Benefits and Issues Surrounding Data Mining and its Application in the Retail Industry

Prachi Agarwal

Department of Computer Science, Suresh Gyan Vihar University, Jaipur, India

Abstract- Today with the advent of technology data has expanded to the size of millions of terabytes. For retail industries, customer's data works as tracks for analysing their buying behaviour. How this data is maintained and used for an effective decision making in retail industry is discussed in this paper. This not only increases profit for companies but also poses a challenge in the field of data mining about how probably "Recommended for You" items are chosen by the customers and how likable is the platform or store according to the customer preferences. Clustering algorithm is used to segregate customer profiles. Business Intelligence and analytics work to bring best decision routes for marketing.

Index Terms- data mining, business intelligence and analytics, retail industry, tesco, decision making, customer behaviour.

I. Introduction

Business Intelligence and Analytics (BI&A) is now recognized and understood as a very significant factor that can enhance organizational performance through intelligently taken decisions and meticulous ideas and perspectives (Chen et al. 2012; Davenport 2006). The evolution of modern day BI&A systems is largely attributable to Decision Support Systems (DSS). Business Intelligence and Analytics systems allow for the collation and transformation of data into information that can act as valuable inputs in the process of business decision-making. These systems allow the organization to emphasize on their decision routes, that is, the format through which decisions are arrived at in a company in order to enhance performance. By integrating data with the decision routes, BI&A systems provide for improved end decisions that in turn imply organizational success.

II. Data Mining

History of Data Mining

Data mining is also known as the knowledge that is discovered from databases. A probable definition of data mining is – a process of extracting previously unknown, implicit, and useful information from the data found from databases. Data mining has come to become as a recognized field of research in the recent past. The Gartner Group had predicted in the year 1998 that almost half of the companies from Fortune-500 would be using technologies based on data mining within the year 2000 (Webster & Watson, 2002). However, most of the techniques are anything but new in the arena since they have been used for ages in statistics.

Database technology was developed in the 1960s which allowed the storage and utilization of data in a systematic manner. These advantages resulted in being not too useful for business since it required a lot of programming in order to generate simple reports. To add to it, computers were slow and expensive during those days.

During the 1970s advances in computer systems and their database, were going through further advances. These proceedings culminated into OLTP or the Online Transaction Processing in the 1980s, which allowed transactions to be put under codes and captured without the intervention of human (Rajagopalan et al, 1993). A new and more significant source of data was introduced with the invention of personal computer (PC). Gradually computer became a tool for everyday use for every employee. Marketers were especially benefited with the spreadsheet software for undergoing data analysis. As a result they discovered how powerful data is for aiding decision making capabilities. On the hand, statistics and artificial intelligence were two other streams of data mining where researches were taking place. Researchers were developing techniques for detection of automatic relationship and data visualization.

In the beginning of the 1990s the above mentioned technologies started to be highly developed. A large amount of database was now accessible to users who had their PCs and a local area network, along with server/client technology (Forgionne, 1999). The statistical and database software was proving to be user-friendly and thus several end users were able to generate their own reports and perform analyses on their own as well. Technologies became more powerful and this resulted in businesses organizing systems for storage of special data in order to provide structured and useful data throughout the company. The data stores, also known as the warehouses for data, provide the required raw material for data mining on a large scale.

Data Mining and Business Intelligence

Software solutions are instrumental to utilizing business applications. Risk management and enterprise decision-making now cannot be separated from mining tools. Business Intelligence (BI) is acquired by using mining. Use of data warehousing and Information Systems (IS) have made it possible for enterprise datasets to grow rapidly. Credit card companies generally log millions of transactions in any given year. Mobile operators and telecommunications companies usually generate largest data sets. User accounts exceed 100 million which generate billions of data per year. Against such numbers, OLAP or other analytical processing and manual operation do not have any chance to stand up. BI, on the other hand, makes the task possible.

