SOUTHEAST ASIAN JOURNAL OF SCIENCE AND TECHNOLOGY

Volume 3, Issue 1, 2018
P-ISSN: 2672-2984
E-ISSN: 2672-2992
www.sajst.org

Customer Profile Analysis Using Time Series Model and Clustering

Julius Cesar O. Mamaril
Pangasinan State University
College of Computing Sciences

Abstract *- The use of data mining in sales and marketing processes of businesses has grown in the past years. It has been an advanced technique to bolster income while at the same time meet customer expectations by analyzing shoppers' buying behavior and profile [1]. This study focused on the use of customer segmentation and market basket analysis data mining techniques to determine customer profile, purchases, and product stock movements by analyzing sixteen months of sales transactions of a local business and subjecting such dataset in an open-source data mining and visual programing tool called RapidMiner Studio and Shrich MBA to produce different helpful visualizations. After carefully interpreting and validating the outputs of the data mining tool, the researchers were able to successfully identify the saleable products and its subsequent loyal customers and the period/time pattern in which it is saleable.*

Keywords - *Market-Basket Analysis, Affinity Model, Visualization Techniques, Apriori Algorithm, Design, Algorithms, Data Mining, Clustering Model*

INTRODUCTION

Sales transaction is one of the most important business processes in any forms of enterprise as it is the core business activity in which income can be generated. In this activity, goods available for sale are exchanged for equivalent cash payment from either first-time or frequent customers [2]. To have a successful enterprise, one should consider the different factors that directly and indirectly affect the sales transaction.

Among the factors that influence a sale transaction are the goods offered for sale and the customers [3]. Without goods to offer for sale, no matter how massive your customer base is, one cannot convert the number of customers into a significant figure of sales. Meanwhile, a high inventory of product stocks cannot fully guarantee an enormous digit of sales if customers are nowhere to be found or are present in the store (either in the physical or virtual/online store) but are not willing to make a purchase. In this scenario, a third factor of sales transaction should be at work influencing sales – marketing [4].

The relationship of products, customers and marketing activities and promotions is like a three-string rope, each string becomes integral in the strength and good shape of one another as a whole. Setting a good balance between the three strings will make a strong and physical wellness of the rope in general [5].

These three factors and the relationship of each with one another to produce higher sales revenues are the main aspects of the goals of this study which are: to determine the profile of customers in terms of demographics, gender, age bracket, job, and goods consumption or buying behavior using customer segmentation techniques and market basket analysis; and, identify the fast moving products and its corresponding season or time pattern using time series model of data mining.

RELATED WORKS

Data mining is a sub-field of computer science which uses digital computers to discover repetitions or arrangements in an arbitrary large input dataset to extrapolate information and convert the same in a well-structured resultset by way of the commonly overlapping methodologies of statistics, machine learning, database systems, and artificial intelligence [6].

The term data mining was first coined in the 1990s in the field of computer databases but prior to that the words data dredging or data fishing are already being used by statisticians in the 1960s to denote a bad procedure of establishing a data hypothesis prior to data collection [7]. In the late 90s, although data mining was first introduced in the world of machine learning and artificial intelligence in the form of a workshop on Knowledge Discovery in Databases or KDD started by Gregory Piatetsky-Shapiro, it became a hit in the press and business communities. Since then, knowledge discovery and data mining became synonymous with one another [8].

The KDD model is composed of 5 stages, namely: Selection - which involves the actual querying of datasets available in the database that will be subjected for data mining; Pre-Processing – the method of clearing unwanted or null values in each row of dataset; Transformation – the processes of exchanging numerical IDs to their related string values for each row of dataset or vice versa; Data Mining – the systematic application of algorithms in order to search for pattern among data; and Interpretation or Evaluation – the manner of visualizing data mining results using statistical graphs or charts, inferring significant information from it and evaluating the validity and usage of the newly collected information [9]. Fig. 1 shows the graphical representation of the KDD stages.

