Analysis of Sales Data for Decision Making Using Association Rules

O. O. Oladimeji[1]; O. O. Catherine[2]; G. Mutiu[3]

[1.3]Department of Computer Science,
Federal Polytechnic Ile-Oluji,
Ondo State, Nigeria.
e-mail: Olasehindeolayemi@yahoo.com[1]; mutgan2002@yahoo.com[3]

[2]Department of Computer Science,
Joseph Ayo Babalola University,
Ikeji-Arakeji, Nigeria.
e-mail: Olafunyemi@yahoo.com[3]

***Abstract* - This paper aimed at discovering pattern of items (products) that are bought together by customers using one year sales dataset of a major distributor of Golden Penny products in Nigeria using association rule. An apriori algorithm was implemented to generate strong and interesting rules from the dataset. Result from this study would help in developing promotional strategies so that items with low level of patronages but have the capacity to enhance the revenue of the organisation can be placed adjacent to other products/items that are likely to be bought together with them on the shelves. This arrangement, if employed, could assist and remind customers to buy those items together.**

Keywords - *Data mining; Association rule; Classification; Algorithm.*

I. INTRODUCTION

Due to the stiff competition in market place, retailers are looking for ways to have a competitive edge over competitors through technology. Many organisations today keep huge amount of historical data and daily generate a lots of transaction data. Transaction data are recorded business events that have monetary impact on an organisation financial statement. These data are updated regularly and they are key information that supports the operational activities of an organization.

Daily sales record are kept by retailer without knowing how to extract important and meaningful information from it to gain competitive advantage.

The market basket is defined as an item-set bought together by a customer on a single visit to a store, market basket analysis, also known as affinity analysis, has emerged as the next step in the evolution of retail merchandising and promotion.. Market basket analysis have been used to determine customer interest profile and interests on particular products for one to-one marketing [1], and to determine purchasing patterns in a multi-store environment to improve the sales [2].

Market basket analysis has been intensively used in many companies to discover product associations and base a retailer's promotion strategy on them. Market Basket Analysis (Association Analysis) is a mathematical modeling technique based upon the theory that if you buy a certain group of items, you are likely to buy another group of items. It is used to analyze the customer purchasing behavior and helps in increasing the sales and maintain inventory by focusing on the point of sale transaction data.

Data mining is *the extraction of hidden predictive information from large dataset such as the transaction data.* Association rule mining is one of the most important and well researched techniques of data mining [3]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the sales dataset. it finds all frequent item-sets and then generate strong association rules that satisfy both a minimum support threshold and a minimum confidence threshold from the frequent item-sets, the rules generated can be used to predict the occurrence of an item based on the occurrences of other items in the sales record.

Mining interesting rules by Association and Classification Algorithm was put forth by [4]. The main intention in data mining is to disclose hidden knowledge from dataset, the drawback of this techniques is that it generates a lot of irrelevant and rules that are not interesting, numerous measures such as confidence, support, lift, information gain, etc., have been suggested to remove this drawback.

Trnka [5] uses Data Mining Methods for Market Basket Analysis. Reference [7] compares the percentage of shoppers buying a certain product and the percentage of all total sales generated by this product. By making such comparisons, one can easily find out the leading products and what is their share of sales. Chenqiz and Shichao [8] reviews the most well known algorithm for producing association rules - Apriori and discuss variants for distributed data, inclusion of constraints and data taxonomies. The review ends with an outlook on tools which have the potential to deal with long item-sets and considerably reduce the amount of (uninteresting) item-sets returned.

II. MATERIALS AND METHODS

A. *Description of Dataset*

The dataset used for this research is a transactional data of Golden penny products, a total of 3405 transaction records were obtained from six months sales record, Table 1 shown sample of the raw dataset. Each row in the table represents a single transaction, the Transaction Detail columns are the columns needed and were used for our analysis. The full description each item in the transaction detail column is show in Table 2.

