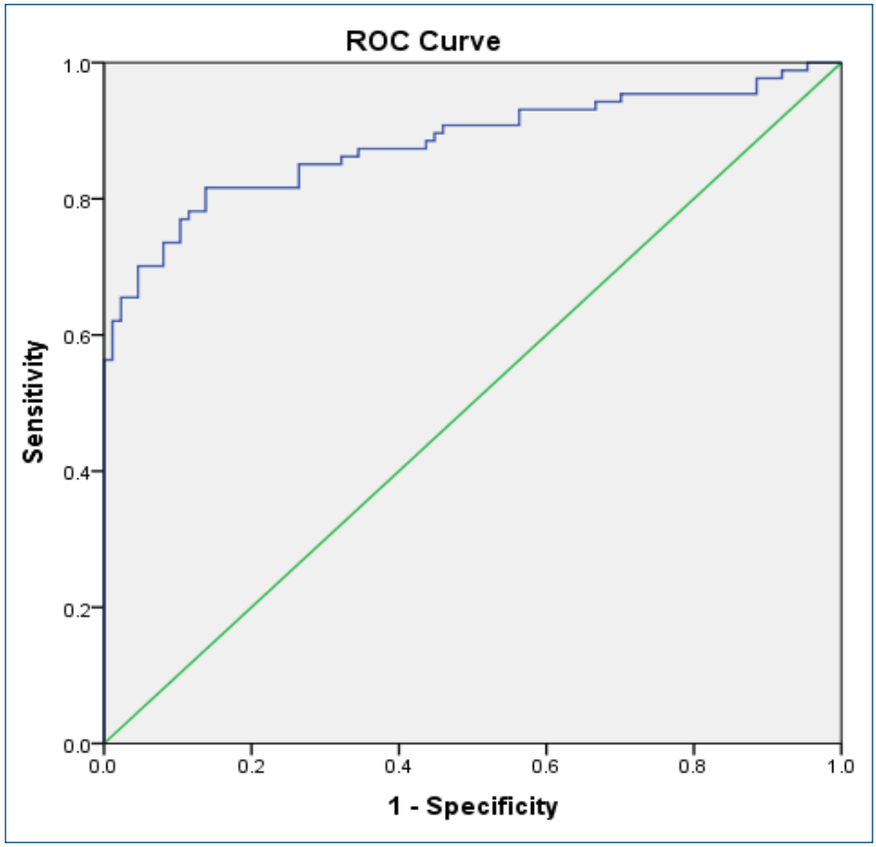
Both the models showed the highest classification accuracy for the year t0 and the lowest for the year t5. This was because of diminishing discriminating capability

Table 2: Predictors Prediction Capability

Years	*Numbers of Poor Predictors*	*Poor Predictor's p-value of F statistics/Chi-square statistics [>0.05]*
T0	2	CL/CA [0.733], CHIN [0.101]
T1	1	CL/CA [0.623]
T2	3	WC/TA [0.346], CL/CA [0.271], CHIN [0.187]
T3	4	SIZE [0.095], WC/TA [0.504], CL/CA [0.883] & ***OENEG [0.202]***
T4	4	SIZE [0.223], WC/TA [0.975], CL/CA [0.485] & ***OENEG [0.469]***
T5	6	SIZE [0.272], WC/TA [0.892], CL/CA [0.289], CHIN [0.136], FO/TL [0.097] & ***OENEG [0.35]***

Source: *SPSS Output.*

Figure 2. Receiver Operating Characteristic Curve



Source: SPSS Output.

of predictors across the years. The year of bankruptcy, t0 had maximum discriminating capability possessed by predictors. This was obvious as bankrupt companies were financially strained at the year of filing for bankruptcy whereas their counter parts (in the analysis, non-bankrupt and healthy companies) were not experiencing any financial strain as captured by predictors or financial ratios. However, 5 years prior to year of bankruptcy, both categories of bankrupt and non-bankrupt did not have *so* differentiable predictors or, in other terms, both the categories were almost the same. Table 2 shows poor prediction capability of predictors judged through F-test for continuous predictors and through Chi-square test for categorical predictors at 5% level of significance.

The comparison of overall classification accuracies clearly favoured the supremacy of ANN over LR. All years showed marginally higher classification accuracies by ANN models. In the year's t2 and t4, the differences were impressive in the tune of 3.91 and 4.88 respectively. The differences were found significant at 5% level of significance with p-value as 0.032 associated with Paired Sample t-test. Prior to Paired Sample t-test, the classification accuracies displayed by ANN and LR were tested by One Sample Kolmogorov Smirnov test and p-values were found as 0.939 and 0.987 respectively. Thus, necessary condition for applying Paired Sample t-test was met.

The above results were crosschecked by the areas under Receiver Operating Characteristic Curves (ROC) captured by ANN and LR. ROC curves are plotted against *(1-specificity)* on X-axis and *sensitivity* on Y-axis for a range of cut-offs. Sensitivity is the probability of classifying a case wrongly when the case belongs to category 1. This is termed as Type I error in the domain of terminologies used for explaining classification results. Similarly, specificity is the probability of classifying a case wrongly when the case belongs to category 0. This is termed as Type II error. ROC curves are used for comparing different discriminating schemes. The closer the ROC curve towards left top corner, the better the

Table 3: Area under Receiver Operating Characteristic Curves through ANN and LR Models

Years	*Area under ROC Curve ANN Model*	*Area under ROC Curve LR Model*	*Difference: ANN-LR*
T0	95.40	95.40	0.00
T1	91.60	91.30	+0.30
T2	89.30	86.30	+3.00
T3	81.40	80.60	+0.80
T4	81.20	80.80	+0.40
T5	74.60	75.70	-1.10

Source: SPSS Output.

curve is. Judging closeness of *two* ROC curves towards left top corner is a subjective matter which is resolved by the term 'area under ROC curve'. Area under ROC curve is an indication of efficiency of classification scheme. Thus, the higher the area under ROC curve, the better is the discriminating scheme. ROC curves are generated for ANN by SPSS V 20; however, for LR, these are not default output. ROC curves for LR were generated by separate commands through SPSS V 20.