SOUTHEAST ASIAN JOURNAL OF SCIENCE AND TECHNOLOGY

Volume 3, Issue 1, 2018
P-ISSN: 2672-2984
E-ISSN: 2672-2992
www.sajst.org

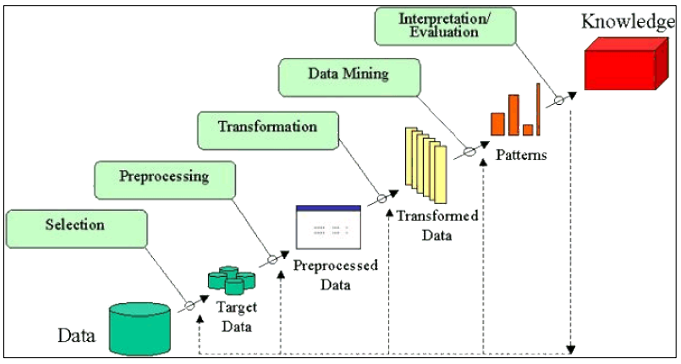

Fig. 1. Knowledge Discovery in Databases (KDD) Model

Data mining includes six common classification of tasks which are: anomaly detection which is about the investigation of interesting data errors; association rule learning which focuses on finding correlations among known variables; clustering which involves the use of unknown data structure to discover similar structure and groups; classification which refers to the task of simplifying or summarizing known data properties that can be applicable create new data; regression pertains to the search for data models in the form of algorithms or functions to estimate data relations in the least possible error; and summarization which involves the use of statistical techniques to visualize data in the form of graphs or charts [10].

Among the data mining classes of tasks, association and clustering have been utilized mostly to favor business activities and stakeholders' decisions. Association task or association rule learning, most of the times referred to also as dependency modelling, involves finding relationships between known attributes. This model is famously known in the business community as market basket analysis [11]. In this type of analysis, percentages of support and confidences are monitored and given weight. Confidence refers to the rate of probability that when a customer buys Product X, he will buy Product Y[12] while support refers to the ratio between all sales transactions with Product X over the total number of sales transactions [13].

On the other hand, clustering task of data mining involves the discovery of similar structures and groups in an arbitrary voluminous dataset using known attributes from such dataset [14].

Patterns, similarities, relationships and correlations discovered during data mining tasks are subjected to different data visualization techniques in order for non-technical people to comprehend, interpret, and infer knowledge from such results. Different charting and graphing methods are utilized to give data a face [15].

METHODOLOGY

The researchers employed the KDD model in this study in order find patterns and relationships among the sixteen (16) months of sales transactions of a local convenience store from January 1, 2016 to April 30, 2017. It is consisted of 24,293 sales records corresponding to 134,554 sales details of 2,071 unique products which are categorized into 12 major categories and are subcategorized into 76 subcategories. The data was provided in a spreadsheet format (Microsoft Excel .xlsx) which also included a worksheet of the grocery's 100 loyal customers where profile such as age, address, occupation, gender, total points earned, and retail store ownership are also included.

The customers of the subject business is grouped into two – loyal members who availed of loyalty membership card, and walk-in customers. The latter cannot be accounted in this research as to any of their profile since point-of-sale software do not record any of their profile during sales payment. However, the dataset provided identifies whether a sales transaction is from a loyal member or from a walk-in customer.

In order to pre-process the dataset thoroughly, the researchers created a MySQL schema and tables for each spreadsheet worksheet whose fields are similar to those of the columns of each MS-Excel worksheet. SQL queries were developed to generate datasets needed according to each research objective. The query resultsets are then exported in a comma-separated text file (.csv format).

The researchers used RapidMiner Studio version 7.5 and Shrich MBA version 1.0 developed by Shrich Enterprises as data mining software tools. RapidMiner Studio is a popular open-source visual programming software which has built-in dataset pre-processing operators and algorithms for rapidly creating predictive-analytic workflows. It has also several basic integrated visualization features and an advanced charting module [16]. Meanwhile, Shrich MBA by Shrich Enterprises is a data mining tool exclusively developed to support this research and to perform market basket analysis without requiring huge and expensive hardware resources.

After pre-processing the datasets, they were subjected to the visualization features of RapidMiner Studio. Meanwhile, using Shrich MBA, the researchers connected to the MySQL database and perform the Market Basket Analysis. Frequency, percentages of support and confidence, and correlated items were noted and results were given form and faces using the tool's embedded visualization network graph feature.

