Application of Association Rule Mining in a Clothing Retail Store

Akshay Jain, Shrey Jain, and Nitin Merh

1 Introduction

Retailing is a sale of goods or commodities in small quantities directly to the customers. Retailers provide important functions that increase the value of the products and services they sell to customers. These value-creating functions are providing assortment of products and services, breaking bulk, holding inventory and providing services and experiences (Avcilar et al. 2014). Retail industry is a business which has a high market concentration but has the potential to generate high profits if managed properly with the help of certain tools. The tools which contribute to the success of retail store help in understanding the consumer behavior, buying pattern, relativity between the products. Consumer behavior is a consumer activity in deciding to purchase, use, as well as consume the purchased goods and services including the customer factors which can give a rise to their decisions whether to purchase and use products (Kurniawan et al. 2018).

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large databases in order to find useful patterns that might otherwise remain unknown (Tan et al. 2016). Association rule mining, first introduced by Agrawal et al. (1993), is useful for discovering interesting relationships hidden in large datasets. The uncovered relationship can be represented in the form of association rule or sets of frequent items (Tan et al. 2016). Market basket analysis is a technique that analyzes customer

A. Jain · S. Jain
SVKM's Narsee Monjee Institute of Management Studies (NMIMS), Indore, India
e-mail: akshay2602jain@gmail.com

S. Jain
e-mail: shrey.1425@gmail.com

N. Merh (✉)
Jaipuria Institute of Management Indore, Indore, India
e-mail: nitinmerh0812@gmail.com

A. K. Laha (ed.), *Applied Advanced Analytics*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-981-33-6656-5_9

buying habits with the help of associations between the products that the customer place in their shopping basket. This will help the retailers to develop marketing strategies by gaining insights into the items that are frequently purchased with each other.

Business analytics is a scientific process of transforming data into insight for making better decisions. Business analytics is used for data-driven or fact-based decision making which is often seen as more objective than other alternatives. It is of three types: descriptive, predictive and prescriptive; out of these, this research is focusing on predictive analytics (Cochran et al. 2015). Predictive analytics includes variety of statistical techniques from modeling, data mining and machine learning that analyze current and historical data to make predictions about future outcomes. Prediction helps organization in making right decision at right time by right person as there is always time lag between planning and actual implementation of the event.

"Try Us," a new retail clothing store located in Indore, Madhya Pradesh, is selected as the store under study. It sells multiple brands for men. The organization wants to expand its business and is planning to open another store on a larger scale.

Apriori algorithm is used for mining datasets for association rules. The name Apriori is used because it uses prior knowledge of frequent item properties. Apriori algorithm uses bottom-up approach where frequent subsets are extended one by one. In Apriori algorithm, iterative approach is used where k frequent item sets are used to find $k + 1$ item sets. It is generally used in market basket analysis as it is useful in finding the relationship between two products. Apriori makes some assumptions like:

- All subsets of a frequent item set must be frequent.
- If an item set is infrequent, all its supersets will be infrequent.

While applying Apriori algorithm, the standard measures are used to assess association rules. These rules are the support and confidence value. Both are computed from the support of certain item sets. For association rules like $A \rightarrow B$, two criteria are jointly used for rule evaluation. The support is the percentage of transactions that contain $A \bigcup B$ (Agrawal et al. 1993; Avcilar et al. 2014). It takes the form support $(A \rightarrow B) = P(A \bigcup B)$. The confidence is the ratio of percentage of transactions that contain $(A \bigcup B)$ to the percentage of transactions that contain A. It takes the form confidence $(A \rightarrow B) = P(B \backslash A) =$ support$(A \bigcup B)$/support (A). Rules that satisfy both a minimum support threshold (min_sup) and minimum confidence threshold (min_conf) are called strong (Avcilar et al. 2014).

Primary objective of the study is to understand the buying pattern of the customer and to study and analyze proper basket (combos) of products for cross-selling and upselling. Another objective is to explore the relativity between the products for applying the optimal design layout for the clothing retail store.