A typical ROC Curve has been shown in Figure 2.

The Table 3 shows the areas under Receiver Operating Characteristic Curves captured by ANN and LR across the years. The last column shows the differences in areas under ROC curves captured by ANN and LR models.

As evident from the above table, areas under ROC curves were marginally higher in case of ANN models across the years. Besides, year t3 which showed higher area captured by ANN model by 3.00 percent, rest of the years were marginally higher in case of ANN models. The differences were found not significant at 5% level of significance with p-value as 0.352 associated with Paired Sample t-test. Prior to Paired Sample t-test, the areas under ROC curves captured by ANN and LR were tested by One Sample Kolmogorov Smirnov test and p-values were found as 0.964 and 0.943 respectively. Thus necessary condition for applying Paired Sample t-test was met.

The Overall classification results and Areas under ROC curves captured by ANN and LR models were better for ANN models; however, the final verdict can be framed only after analyzing Validation results which are discussed below.

Validated Overall Classification Accuracies

We took out 44 cases (22 bankrupts and 22 non-bankrupts) out of initial paired samples of 218 cases for each of six years for validation purpose. These 44 cases were not used for model building. The models were first saved in ***xml*** files and then later applied by *Scoring Wizard* commands available under *Utilities* in SPSS V 20 software.

Table 4 shows the Validated Overall classification accuracies displayed by ANN and LR. The last column shows the difference between overall validated results displayed by ANN and LR.

Table 4: Overall Validated Classifications through ANN and LR Model

Years	*Overall* **validation** *by ANN Model*	*Overall* **validation** *by LR Model*	*Difference: ANN – LR*
T0	100.00	97.73	+2.27
T1	79.56	81.82	-2.26
T2	79.56	81.82	-2.26
T3	72.73	72.73	0.00
T4	63.64	52.30	+11.34
T5	63.64	59.10	+4.54

Source: SPSS Output.

Validated results were the same for year t3, higher by LR models in years t1 and t2 both by 2.26 percent and higher by ANN models for years t0, t4 and t5. The lead taken by ANN model over LR model for year t4 was spectacular in the tune of 11.34 percent. These mix results could not show supremacy of any model over another. Paired Sample t-test showed p-value as 0.331 which was not significant at 5% level of significance, thus confirming the statistically same performance by both the models. Prior

to Paired Sample t-test, the validation results by ANN and LR were tested by One Sample Kolmogorov Smirnov test and p-values were found as 0.866 and 0.993 respectively. Thus, necessary condition for applying Paired Sample t-test was met.

As a commonly accepted/practiced rule of thumb (gathered through extensive literature review) of considering a classification accuracy more than 70% as good classification accuracy depicts that both ANN and LR models could display good validated results till year t3 (three years prior to bankruptcy).

Conclusions

In this work, we compared the classification accuracies of bankruptcy prediction models based on Logistic Regression and Artificial Neural Networks. Based on *Overall Classification results*, *areas under Receiver Operating Characteristic Curves* and *Validated Overall Classification results*, models based on Artificial Neural Network were found marginally better. Our findings were in line with the findings of Luther (1994), Zhang (1999) and Lin (2009). However, stability of ANN remains an issue which needs attention by technique developers. Altman had mentioned in his paper pertaining to study with Italian data that the behavior of the network became at times unexplainable and unacceptable. Altman had further mentioned in the same paper that ANN had shown enough promising features to provide an incentive for better implementation techniques and more creative testing.

References

Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparison using linear discriminant analysis and neural networks (the Italian experience*)*. *Journal of Banking and Finance*, 18(3), 103.

Hair, J. F. Jr., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate Data Analysis* (6th ed.). Upper Saddle River, NJ: Pearson/Prentice Hall.

Javanmard, H., & Saleh, F. (2009). The comparison artificial neural networks and multi decimal analysis models for forecasting bankruptcy and financial distress. *Proceedings of the World Congress on Engineering, 2*

Lin, T. H. (2009). A cross model study of corporate financial distress prediction in Taiwan: Multiple discriminant analysis, logit, probit and neural networks models. *Neurocomputing,* 72(16-18), 3507-3516.

Luther, K. R. (1994). An Artificial Neural Network Approach to Predicting the Outcome of Chapter 11 Bankruptcy. *Journal of Business and Economic Studies*, 4(1).

Nargundkar, R. *Marketing Research* (3rd ed.). New Delhi: McGraw-Hill.

Ohlson, J. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1), 109-31.

Registered cases for bankruptcy. (2011, Jan 25). Retrieved from http://bifr.nic.in/casesregd.htm

Wang, Y., & Campbell, M. (2010). Financial ratios and the prediction of bankruptcy: The Ohlson model applied to Chinese publicly traded companies. *Journal of Organisational Leadership and Business.*

Zhang, G., Hu, Y. M., Patuwo, E. B., & Indro, C. D. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European Journal of Operations Research,* 116(1), 16-32.