Table 1: Sample of the Raw Sales Record

Transaction ID	Date	Transaction Details						Amount (in ₦)
R1005467001	12-04-2016	1kgs	2kgs	5kgs	10kgs	1kgw	5kgw	14770
R1005467002	12-04-2016	1kgs	2kgs	1kgw	Spag	70g		14350
R1005467003	12-04-2016	1kgs	5kgw	1kgw	Spag	70g		12600
R1005467004	12-04-2016	1kgs	2kgs	5kgs	Spag	70g		13300
R1005467005	12-04-2016	2kgs	spag	70g				8450
R1005467006	13=04-2016	2kgs	spag					6950
R1005467007	13-04-2016	1kgs	2kgs	5kgs	10kgs	Spag		14760
R1005467008	13-04-2016	1kgs	2kgs	5kgs	10kgs	1kgw		13520

Table 2: Description of items in the transaction detail column of Table 1

Items	Description	Items	Description	Items	Description
1kgs	1kg semo vita	2kgw	2kg Golden vita	Twist	500g Twist
2kgs	2kg semo vita	5kgw	5kg Golden vita	Stwist	200g Twist
5kgs	5kg semo vita	10kgw	10kg Golden vita	Macaroni	500g Macaroni
10kgs	10kg semo vita	500gw	500g Golden vita	70g	70g Noodles
500gs	500g semo vita	Spag	500g Spagethi	100g	100g Noodles
1kgw	1kg Golden vita	SSpag	200g Spagethi	150g	150g Noodles

A. *Association Rule*

Association rule is a data mining technique used to discover associations among sales dataset items. Data mining refers to extracting knowledge from large amount of dataset. It identifies relationship between a large set of data items.

Following the original definition by Agrawal and Sikrant [6], the problem of association rule mining is defined as follow: let
be a set of transactions called the sales record (*database*)

$$I = \{i_1, i_2, \ldots\ldots\ldots, i_n\} \quad (1)$$

be a set of *n* products called *items. let*

$$D = \{t_1, t_{2,\ldots\ldots\ldots\ldots}, t_n\} \quad (2)$$

Each transaction in D has a unique transaction ID and contains a subset of the items in I. A *rule* is defined as an implication of the form

$X \rightarrow Y \quad X, Y \subseteq I \quad$ and $\quad X \cap Y = \emptyset$. (3)
(where X, Y are product items)

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest are used. The best-known constraints are minimum thresholds on Support and Confidence.

The *support* supp (X) of an item set X is defined as the proportion of transactions in the dataset which contain the item set.

supp(X) = item set X / *Total Sales Record* (4)

Support denotes the frequency of the rule within transactions. A high value means that the rule involve a great part of the sales database. Equation 4 can also be written as

Support $(X \Rightarrow Y \ [s, c \ ]) = P(X \cup Y)$ (5)
(where s, c is the support and confidence level)
.

The *confidence* of a rule is defined as

$$\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X). \quad (6)$$

Confidence denotes the percentage of transactions containing A which contain also B. It is an estimation of conditioned probability .

Confidence $(X \Rightarrow Y \ [s, c \ ]) = p(X \mid Y) =$ sup $(X \Rightarrow Y)$ / sup(X). (7)

Given a frequent item set {1kgS, 1kgW, Spag}, the following rules are possible:

Rule1 . {1kgS ⇒ 1kgW, Spag}
Rule2 . {1kgW ⇒ 1kgS, Spag}
Rule3 . {Spag ⇒ 1kgW, 1kgS}
Rule4 . {1kgS, 1kgW ⇒ Spag}
Rule5 . {1kgW, 1kgS ⇒ Spag}
Rule6 . {Spag, 1kgW ⇒1kgS}
Rule7 . {1kgS ⇒ 1kgW}
Rule8 . {1kgS ⇒ Spag}
Rule9 . {1kgW ⇒ Spag}

III. Analysis and Results

The two packages '**arules**' and '**arulesviz**' in R software were used to implement the association rule. Minimum support of 0.0029 signifying minimum of one rule out of every 10 transactions and confidence of 90% were selected to generate strong rules, a total of 440 rules were generated , The rule were pruned to remove redundant (duplicate) rules, a total of 399 redundant rules were pruned, leaving behind 47 distinct strong rules. (Table 3). We generated recommended rules from the rule set, Table 4 shows the six (6) recommended rule from the rule set . Visualisation of the 47 distinct strong rules are shown in Figure 1.