According to Gartner Group "Data mining and artificial intelligence are at the top five key technology areas that will clearly have a major impact across a wide range of industries within the next three to five years". This conclusion was arrived at in a 1997 report (Jun Lee & Siau, 2001). Another Gartner group report that came out in 2008 (Godbole & Roy. 2008), says that 80% of information systems data is unstructured and an increase of double the size was envisaged every three months. Decision support systems had taken up BI in a big way. BI's domination was most visible in insurance, retail and banking.
To cite a success story, The First American Corporation has effectively improved investment climate and customers' loyalty by using BI. Employees in the categories of knowledge workers, middle management, executives, analysts and operational management benefited from BI.

BI is acquired by data in today's environment. Mining tools are used to implement BI. Rivals are left behind, business operations are better managed, risk management and survivability are given a fillip by the data mined by the mining tools. Customer Relationship Management (CMR) is enabled through mining customers churn, their habits and patterns.

Customer churn happens when a certain percentage of customers leave the enterprise, perhaps to take advantage of their perceived and real better climate at a rival enterprise to keep the customers satisfied. Discovering alpha consumers who play ambassadors for the product and make the product a success is made possible by mining tools. Customers' segmentation depending upon their habits and trends is necessary to target these alpha consumers. Advertising agencies and Catalogue marketing cannot do without mining tools. Mining tools can also provide market analysis from which information regarding the products that are usually bought together can be put together. DM is especially good at it.

III. DATA MINING IN THE RETAIL SECTOR: CASE STUDIES

Data mining involves risks and returns in equal proportions. The following discussion shall analyze two organizations, Tesco and Amazon, with respect to their data mining operations, to understand in practical light, whether or not data mining results in improved performance for retail organizations.

A. Tesco

Tesco Plc, a globally recognized and rated retail chain deals in grocery as well as general products. Its headquarters are in Hertfordshire (Chestnut) and is presently the largest supermarket in the United Kingdom. It also has the highest recruitment rate in the whole of UK. According to its 2009 annual report, Tesco had a whopping 320000 employee strength globally with 2320 running stores spread over the 7 continents. Tesco touched the Asian markets only at the start of this decade with setting up stores in Malaysia, Taiwan and China (to name a few). It was in 1995 that Tesco succeeded in overtaking its rival Sainsbury's to become the highest UK superstore. This achievement made their market shares take a huge leap from 15.4% in 1995 to 29% in 2004. Tesco also managed to acquire the famous retail superstore T& S Plc which is stated to have had 90 stores around the UK (Bandura, 1986). Specifically the last decade has been of growth and progress. They ventured into newer verticals like e-commerce, major diversification and increase in the variety of products and services which they have been offering (Also including diversifying into garments brand etc).

Tesco is at present UK's leader in the grocery market sector and enjoys a 30% chunk of the grocery business. Also it is noteworthy that Tesco has presence in 13 different countries worldwide. Tesco is believed to have followed Ian McLaurin's achievements and the CEO dreamt of an organization whose prime motive would customer satisfaction. They also play on their idea of extending "every little helps" that the organization can. This policy was adopted during the transformation of Tesco. Under the leadership of Leahy, Tesco has tasted great heights in the business history map and is only behind 2 other superstore chains worldwide. (America's Wal-Mart and France's Carrefour).

Presently, Tesco has comprehensive consumer data and profiles on approximately 145 million Americans – which accounts for more than 65% of its adult population. This information consists of their search history, browsing routes, products viewed, purchase records, location information, unique identification codes for customers and their devices and information about their computer/device systems (Davenport, 2006). Through the in-store Wi-Fi available at all Tescot stores, it is able to maintain records of customer movements which are applied to enhance layouts and product placements. Besides the use of the traditional cookies, which track user browsing preferences for the use of the online advertising industry, Tescot uses unique identifiers to maintain exhaustive profiles of its customers. Tescot also capitalizes on the rich information left behind by smartphone users. Data mining is conducted on smartphone data to obtain the unique MAC address, which is device-specific (Arnott & Pervan,2008).

Tesco also uses advanced video software's and cameras to track in-store movements of customers. Mining of such data reveals significant information on the number of adults and children, number of cash counters idle/overloaded, parts of the store that are more crowded and popular, facial expressions of the customers in response to the products, etc. It is also reported to be developing a store-only social network, where buyers can interact to discuss, recommend or provide their opinion on different products, services, special schemes, offers, etc. Tesco's extensive investment in its exhaustive data mining practices, and the continuing success they have have resulted in, is ample evidence of the fact that if carried out smartly and efficiently, data mining can turn out to be extremely profitable for a retail organization.