Zurada, J. M. (1996). *Introduction to Artificial Neural Systems.* Mumbai: Jaico Publishing House.

Sentiment Analysis of Swachh Bharat Abhiyan

Sahil Raj*, Tanveer Kajla**

Abstract

The present paper is about the social media analytics. It is a new tool to analyse the behaviour of the users who use social networking sites and other social sites like blogs, forums etc. Every organisation uses this tool to analyse their customers. Even the government agencies are using these analytical tools to get the feedback of their newly launched missions and their policies. In this paper the sentiment analysis of Swachh Bharat Abhiyan is done with the help of tweets extracted from twitter. Tweets regarding Swachh Bharat Abhiyan are extracted with the help of an open source software R-studio. The geo-locations of tweets are also extracted in the software and the results are plotted on the map of India. The pattern of tweets is analysed and the popularity of the mission is evaluated. The word cloud of the popular and the most used words is also formed in the R-studio software. With the overall analysis, the popularity of the mission is perceived according the regions on the map of India, and the strategies can be applied to popularize the campaign in the lesser known regions of India.

Keywords: Swachh Bharat, Word Cloud, Geo-Location, Campaign

Introduction

Swachh Bharat Abhiyan is a special campaign by the BJP government to clean the roads, streets and infrastructure of the country. It is the visionary mission launched by our honourable Prime Minister Shri Narendra Modi. It was launched on 2nd October, 2014. This campaign is one of the India's biggest campaign, covering around 3 million government employees. This mission is widely popular among the citizens of the country as it directly gives them the responsibility to clean up their own country. The cleanliness campaign is also covering the schools, colleges and universities. The Prime Minister also nominated the nine big personalities of the country and also gave them the responsibility to nominate nine more people to join the campaign, making a chain to increase the participants of the mission. The aim of the campaign is to achieve the vision of cleanIndia by the year 2019. The main objectives of the campaign are to finish up the manual scavenging and to eliminate the open defecation which is the main cause of the tuberculosis in India. The construction of individual, community and cluster toilets were also included. The villages should be cleaned and to lay water pipelines in the villages to ensure 24 hour water supply to all the households by 2019.

Review of Literature

Social media produces massive amount of data (Ediger *et al.*, 2010).Twitter is a micro blogging service where users create messages called tweets. These tweets sometimes express opinions about different topics (Go, Huang & Bhayani, 2009). Social media is used as the official media platform by the celebrities, politicians but the research can be centred around the events also (Tumasjan *et al.*, 2010). Millions and trillions of users share their opinions on social media sites. Athletes use tweets to interact with their fans. Twitter feeds can also be used for emergency events like natural disaster and crisis management during or after the time of disaster(Zielinski *et al.,* 2012). The research can be done in English or any other language, where relevant tweets can be classified and extracted. Sentimental classifier is able to determine the neutral, negative and positive tweets. Algorithm can be made accurately to classify Twitter messages as positive or

* Assistant professor, School of Management Studies, Punjabi University, Patiala, Punjab, India.
E-mail: er_sahil@yahoo.com
** Research Scholar,School of Management Studies, Punjabi University, Patiala, Punjab, India.
E-mail: er.tanveer47@gmail.com

negative, with respect to a query term (Go *et al.*, 2009). Searching the tweets can be more easily done by using the hashtags ahead of the subject to be searched. Hashtags are used to categorise the messages in the twitter. With the use of the hashtags, twitter users can propagate the ideas and promote specific topics and people. Hashtags are used to streamline the search and at the same time, to increase the effectiveness of the research (Wang *et al.*, 2011). But the biggest problem is the size of an unstructured data, which is in chunks, and there can be lot of repetition in the data, so this unstructured data should be taken in large volumes and therefore more complex algorithms are used to classify very high number of tweets. Linguistic features can also be used to identify the language used in the tweets. The research is also done on how many retweets came and what are the factors contributing to the retweets (Naveed *et al.*, 2011). The researchers brought much attention to the data from the tweets as the data is very irregular due the 140 character limits put on the tweets (Saif, He & Alani, 2012). The authors show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. Using the corpus, the authors build a sentiment classifier that is able to determine positive, negative and neutral sentiments for a document (Pak & Paroubek, 2010).