1.1 Literature Review

Research work done by Kurniawal et al. (2018) suggests that market basket analysis performs better results over association rule mining using Apriori algorithm. The research done by Tatiana et al. (2018) on a study of integrating heterogeneous data sources from a grocery store based on market basket analysis, for improving the quality of grocery supermarkets, shows positive results for increasing the performance of the store.

Szymkowiak et al. (2018) propose theoretical aspects of market basket analysis with an illustrative application based on data from the national census of population and housing with respect to marital status, through which it was made possible to identify relationships between legal marital status and actual marital status taking into account other basic socio-demographic variables available in large datasets. Study (Roodpishi et al. 2015) conducted on various demographic variables for an insurance company in the city of Anzali, Iran, provides various associations with clients of an insurance company. The study used association rules and practice of insurance policy to find hidden patterns in the insurance industry.

In the study done by Sagin et al. (2018), market basket analysis was conducted on a data of a large hardware company operating in the retail sector. Both the Apriori and FP growth algorithms (Sagin et al. 2018) were run separately and their usefulness in such a set of data was compared. When both the algorithms were compared in terms of performance, it was seen that FP growth algorithm yielded 781 times faster results but resulting rules showed that FP growth algorithm failed to find the first 14 rules with high confidence value. In the study done by Srinivasa Kumar et al. (2018), product positioning of a retail supermarket based at Trichy, Tamil Nadu, was examined using data mining to identify the items sets that were bought frequently and association rules were generated. The study done by Santarcangelo et al. (2018) focused on visual market basket analysis with the goal to infer the behavior of the customers of a store with the help of dataset of egocentric videos collected during customer's real shopping sessions. They proposed a multimodal method which exploited visual, motion and audio descriptors and concluded that the multimodal method achieved an accuracy of more than 87%. In the study done by Avcilar et al. (2014), association rules were estimated using market basket analysis and taking support, confidence and lift measures. These rules helped in understanding the purchase behavior of the customers from their visit to a store while purchasing similar and different product categories. The objective of the research study done by Seruni et al. (2005) was to identify the associated product, which then were grouped in mix merchandise with the help of market basket analysis. The association between the products was then used in the design layout of the product in the supermarket.

1.2 Pricing Intelligence

Pricing intelligence consists of tracking, monitoring and analyzing pricing data to understand the market and make educated pricing changes at speed and scale (Ballard 2018). Pricing intelligence can help in determining the effective price for various products that will give an edge over the competitors and also help in boosting up the sales. If used smartly, it can also act as a tool to clear the pending stock in an outlet. Pricing of a product can be determined by keeping various factors in mind such as time duration of a particular product in a shelf and competitor's price for the same product. Discount percentage could also be determined by the time a product is on the shelf.

2 Methodology

In the current study, an attempt is made to find the relationship between the different products using Apriori algorithm.

At the first stage, data is preprocessed and transformed, values are handled and the data is cleaned before selecting the components. After transforming the data, Apriori algorithm was used to find the relationship between different apparels using association rules. The data is analyzed on the basis of results obtained from Frontline Analytic Solver® Data Mining (XLMiner).

Various parameters used for evaluation of the model are antecedent support (if part) which is the number of transactions in which item/s is present, consequent support (then part) which is number transactions in which item/s is present, support which is number of transactions that include all items in the antecedent and consequent. Antecedent (the "if" part) and the consequent (the "then" part), an association rule contains two numbers that express the degree of uncertainty about the rule.

The first number is called the support which is simply the number of transactions that include all items in the antecedent and consequent. The second number is confidence which is the ratio of the number of transactions that include all items in the consequent as well as the antecedent (namely, the support) to the number of transactions that include all items in the antecedent.

Lift is another important parameter of interest in the association analysis. It is the ratio of confidence to expected confidence. A lift ratio larger than 1.0 implies that the relationship between the antecedent and the consequent is more significant. Larger the lift ratio, the more significant the association. The following are the parameters used to evaluate the model:

- Support—support which is simply the number of transactions that include all items in the antecedent and consequent.
- Confidence = (no. of transactions with antecedents and consequent item sets)/(no. of transactions with antecedents item sets).