B. Amazon

Amazon is amongst the largest and the most profitable e-commerce giants in the world today. Although the internet offers lucrative avenues for the use of data mining, Amazon does not have a formally developed, in-house data mining programme in place to track and understand the rich data trail left behind by its users (Kannungo, 2009). While it is most common to cite Amazon's "Recommended For You" feature when discussing real-life examples of data mining, the truth is that Amazon's reliance on data mining techniques is extremely restricted, especially when compared to competitor e-commerce

organizations operating on the same scale. Amazon does not track the behavior of buyers who eventually delete items from their cart before their final purchase, it does not mine data on buyers who add items to the cart and do not make a purchase and it also does not attempt to study data on interruptions in browsing experiences and why they occurred (Benbasat & Nault,1990). Rather than following a research and data-backed approach to enhance the experience they provide to users, they follow an intuitive approach to accomplish that goal. In the process, the organization side-steps opportunities to enhance their services.
Whenever the supermarket wars take place, the bullets fly thickest where the loyalty background is. In this regard the grocer who is considered to have the most potent weapon of loyalty is Tesco. Yet, the immense mystery belonging to the data world remains how the customer insights brought about from Tesco Clubcard data mountain help to make it to the UK's biggest grocery chain.

Six years post the launch of Clubcard, in early 2001, Tesco had bought a shares in majority in dunnhumby which is its data analysis supplier. People in the data land were aware that something big was waiting in the future. It was proved so. Until 2001 dunnhumby and Tesco had restricted analysis of customer samples typically to about 10%. This was done in order to control the expenses of transmission and data storage. The huge changes came with lower costs of technology and the growing desire by Tesco to go deeper into data regarding consumer insights. They had taken it as a challenge to decipher the 104 billion rows of data that was stored at one point of time.

Tesco Lifestyle happened to be a result; a modeling system and segmentation based on the customer shopping behavior. Lifestyles ultimately dealt with understanding of the factors which affect shopping behavior, for instance, promotions, price, healthy eating, along with measurement of the share Tesco's hold from a customer's wallet. Their aim was to make people spend more and nudge them towards buying products from Tesco that they would have bought from elsewhere.

Their campaigns included the famous - "What is in the basket?". Dunnhumby started to look for products in the shopping baskets of random customers to find out certain ones that are predictive of a lifestyle or a need - for example, weight-watching goods. Thereafter, the data was mined to find out which other products were related to this and how much of it is used by customers regularly. The analysis found out 25 different dimensions of shopping or typologies - these include factors like how ‘green', family oriented, and healthy a certain product was. These areas were such where traditional methods of calculation could hardly help, so there was a need to devise people's own approaches to collect data. Most of the 40K products that Tesco stocks were given a thumbs up or down across these dimensions.
data and methodology

In this section, Tesco grocery section would be surveyed before and after the use of data mining techniques would be surveyed in detail. The Grocery section of Tesco was chosen because it was the busiest section which attracted the maximum number of consumers.

C. Survey and Data

Customer woes continue to plague the executives and managers who stare at unresolved questions that beg their attention day in and day out. In a revelation of a kind, discussions between the Retail Food Industry Center’s (TRFIC) Board of Advisors and Walmart, it was realized that there is a disconnect between consumers who refuse payment for things the retailers had assumed that the former would want and the retailers themselves. (Wolfson,2000). This initiated a research in order to find out the motivations that drove consumers’ purchases. The questions asked were:

1. What makes the shoppers to choose a particular store to buy groceries?
2. What goes into making the decision about the stores?
3. Are all the shopping trips undertaken to the same shops for different purchases?