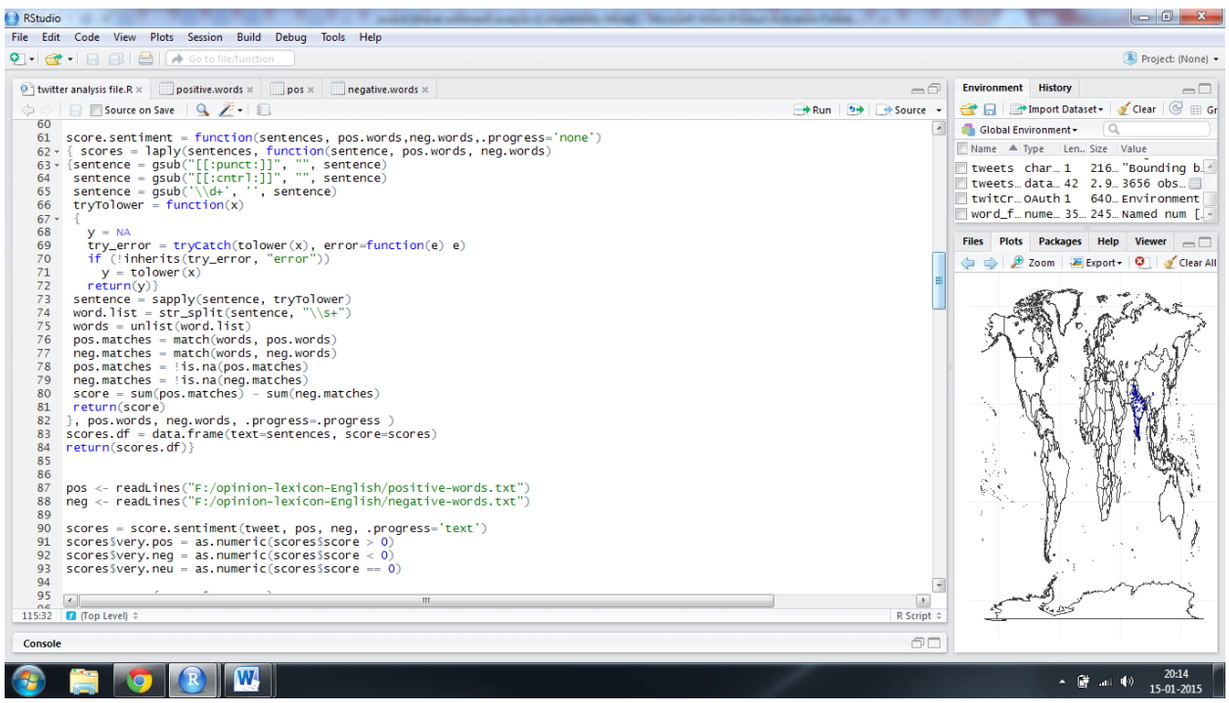
Objectives

The objective of the research paper is to do the sentimental analysis on the cleanliness campaign launched by the present government. The analysis will give the perception of the citizens regarding this new campaign. The second objective of the paper is to find the location of the tweets regarding the Swachh Bharat, and then plot them on the map of India to check the popularity of the mission. The third objective of the paper is to make a word cloud of the tweets which will prompt the most used words in the tweets.

Research Methodology

The data related to the research are the unstructured data. The unstructured data are extracted from twitter in the form of tweets. For extracting the data from the twitter we used open-source R-Studio software which is based on windows platform. First of all the developer account has to be created in the twitter, and then the application to mine the text is created in the application account of the twitter. In this application, keys and tokens were generated. With these keys and tokens, authorisation is provided by twitter to extract the tweets. In R-studio software, these tweets were extracted with the help of

Figure 1: Function for Sentiment Analysis



some text mining packages. When the extraction of tweets is done, the analysis is done on the set of tweets. In the analysis part, we are categorising the tweets into the positive tweets, negative tweets and neutral tweets. It is done with the assistance of text files having the list of positive and negative words. With these lists tweets are compared and the positive and negative ratings are given to the tweets. The rated tweets are plotted on the pie chart with the help of which analysis is done.

Figure 1: shows the function for sentiment analysis. With this function the tweets are rated to positive, negative and neutral.

The word cloud of the tweets is the pictorial representation of the words in the software. It will highlight the most used or talked about words in the middle of the cloud. With this cloud some prominent parts of the mission can be detected. The cloud can also prompt us about the important subjects of the mission. The word cloud is created with the help of the word cloud package and then the word cloud function is created to display the word cloud.

The Highlighted Portion of the Screenshot in Figure 2 Shows the Word Cloud Function

The geolocation of tweets is done with the assistance of ggplot package and maps package.The tweets are extracted according to the location. According to the research, tweets should be extracted from the Indian region. So the longitude and latitude of the area is given from the south-west corner to the north-east corner in the command interface and the tweets will be extracted from this bounded region. Maps package will print the map in the software and ggplot package will plot the extracted tweets from the twitter on the map. The geo-located tweets will give the pattern on the map, and will give the region from where more number of tweets are coming.

The tweets which are fetched according to the Indian region will be saved in bharat2.json file. Then these tweets be cleaned for the relevant text, and then plotted on the world map.

Analysis And Discussion

When the tweets are extracted from twitter on swachhbharat, then the sentiment analysis is done on the tweets. Analysis is done through pie chart, as shown in Figure 4. Pie chart clearly explains that twitter users are having a liking towards the swachhbharat mission as there are very less negative responses towards the campaign.