SOUTHEAST ASIAN JOURNAL OF SCIENCE AND TECHNOLOGY

Volume 3, Issue 1, 2018
P-ISSN: 2672-2984
E-ISSN: 2672-2992
www.sajst.org

The time series analysis technique has been employed to validate the results of market basket analysis by comparing sales transactions that occurred from one period of a particular year to the same period on the succeeding year. If a high degree of similarity occurs, then the results of the market basket analysis can be inferred as highly accurate.

RESULTS AND DISCUSSION

4.1 Customers Profile

Fig. 2 presents the sales transactions that occurred from January 1, 2016 to April 30, 2017 among loyal members and walk-in non-member customers. Among the 24,293 sales records from the said period, 18,926 (77.91%) are from non-members and 5,397(22.09%) are from loyal members. It can be noted from the graph that although frequency of sales among members are fewer than those of walk-in customers, in terms of the amount being spent, loyal customers consistently drive more sales to the convenience store.

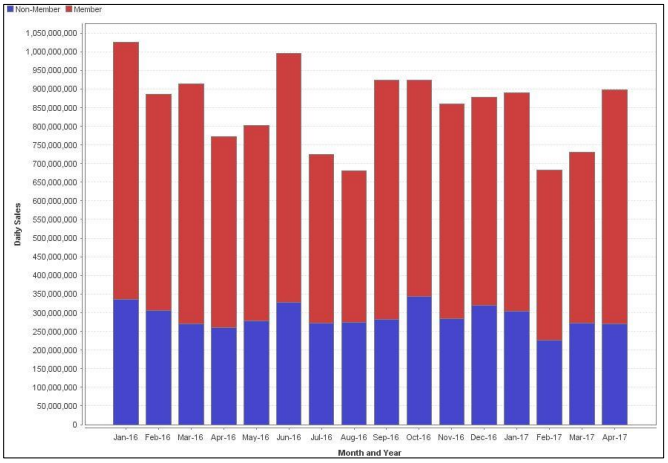

Fig. 2. Sales Transactions between members and non-members.

Fig. 3 exhibits the daily average sales of members and non-members in terms of their address. Non-members' address is labeled with "N/A" in the x-axis. The graph is clear that majority of sales transactions are coming from members who reside in Lingayen, Pangasinan.

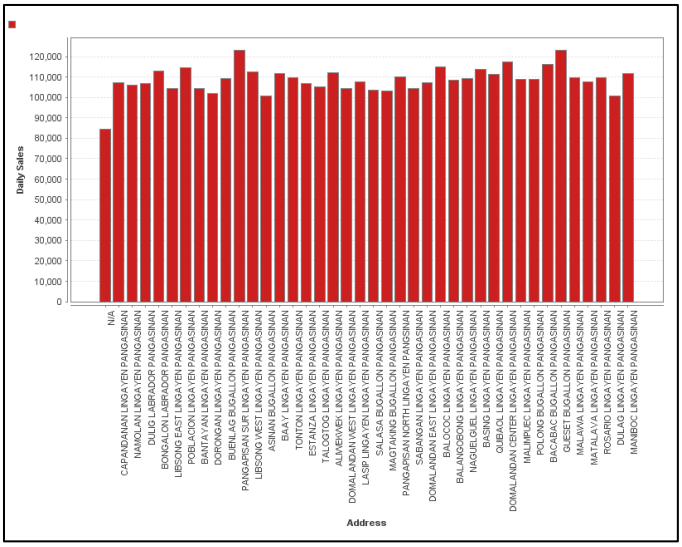

Fig. 3. Daily average sales of members and non-members in terms of address.

In terms of Gender, majority of the members are males (52) in contrast to females (48) as shown in Fig. 4.

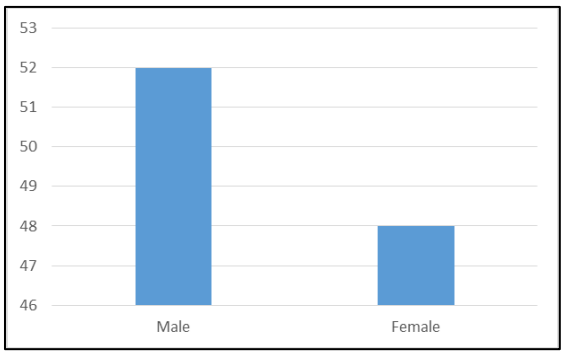

Fig. 4. Graph of Members in terms of gender.