IV. Discussion of Results

From the result of association rules, the following was inferred;

- 100% of 500g semo and 200g twist also buys 2kg semo
- 100% of 1kg semo and 500g spagetti also buys 1kg wheat
- 100% of 2kg semo and 200g spagetti also buys 2kg wheat
- 100% of 1kg semo and 500g spagetti also buys 2kg wheat
- 87% of 10kg semo and 200g twist also buys 500g semo
- 98% of 1okg and 500g semo also buys 2kg and 5kg semo
- 96% of 2kg and 500g semo with 500g spagetti and 100g noodles also buys 10kg semo and 70g noodle

Table 3. Sampled 47 distinct strong rules

Sample Rules Generated and Pruning for supp=0.0029 and conf=.09

	lhs		rhs	support	confidence	lift
[1]	{X10kgsemo=talia}	=>	{X2kgwheat=}	0.06566784	0.9809524	1.150803
[2]	{X2kgsemo=10kgsemo}	=>	{X2kgwheat=}	0.06949315	1.0000000	1.173149
[3]	{X10kgsemo=1kgwheat}	=>	{X2kgwheat=}	0.08128785	0.9172662	1.076090
[4]	{X10kgsemo=}	=>	{X500gsemo=}	0.09276379	1.0000000	1.521339
[5]	{X10kgsemo=}	=>	{X2kgwheat=}	0.09276379	1.0000000	1.173149
[6]	{X10kgsemo=70gnoodle}	=>	{X2kgwheat=}	0.09977686	0.9968153	1.169413
[7]	{X2kgsemo=500gspaghetti}	=>	{X2kgwheat=}	0.13069812	0.9903382	1.161814
[8]	{X10kgsemo=500gspaghetti}	=>	{X2kgwheat=}	0.20051004	0.9797508	1.149393
[9]	{X500gsemo=}	=>	{X2kgwheat=}	0.65731591	1.0000000	1.173149
[10]	{X10kgsemo=, X500gsemo=}	=>	{X2kgwheat=}	0.09276379	1.0000000	1.173149
[11]	{X10kgsemo=, X2kgwheat=}	=>	{X500gsemo=}	0.09276379	1.0000000	1.521339
[12]	{X10kgsemo=70gnoodle, X500gsemo=}	=>	{X2kgwheat=}	0.07969398	1.0000000	1.173149
[13]	{X2kgsemo=500gspaghetti, X500gsemo=}	=>	{X2kgwheat=}	0.11762831	1.0000000	1.173149
	.....					
	.....					
	.....					
[44]	{X2kgsemo=500gspaghetti, X2kgwheat=}	=>	{X500gsemo=}	0.11762831	0.9000000	1.369205
[45]	{X2kgsemo=5kgsemo, X10kgsemo=500gspaghetti}	=>	{X2kgwheat=}	0.06216130	0.9512195	1.115922

Table 4. Recommended Rules

Recommendation Rules

	lhs		rhs	support	confidence	lift
[1]	{X2kgsemo=10kgsemo}	=>	{X2kgwheat=}	0.06949315	1	1.173149
[2]	{X10kgsemo=}	=>	{X500gsemo=}	0.09276379	1	1.521339
[3]	{X10kgsemo=}	=>	{X2kgwheat=}	0.09276379	1	1.173149
[4]	{X500gsemo=}	=>	{X2kgwheat=}	0.65731591	1	1.173149
[5]	{X10kgsemo=,X500gsemo=}	=>	{X2kgwheat=}	0.09276379	1	1.173149
[6]	{X10kgsemo=,X2kgwheat=}	=>	{X500gsemo=}	0.09276379	1	1.521339