Figure 2: Word Cloud Function

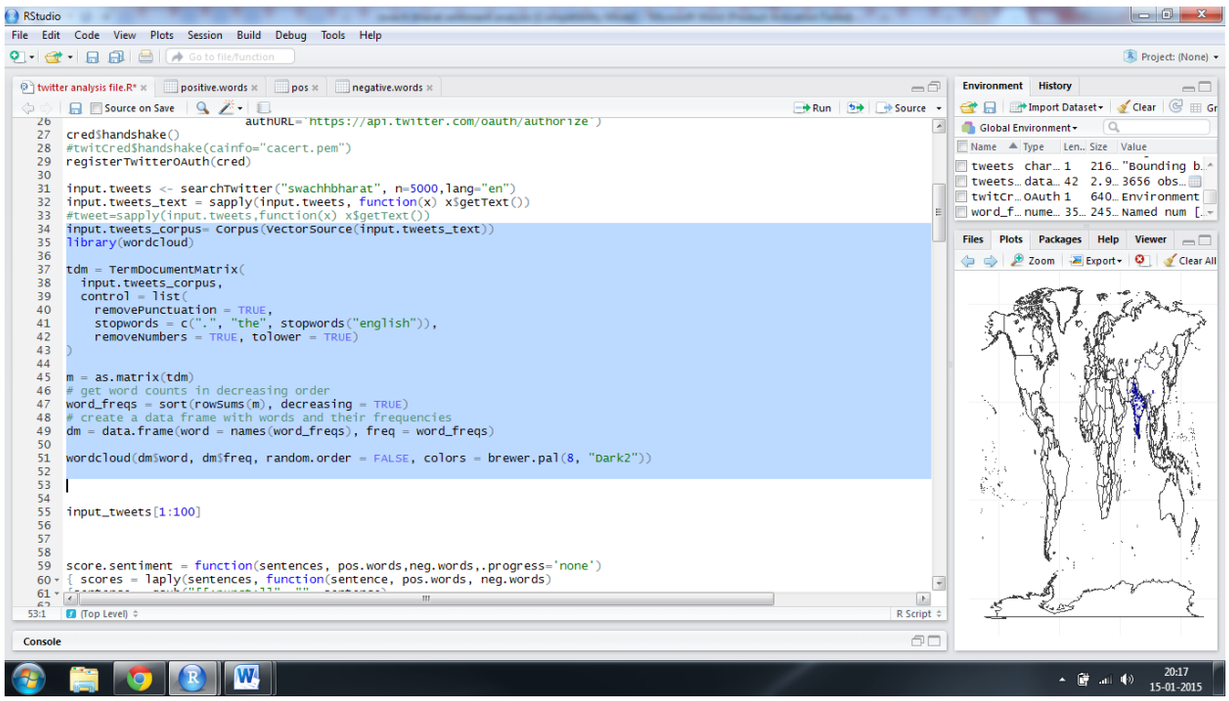


Figure 3: Steps to Save Tweets in bharat2.json file