Fig. 5 showcases the average daily sales of loyal members in terms of their age bracket. The graph is indicative that majority of the loyal members belong to the age group between 35-50 years old. (The isolated blue scattered points refer to the non-members automatically aged 0 by the POS software). The image also portrays that loyal members under the said age group spent a minimum of P17.00, maximum of P8,300.25 and average sales transaction per day of P4,309.00.

The classification of occupation of members are shown in Fig. 6. It shows that majority of members are in the business sector, government and education (13 frequency count each forming 39% of the total members). It is followed by office clerks, drivers, housekeepers and fishermen (having 8-11 frequency count forming 49% of the total members).

SOUTHEAST ASIAN JOURNAL OF SCIENCE AND TECHNOLOGY

Volume 3, Issue 1, 2018
P-ISSN: 2672-2984
E-ISSN: 2672-2992
www.sajst.org

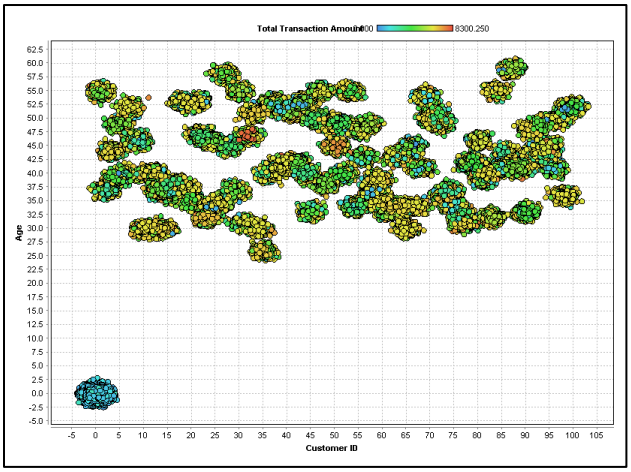

Fig. 5. Average daily sales of loyal members in terms of age bracket.

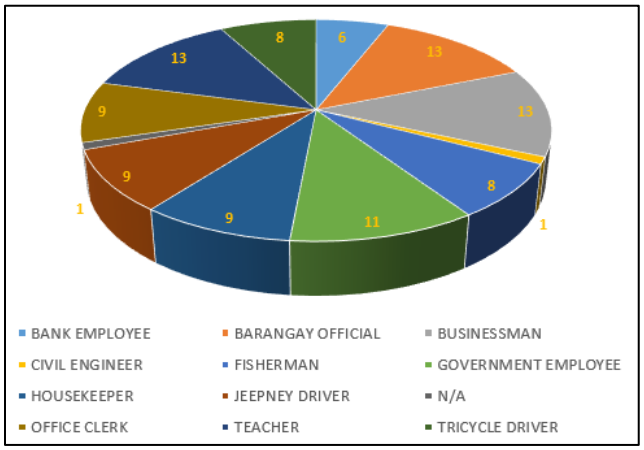

Fig. 6. Chart of Members in terms of Occupation.

Fig. 7 presents the average daily sales transactions of members in terms of occupation. The figure is indicative that daily sales volume is coming from members whose jobs are businessmen, barangay officials and teachers (Non-members is indicated with N/A legend).

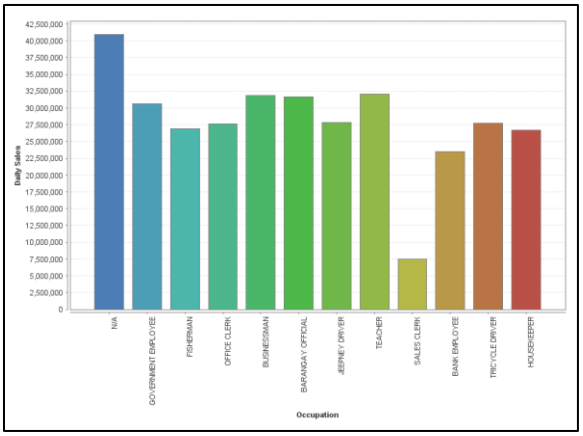

Fig. 7. Average daily sales of members in terms of job.

4.2 Customers Buying Habit.

The researchers employed the Market Basket Analysis technique in order to identify the buying habit of a local convenience store to find association/affinity and relationship of product items being bought. Under the Market Basket Analysis, all sales transactions are tabulated and scored per product through frequency count. The higher the frequency means the product is being regularly bought.