A nationwide telephone survey with 900 households as respondents (and in Atlanta, 300 households), was taken as a first step in the summer of 1999. The primary shopper for food was interviewed and the age group covered was between 18 and 75 years. Four different shopping aspects were covered: ready-to-eat/take out, stock up, fill-in and special occasion. The respondents had to rate a store on a scale from 10 (very Important) to 1 (not important) in correspondence with 30 factors given in the survey. The method of cluster analysis was employed to figure out the shoppers’ preferences for a shopping trip for stock up. For the purpose, six types of grocery shoppers were chosen.

D. Analysis

Cluster analysis is a technique to employ in data mining. The original data containing preferences was applied to arrive at a conclusion about the shoppers’ individual preferences. Market segmentation based on clusters has been put to use since the beginning of 1970s increasingly (Green 1995, Wind 1978). Customers were profiled based on hierarchical as well as non-hierarchical techniques. Segment of large data sets come across during marketing was carried out by k-means, a non-hierarchical method towards the end of 80s (k denotes the number of clusters).

Something that may be called “path dependence” may impact hierarchical approach and so non-hierarchical clustering techniques are being used. Path dependence involves the tendency of the objects grouped to stay together as in the beginning even if one does not concur with cluster average. In the event, the increase in group homogeneity can be rearranged by k-means.

IV. RESULTS AND DISCUSSION

Different customer profiles were segregated by making use of three (Arnottt & Pervan, 2008) clustering algorithms. The results were reviewed carefully and they were instrumental in identifying different groups. The six consumer profiles have been described with the Minitab results as reference point.

Cleanliness along with sanitation has emerged as the most important aspect of shopping when consumers go for stock-up shopping. Quality or fresh farm products or meat are the second most important point. Price does not seem to matter too much in anybody’s case. About 60% of all the shoppers rank price a little above mid-point. Even for people who did give a thought to

price, the price did not reach a score of much more than fifth or sixth. If you were to look beyond this, consumers do not seem to agree much on most things. In fact, they display marked differences.

V. CONCLUSION

Through the preceding analysis, it is evident that data mining has wide and varied applications for retail organizations, especially those that function on a large scale and those that have a well-developed web presence. While it does involves substantial expenditure, if implemented with care and after a thorough cost-benefit analysis, it is bound to show results. Any organization that operates in the dynamic markets of present times cannot afford to neglect the value of data mining in creating personalized shopping experiences for its customers and in optimizing its supply chain, operations and products – and thus, in enhancing organization performance.

Retail industries make profits out of efficiency from their supply chain managements. BI tools help in supporting their supply chain management which makes enterprises have better management over the supplies. When integrated in IMS or WMS, BI tools help in finding patterns, possible over production, shortages, and underproductions and so on; but in most cases quick response towards demand spike which are very important to catch. Originally, models that are complexly mathematical were used for logistical problems and supply chain management. Logistics are capable of affecting inventories specially to lack of precise forecasting and spikes in demands from slow response.

ACKNOWLEDGMENT

Grateful and indebted for the guidance and motivation provided by Dr. Savita Shiwani, Head of the department(I.T), who provided great technical assistance for the research paper. This paper is based on a dissertation for author's final year in dual course of M.tech.

REFERENCES

[1] Arnott, D. and Pervan, G. (2008). Eight key issues for the decision support systems discipline. Decision Support Systems, 44, 657–672.

[2] Amir F. Atiya, (2001) "Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results" IEEE Transactions on Neural Networks, vol. 12, no. 4.

[3] Chen, H., Chiang, R. and Storey, V. (2012). Business Intelligence and Analytics: From Big Data to Big Impact, MIS Quarterly, 36 (4), 1165-1188.

[4] Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews.

[5] Knowledge, Technology & Policy, 1(1), 104-126.

[6] Davenport, T. (2006). Competing on Analytics. Harvard Business Review. Davenport, T. (2010). Business Intelligence and Organizational Decisions. International Journal of Business Intelligence Research (IJBIR), 1, 1–12.

[7] Dean, J.J.W. and Sharfman, M.P. (1996). Does decision process matter? A study of strategic decision making effectiveness. Academy of Management Journal, 39, 368–392.

[8] Dien D. Phan, Douglas R. Vogel, (2010) "A Model of Customer Relationship Management and Business Intelligence Systems for Catalogue and Online Retailers", Information & Management,Vol. 47, Issue 2, Pages 69-77.