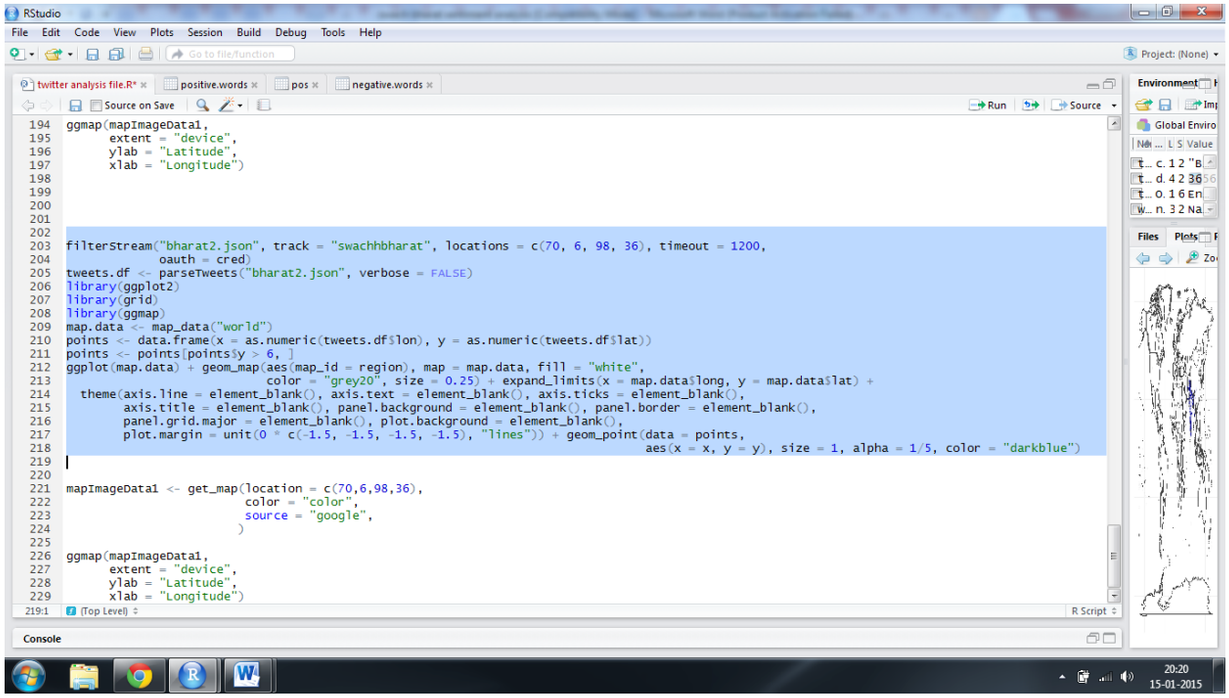


Figure 4: Analysis Through Pie Chart

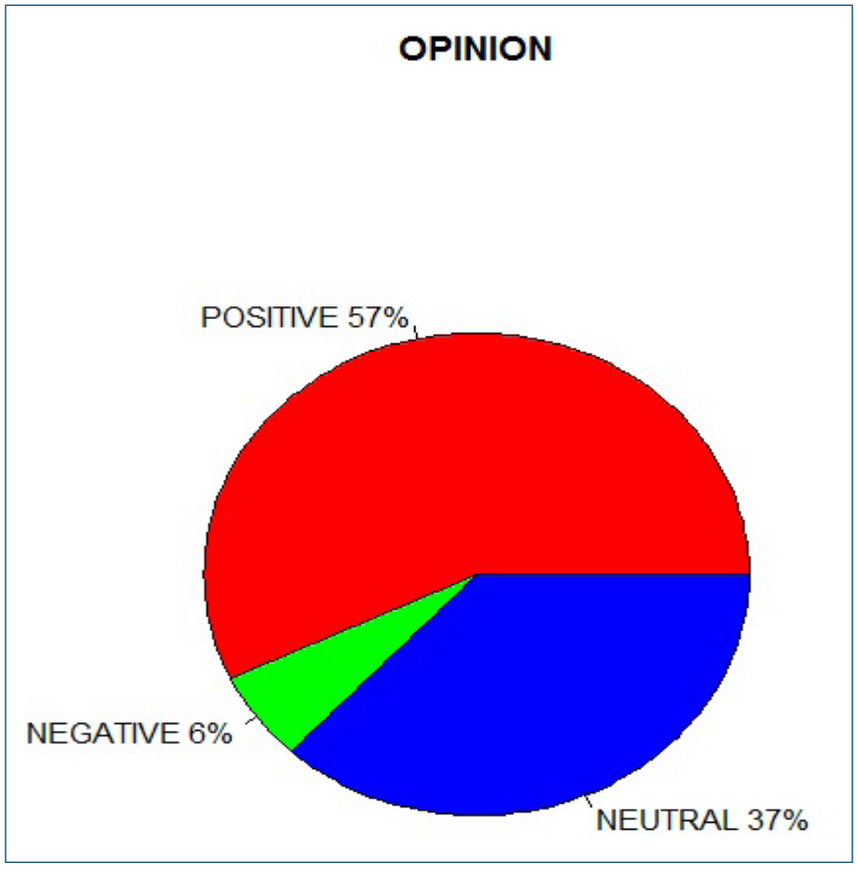
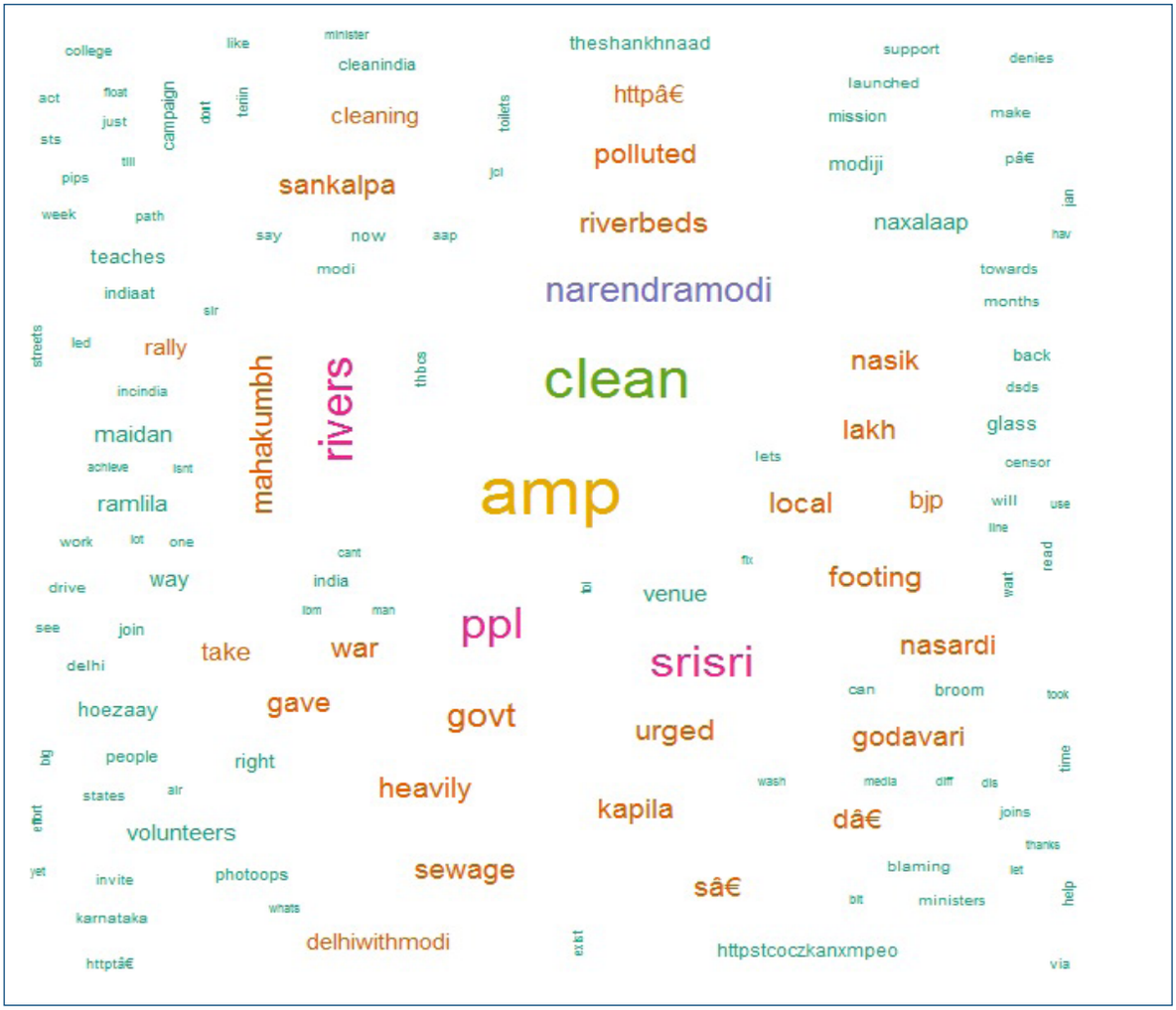