Fig. 9 presents the ranking of the top 20 fast moving products identified together with each product's corresponding frequency, support, and pairing count. It can be noted that fast moving products are daily household needs.

The product's support refers to the number of times the product has been involved in a sales transaction (frequency) divided by the total number of sales transactions. Meanwhile, a product's pairing refers to the number it has become part of an itemset. If the product's frequency is greater than the pairing, their difference is the number of times the product has been bought as a solo item, conversely, if they are equal, it means that the product is always bought in pair with other products.

SOUTHEAST ASIAN JOURNAL OF SCIENCE AND TECHNOLOGY

Volume 3, Issue 1, 2018
P-ISSN: 2672-2984
E-ISSN: 2672-2992
www.sajst.org

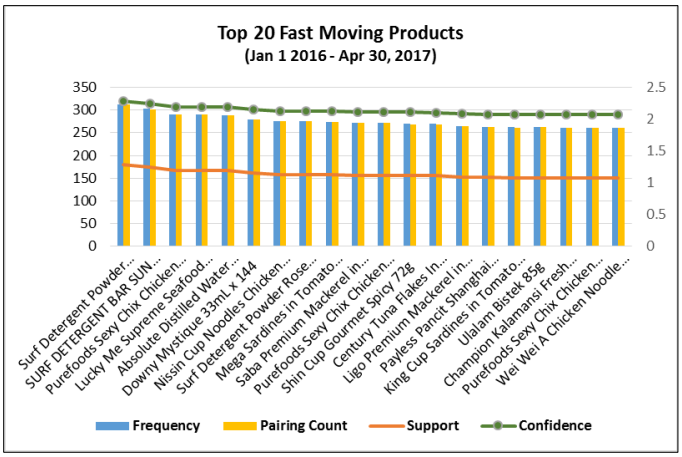

Fig. 9. Product frequency of the top 20 fast moving products from January 1, 2016 to April 30, 2017.

An itemset is a group or set of items being bought by customers which creates a unique binary pattern in terms of all the possible combination of the products that can be bought in a store where 1 corresponds to the items in the itemset bought and zero to those products which are not bought. These itemsets or their binary representations are referred to as organic itemsets as the pairing of such items are unique and inherent from the customers themselves.

Fig. 10 shows the table of organic itemsets preferred by customers in a local convenience store. It is remarkable that only 9 itemsets have been purchased twice. This is due to the very low items available in an itemset. With around 2,071 available products for sale, the probability of unique itemsets can be up to 22071 - 2.

Aside from identifying the fast moving products, the researchers found out the related products bought by customers together with the fast moving items. Figure 11 presents the network graph of the top 10 fast moving products (represented by their ProductID) with their correlated products whose frequencies with the fast moving items are more than 50.

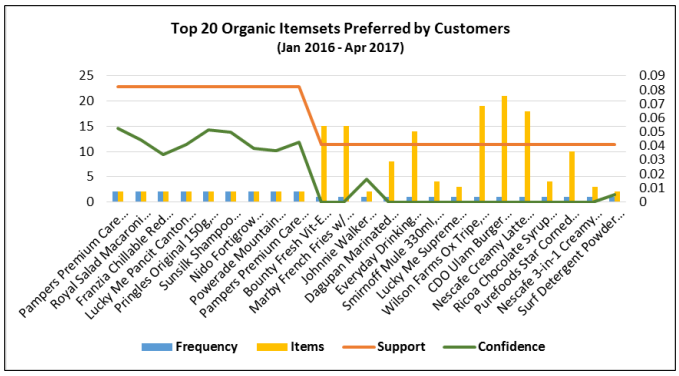

Fig. 10. Organic itemsets reflecting frequency, support, confidence and items count.