[9] Forgionne, G.A. (1999). An AHP model of DSS effectiveness. European Journal of Information Systems, 8, 95–106.

[10] Fitzsimons, M., Khabaza, T., and Shearer, C. (1993) "The Application of Rule Induction and Neural Networks for Television Audience Prediction" In Proceedings of ESOMAR/EMAC/AFM.

[11] Symposium on Information Based Decision Making in Marketing, Paris, pp 69-82.

[12] Godbole, S., Roy, S.(2008) "Text Classification, Business Intelligence, and Interactivity: Automating C-Sat Analysis for Services Industry" KDD'08, ACM Las Vegas, USA.

[13] Jun Lee, S., & Siau, K. (2001) "A Review of Data Mining Techniques" Industrial Management and Data Systems, 101/1, MCB University Press.

[14] Kanungo, S. (2009). The Centrality of Processes in IT-Enabled Decisions. In Proceedings of AMCIS 2009.

[15] MAIA Intelligence (2009) "Business Intelligence in Manufacturing".

[16] M. Crouhy, D. Galai, and R. Mark, (2000) "A comparative analysis of current credit risk models," J. Banking & Finance, vol. 24, pp. 59–117.

[17] Benbasat, I. and Nault, B.R. (1990). An evaluation of empirical research in managerial support systems. Decision Support Systems, 6 (3), 203–226.

[18] Mocanu, Aura-Mihaela, Litan, D., Olaru, S., & Munteanu, A. (2010) "Information Systems in the Knowledge Based Economy" WSEAS Transactions on Business and Economics, Issue 1, Vol. 7.

[19] Mintzberg, H., Raisinghani, D. and Theoret A. (1976). The structure of "unstructured" decision processes. Administrative Science Quarterly, 21, pp. 246–275.

[20] Nutt, P.C. (2008). Investigating the Success of Decision Making Processes. Journal of Management Studies, 45, 425–455.

[21] Papadakis, V., Thanos, I. and Barwise, P. (2010). Research on Strategic Decisions: Taking Stock and Looking Ahead? In Handbook of Decision Making, John Wiley & Sons.

[22] Park, Byung-Kwon, & Song, Il-Yeol (2011) "Toward total business intelligence incorporating structured and unstructured data" In Proceedings of the 2nd International Workshop on Business intelligence and the WEB (BEWEB '11), ACM, NY, USA.

[23] Phillips-Wren, G.E., Hahn, E.D. and Forgionne, G.A. (2004). A multiple-criteria framework for evaluation of decision support systems. Omega, 32, 323–332.

[24] Phua, C.et al. (2010) "A Comprehensive Survey of Data Mining-based Fraud Detection Research" Cornell University library, CoRR.

[25] Rajagopalan, N., Rasheed, A.M.A. and Datta, D.K. (1993). Strategic decision processes: Critical review and future directions. Journal of Management, 19, 349–384.

[26] Ramakrishnan, T. et al.(2011) "Factors Influencing Business Intelligence and Data Collection Strategies: An empirical investigation", Decision Support Systems.

[27] Shanks, G., Sharma, R., Seddon, P. and Reynolds, P. (2010). The Impact of Strategy and Maturity on Business Analytics and Firm Performance: A Review and Research Agenda.

[28] In Proceedings of ACIS 2010, 1-3 Dec 2010, Brisbane.

[29] Watson, H.J. (2010). Business Analytics Insight: Hype or Here to Stay? Business Intelligence Journal, 16 (1), 4–8.

[30] Watson, H.J., Goodhue, D.L. and Wixom, B.H. (2002). The benefits of data warehousing: why some organizations realize exceptional payoffs. Information & Management, 39, 491–502.

[31] Webster, J. and Watson, R.T. (2002). Analyzing the past to prepare for the future: Writing a literature review. MIS Quarterly, 26 (2), pp. xiii-xxiii.

AUTHORS

First Author – Prachi Agarwal, Department of Computer Science, Suresh Gyan Vihar University, Jaipur, India, Email: prachi091@gmail.com