Figure 5: Word Cloud of the Tweets

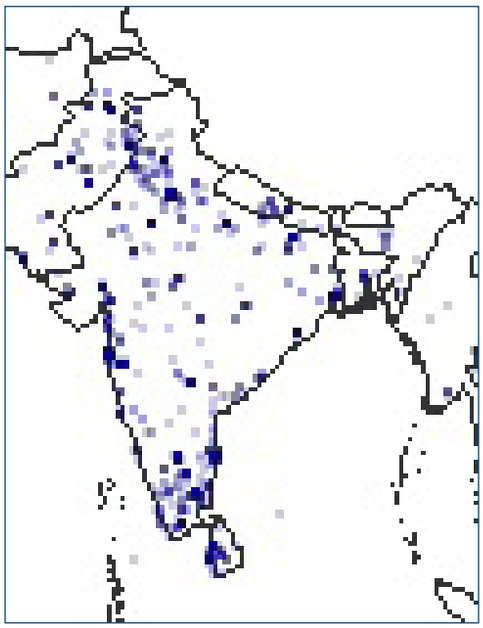


There is more number of neutral and positive responses on this campaign. This clearly explains that the people of India are adapting to this kind of campaign.

Word cloud of the tweets is also made in the software. In the word cloud it is clearly seen that the Narendra Modi is a heavily used word, as it is seen in the middle of the cloud. This shows that he is an integral part of the mission. It is a well-known fact that he is the creative mind behind this whole campaign. The word "clean" is the most highlighted word in the word cloud. This indicates that the twitter users, who are sending tweets on the twitter about the Swachh Bharat were more concerned about cleanliness, which is the main objective of this campaign.

Tweets are extracted from the Indian region to see the popularity of the region, from where the people are more concerned about the Swachh Bharat. When the tweets are plotted on the map of India then the tweets are easily distinguished on the basis of their recurrence. More number of tweets are coming from northern and southern region of India as it is clearly seen on the map. That means the campaign is more popular in these regions. The campaign has started from New Delhi, and that is the only reason it is very popular among them. In the southern region, the campaign is popular in the state of Tamil Nadu. The campaign is not so popular in the Gujarat region and in other states of India. Many tweets have been received from the neighbouring countries also. This shows that campaign got worldwide popularity and the new government is connecting with the neighbouring countries in a positive way.

Figure 6: Origin of Tweets



Conclusion

Twitter is the emerging social media tool, and every organisation is using this tool for marketing of their product. Internet users from different background and countries come under a single medium to share their views and opinions about any new happenings, and can also recommend to other users or some experts, who can even give their advice to internet users, who are new to the world of gadgets. The sentiment analysis clearly shows that this campaign is a success among the people of India. People have given a very positive response on this initiative by the Indian Prime Minister.

In this analysis many loopholes are also found for the campaign. Though the campaign is popular and is very much appreciated, but still it is not popular in the central region of India. Very less number of tweets has been received from these parts of India. Northern part of the India is also less involved in this campaign. The main aim of this campaign was cleanliness and the tweets show that 'clean' is a highly used word in the tweets as shown in the word cloud.

Recommendations

The government should not focus on the urban areas, as this campaign is popular in urban areas only. The campaign should be encouraged in the central parts of India also. The campaign has been accepted by the people, so this initiative can be easily extended to the lesser known regions also.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media* (pp. 30-38). Association for Computational Linguistics.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market.*Journal of Computational Science*, 2(1), 1-8.

Bruns, A., & Stieglitz, S. (2013). Towards more systematic Twitter analysis: Metrics for tweeting activities. *International Journal of Social Research Methodology*, 16(2), 91-108.

Ebner, M., Altmann, T., & Softic, S. (2011). @ twitter analysis of# edmedia10–is the# informationstream usable for the# mass. *Form@ re-Open Journal per la formazione in rete*, 11(74), 36-45.

Ediger, D., Jiang, K., Riedy, J., Bader, D. A., Corley, C., Farber, R., & Reynolds, W. N. (2010, September). Massive social network analysis: Mining twitter for social good. In *Parallel Processing (ICPP), 2010 39th International Conference on* (pp. 583-593). IEEE.

Go, A., Bhayani, R., & Huang, L. (2009a). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1-12.

Go, A., Huang, L.,&Bhayani, R. (2009b). Twitter sentiment analysis. *Entropy,* 17.

Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D. A., Haug, L., & Hsu, M. C. (2011, October). Visual sentiment analysis on twitter data streams. In Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on (pp. 277-278). IEEE.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg!. *ICWSM*, 11, 538-541.

Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C. (2012, April). Tedas: A twitter-based event detection and analysis system. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on* (pp. 1273-1276). IEEE.

Mathioudakis, M., & Koudas, N. (2010, June).Twitter monitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD*

International Conference on Management of data (pp. 1155-1158).ACM.

Naveed, N., Gottron, T., Kunegis, J., &Alhadi, A. C. (2011, June). Bad news travel fast: A content-based analysis of interestingness on twitter. In*Proceedings of the 3rd International Web Science Conference* (p. 8).ACM.

Pak, A., &Paroubek, P. (2010, May).Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREC*.

Saif, H., He, Y., &Alani, H. (2012).Alleviating data sparsity for twitter sentiment analysis. CEUR Workshop Proceedings (CEUR-WS. org).

Thomas, K., Grier, C., Song, D., & Paxson, V. (2011, November). Suspended accounts in retrospect: An analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference* (pp. 243-258).ACM.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10, 178-185.

Wang, X., Wei, F., Liu, X., Zhou, M., & Zhang, M. (2011, October). Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1031-1040). ACM.

Yardi, S., & Boyd, D. (2010). Dynamic debates: an analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society*, 30(5), 316-327.