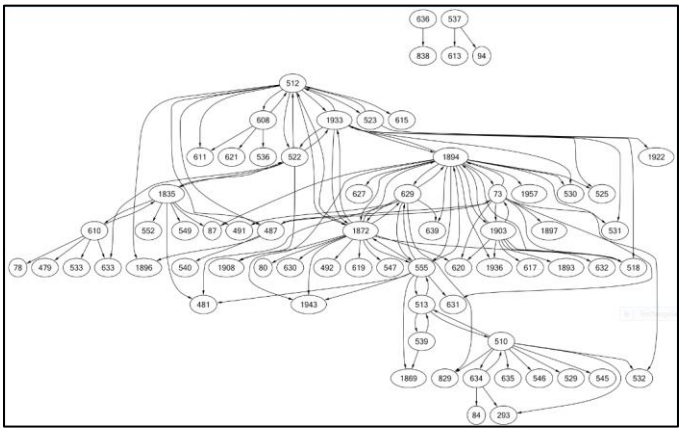

Fig. 11. Network or affinity graph of top 10 fast moving products (represented by their respective ProductID) connected thru line with their correlated products.

Fig. 11 depicts the interconnection of products to their respective correlated products. A node (a circle or oblong) represents a product labeled by its corresponding ProductID while connecting lines refer to the relationship of a node to other nodes (correlation of a product to another product).The line's arrow determine which product is the primary product and which is secondary. Arrows (→) denote that the node where the arrow is pointing is the secondary product and the node from the other end is the primary product (Primary product is the first product being bought and the secondary ones are those being bought after deciding to buy first the primary product).

It can be inferred then from Figure 11 that nodes 1872, 1894, and 512 (which corresponds to the ProductIDs of Surf Detergent Powder Cherry Blossoms, Surf Detergent Bar Sun Fresh and Purefoods Sexy Chix Chicken Chunks, respectively) are products with the most correlated secondary products as

SOUTHEAST ASIAN JOURNAL OF SCIENCE AND TECHNOLOGY

Volume 3, Issue 1, 2018
P-ISSN: 2672-2984
E-ISSN: 2672-2992
www.sajst.org

many nodes are inter-connected with them. Conversely, these products are also the top 3 fast moving products.

4.3 Time Series Technique Validation.

To validate that the volume and frequency of sales transactions that occurred in a certain period as compared to other period, the researchers utilized the Time Series Technique to compare and contrast the sales transactions from January 1, 2016-April 30, 2016 from those sales transactions during January 1, 2016-April 30, 2017.

Fig. 12 presents the top 20 fast moving items from January 1 to April 30, 2016 in comparison to the top 20 fast moving items from January 1 to April 30, 2017. The table is indicative that majority of top 20 saleable items in the first period appeared also in the second period and almost maintained a ranking near to its previous rank. Majority of the most frequently bought products belong to the daily household needs product classifications.

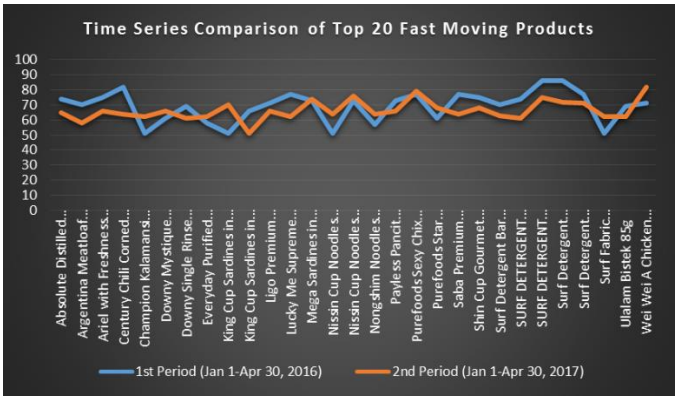

Figure 12. Time Series Comparison of Top 20 Fast Moving Products from Jan 1-Apr 30, 2016 (1st Period) and Jan 1-Apr 30, 2017 (2nd Period).

To provide a visual perspective of the contrast between the two periods' sales transactions amount, Fig. 13 portrays a graphical validation that sales transactions peaked during the month of January from both periods and will plateau on succeeding months.

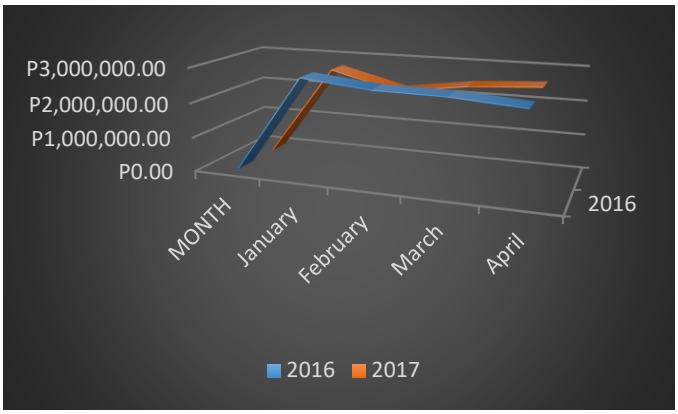

Fig. 13. Series graph of sales comparison between the first 4 months of 2016 and 2017.