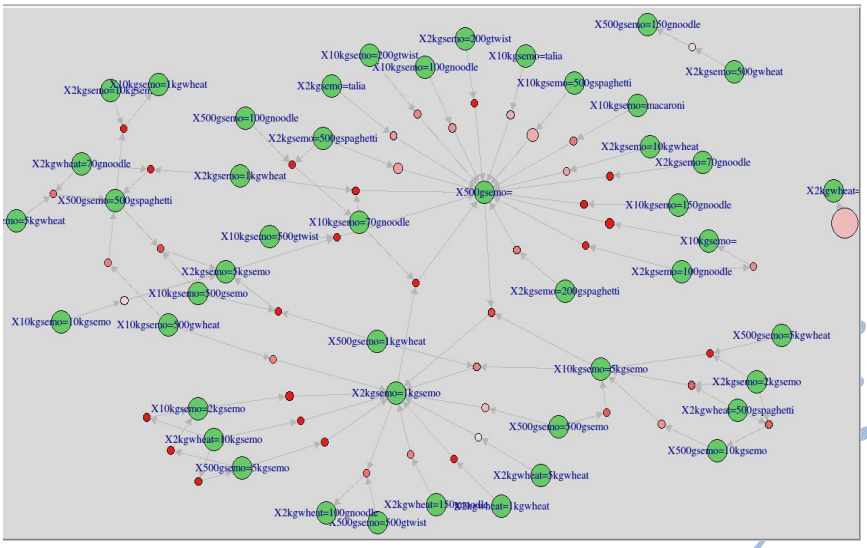

Figure 1: Visualization of the 47 Distinct Rules

V. CONCLUSION

In conclusion, our analysis has shown that Association rule can be used to i.) manage placement of goods on the shelve for ease of access by customers, ii.) manage inventory, iii.) increase sales, iv.) influence customer buying pattern, and v.) identify products that needs promotional strategies to boost their patronage.

VI. RECOMMEDATION

From the analysis carried out in this study, the following recommendations can be offered;

1. Products that appeared together in the pruned rules **must** never be allowed to get out of stock and **must** be arranged closely together to help increase customer satisfaction and increase sales.
2. Products that did not appear in the pruned rules must be promoted.

REFERENCES

[1] Weng, S, S., Liu, J., L.(2004) "Feature-based recommendations for one-to-one marketing" Journal of Expert Systems with Applications", Vol. 26, 2004, pp. 493-508.

[2] Chen, Y.-L., Tang, K., Shen, R, J., Hu, Y., H.(2005) "Market basket analysis in a multiple store environment," Elsevier Journal of Decision Support Systems, Volume 40, Issue 2, August 2005, Pages 339–354 www.sciencedirect.com/science/article/pii/S0167923604000685

[3] Agrawal, R., Imielinski, T., and Swami, A. N. (1993). "Mining association rules between sets of items in large databases". In Proceedings of the 1993 ACM SIGMOD. *International Conference on Management of Data, Washington* DC (USA) pp 207-216.

[4] Yanthy W., Sekiya T.,Yamaguchi K.. (2009) "Mining Interesting Rules by Association and Classification Algorithms," FCST '09. *Fourth International Conference on Frontier of Computer Science and Technology*, Pp. 177 – 182, 2009.

[5] Trnka, A. (2010) ,"Market Basket Analysis with Data Mining Methods", International Conference on Networking and Information Technology (ICNIT), Pp. 446 - 450, 2010.

[6] Agrawal R., Srikant R.,(1994) "Fast algorithms for mining association Rules", Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994, pp. 478– 499

[7] Julander, (1992) "Basket Analysis: A New Way of Analysing Scanner Data", International Journal of Retail and Distribution Management, Vol. 20 Iss: 7 DOI http://dx.doi.org/10.1108/09590559210022362

[8] Chenqiz, Z., Shichao, Z., (2008) "Association Rules Mining" , Lecture Notes in Computer Science, Paperback – June 2008 2003, Pages 226 – 234.