Zielinski, A., Bügel, U., Middleton, L., Middleton, S. E., Tokarchuk, L., Watson, K., & Chaves, F. (2012, April). Multilingual analysis of twitter news in support of mass emergency events. In *EGU General Assembly Conference Abstracts* (Vol. 14, p. 8085).

Guidelines for Authors

International Journal of Business Analytics and Intelligence welcomes original manuscripts from academic researchers and business practitioners on the topics related to descriptive, predictive and prescriptive analytics in business. The authors are also encouraged to submit perspectives and commentaries on business analytics, cases on managerial applications of analytics, book reviews, published-research paper reviews and analytics software reviews based on below mentioned guidelines:

Journal follows online submission for peer review process. Authors are required to submit manuscript online at http://manuscript.publishingindia.com

Title: Title should not exceed more than 12 Words

Abstract: The abstract should be limited to 150 to 250 words. It should state research objective(s), research methods used, findings, managerial implications and original contribution to the existing body of knowledge

Keywords: Includes 4–8 primary keywords which represent the topic of the manuscript

Main Text: Text should be within 4000-7000 words Authors' identifying information should not appear anywhere within the main document file. Please do not add any headers/footers on each page except page number. Headings should be text only (not numbered).

Primary Heading: Centered, capitalized, and italicized.

Seconday Heading: Left justified with title-style capitalization (first letter of each word) and italicized.

Tertiary Heading: Left justified and indented with sentence-style capitalization (first word only) in italics.

Equations: Equations should be centered on the page. If equations are numbered, type the number in parentheses flush with the left margin. Please avoid using Equation Editor for simple in-line mathematical copy, symbols, and equations. Type these in Word instead, using the "Symbol" function when necessary.

References: References begin on a separate page at the end of paper and arranged alphabetically by the first author's last name. Only references cited within the text are included. The list should include only work the author/s has cited. The authors should strictly follow APA style developed by American Psychological Association available at American Psychological Association. (2009). Publication manual of the American Psychological Association (6th Ed.). Washington, DC.

Style Check

To make the copyediting process more efficient, we ask that you please make sure your manuscript conforms to the following style points:

Make sure the text throughout the paper is 12-point font, double-spaced. This also applies to references.

Do not italicize equations, Greek characters, R-square, and so forth. Italics are only used on p-values.

Do not use Equation Editor for simple math functions, Greek characters, etc. Instead, use the Symbol font for special characters.

Place tables and figures within the text with titles above the tables and figures. Do not place them sequentially at the end of the text. Tables and figures must also be provided in their original format.

Use of footnotes is not allowed; please include all information in the body of the text.

Permissions

Prior to article submission, authors should obtain all permissions to use any content if it is not originally created by them. When reproducing tables, figures or excerpts from another source, it is expected to obtain the necessary written permission in advance from any third party owners of copyright for the use in print and electronic formats. Authors should not assume that any content which is freely available on the web is free to use. Website should be checked for details of copyright holder(s) to seek permission for resuing the web content

Review Process

Each submitted manuscript is reviewed first by the chief editor and, if it is found relevant to the scope of the journal, editor sends it two independent referees for double blind peer review process. After review, the manuscript will be sent back to authors for minor or major revisions. The final decision about publication of manuscript will be a collective decision based on the recommendations of reviewers and editorial board members

Online Submission Process

Journal follows online submission for peer review process. Authors are required to register themselves at http://manuscript.publishingindia.com prior to submitting the manuscript. This will help authors in keeping track of their submitted research work. Steps for submission is as follows:

1. Log-on to above mentioned URL and register yourself with "International Journal of Business Analytics & Information"
2. Do remember to select yourself as "Author" at the bottom of registration page before submitting.
3. Once registered, log on with your selected Username and Password.
4. Click "New submission" from your account and follow the 5 step submission process.
5. Main document will be uploaded at step 2. Author and Co-author(s) names and affiliation can be mentioned at step 3. Any other file can be uploaded at step 4 of submission process.

Editorial Contact
Dr. Tuhin Chattopadhyay
Email: dr.tuhin.chattopadhyay@gmail.com
Ring: 91-9250674214

Online Manuscript Submission Contact
Puneet Rawal
Email: puneet@publishingindia.com
Ring: 91-9899775880

International Journal of Business Analytics and Intelligence

SUBSCRIPTION DETAILS

Dispatch Address:-
The Manager,
International Journal of Business Analytics and Intelligence
Plot No-56, 1st Floor
Deepali Enclave, Pitampura
New Delhi -110034
Ph - 9899775880

Subscription Amount for Year 2015

	Print	Print + Online
Indian Region	Rs 2500	Rs 3200
International	USD 150	USD 180

Price mentioned is for Academic Institutions & Individual. Pricing for Corporate available on request. Price is Subject to change without prior notice.

Payment can be made through D.D./at par cheque in favour of "Publishing India Group" payable at New Delhi and send to above mentioned address.

Disclaimer

The views expressed in the Journal are of authors. Publisher, Editor or Editorial Team cannot be held responsible for errors or any consequences arising from the use of Information contained herein. While care has been taken to ensure the authenticity of the published material, still publisher accepts no responsibility for their accuracy.

Journal Printed at Anvi Composers, Paschim Vihar.

Copyright

www.manuscript.publishingindia.com

PublishingIndia.com

Publishing India Group

Plot No. 56, 1st Floor, Deepali Enclave
Pitampura, New Delhi-110034, India
Tel.: 011-47044510, 011-28082485
Email: info@publishingindia.com
Website: www.publishingindia.com

ISSN 2321-1857
01
9 772321 185001