CONCLUSIONS

The use of data mining software is an effective tool to perform market basket analysis and identify customer buying habits, fast moving items, correlated products, and visualize sales patterns together with customer profile. The researchers found out that most items bought by customers in a local convenience store are composed of basic daily household needs. This buying behavior has been consistent, validated and proven using time series technique by comparing months of sales transactions belonging to the same period but from different years.

The profile of loyal member customers has been very helpful as parameters in terms of clustering sales transactions and providing clear visualizations and inferred impact as compared to those walk-in customers. The patterns and relationships discovered in this study conform to the general acceptable standards and real day-to-day shopping scenarios.

REFERENCES

[1] Maheshwari, Anil (2014). Business Intelligence and Data Mining. Business Expert Press. ISBN: 9781631571206.

[2] Arganda, Amelia M. & Herrero, Carmen C. (2014). Accounting Principles 1 5th Edition. National Bookstore, Inc. ISBN: 9786218016163.

[3] Simmons, Gene (2014). Me, Inc.: Build an Army of One, Unleash Your Inner Rock God, Win in Life and Business. Dey Street Books. ISBN10: 0062322613/ISBN-13: 978-0062322616.

Volume 3, Issue 1, 2018
P-ISSN: 2672-2984
E-ISSN: 2672-2992
www.sajst.org

[4] McCarthy , Breda (2014). Strategy, Marketing Plans and Small Organisations. BookBoon. ISBN: 978-87-403-1298-0.

[5] Brooks, John (2014). Business Adventures: Twelve Classic Tales from the World of Wall Street. Open Road Media. ISBN-10: 1497644895 / ISBN-13: 978-1497644892

[6] Hastie, Trevor; Tibshirani, Robert; and Friedman, Jerome (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction 2nd Edition. Springer Science and Business Media. ISBN: 0387848584 / 9780387848587.

[7] Klimberg, R. and Lawrence, Kenneth D. (2015). Contemporary Perspectives in Data Mining, Volume 2. Information Age Publishing. ISBN:1681230879/978-1681230870.

[8] Tan, Pang-Ning; Steinbach, Michael; and Kumar, Vipin (2005). Introduction to Data Mining. Pearson. ISBN-10: 0321321367 / ISBN-13: 978-0321321367.

[9] Fayyad, U.M., Smyth, P. and Uthurusamy (1996). Advances In Knowledge Discovery and Data Mining Revised Edition. American Association for Artificial Intelligence Press. ISBN: 0262560976/ 9780262560979.

[10] Han, Jiawei (2006). Data Mining: Concepts and Techniques. Elsevier. ISBN: 1558609016/781558609013.

[11] Poncelet, Pascal (2007). Data Mining Patterns: New Methods and Applications. Idea Group Reference. ISBN-10: 1599041626/ISBN-13: 978-1599041629.

[12] Giudici, Paolo and Figini, Silvia (2009). Applied Data Mining for Business and Industry, Second Edition. John Wiley and Sons Ltd. ISBN: 9780470058862 / 9780470745830.

[13] Blattberg, Robert C.; Kim, Byung-Do; and Neslin, Scott A.(2008). Market Basket Analysis. Springer New York. 978-0-387-72578-9 / 978-0-387-72579-6

[14] Provost, Foster and Fawcett, Tom (2013). Data Science for Business. O'Reilly Media, Inc. 144937428X, 9781449374280.

[15] The Interaction Design Foundation. Article: Data Visualization for Human Perception. https://www. interaction-design.org/literature/book/the-encyclopedia-of-human-computer -interaction-2nd-ed/data-visualization-for-human-perception. Retrieved 2017-05-04.

[16] RapidMiner. Webpage: RapidMiner Studio Lightning Fast Data Science. https://rapidminer.com/products /studio/. Retrieved 2017-05-04.