- Benchmark confidence = (number of transactions in consequent item sets)/ (number of transactions in database).
- Lift ratio = confidence/benchmark confidence.

In Frontline Analytic Solver® Data Mining (XLMiner), minimum support transaction and confidence percentage controlled parameters were used for designing the model and checking the performance of the data mining.

3 Data

The data used for this paper is collected from "Try Us" a multi-brand retail outlet in Indore, Madhya Pradesh, for a period during November 26, 2017, to September 19, 2018. The collected data is used for the study of association between different products, and inferences generated can then be used to arrange shelves in a better way when planned for a bigger retail store. The data collected includes bill number, date, brand name, size, amount, GST, item type which was refined according to our purpose to bill number, brand name and item type.

For applying Apriori algorithm on binary data format, the data was first converted to binary format where if a product was purchased it was recorded as 1 and if no purchase was made then it was recorded as 0. In total, there are 29 columns and 13,065 rows which were refined to 29 columns and 6008 rows since multiple items purchased by a customer were recorded in multiple rows. The brands that were not present in the store from November 26, 2017, were not taken into consideration. Therefore, 185 rows were deleted out of 6193 rows. Multiple purchases made by a single customer were merged in a single row.

3.1 Data Analysis, Results and Findings

Main objective of the study is to study the buying pattern of the customer and to analyze proper basket (combos) of products for cross-selling and upselling. Another objective is to study the association between the products for applying the optimal design layout for the retail store. In the paper, association rule mining through Apriori algorithm is used to find baskets of products which are purchased together. A total of 5223 transactions are included in the analysis. Using combination of various minimum support transactions and minimum confidence percentage, the following results are derived:

Case I

Association rules: fitting parameters	
Method	Apriori
Min support	50
Min confidence	20

Rules

Rule ID	Antecedent	Consequent	A-support	C-support	Support	Confidence	Lift ratio
Rule 1	[TSHIRTSA]	[SHIRTSB]	660	1994	143	21.67	0.57
Rule 2	[TSHIRTG]	[SHIRTSB]	547	1994	116	21.21	0.56
Rule 3	[TSHIRTN]	[SHIRTSN]	173	1121	59	34.10	1.59
Rule 4	[SHIRTSN]	[SHIRTSB]	1121	1994	259	23.10	0.61
Rule 5	[SHIRTSF]	[SHIRTSB]	307	1994	67	21.82	0.57
Rule 6	[SHIRTSJ]	[SHIRTSB]	150	1994	51	34.00	0.89
Rule 7	[SHIRTSA]	[SHIRTSB]	257	1994	70	27.24	0.71
Rule 8	[JEANSC]	[SHIRTSB]	623	1994	240	38.52	1.01
Rule 9	[JEANSN]	[SHIRTSB]	401	1994	87	21.70	0.57
Rule 10	[JEANSO]	[SHIRTSB]	234	1994	101	43.16	1.13
Rule 11	[JEANSM]	[SHIRTSB]	208	1994	70	33.65	0.88
Rule 12	[TROUSERD]	[SHIRTSB]	438	1994	168	38.36	1.00
Rule 13	[JEANSN]	[SHIRTSN]	401	1121	187	46.63	2.17
Rule 14	[TROUSERD]	[SHIRTSN]	438	1121	96	21.92	1.02
Rule 15	[JEANSF]	[SHIRTSF]	196	307	60	30.61	5.21

Lift ratio—Lift value of an association rule is the ratio of the confidence of the rule and the expected confidence.

Confidence percentage—The confidence of an association rule is a percentage of number of transactions with antecedents and consequent item sets divided by number of transactions with antecedents item sets.

In **case I,** the rules having lift ratio more than 1 are rule 3, rule 8, rule 10, rule 12, rule 13, rule 14, rule 15. A brief description of these rules is given below.

Rule 3

A customer who purchases T-shirt N (Mufti) purchases a shirt N (Mufti).

Rule 8

A customer who purchases jeans C (Nostrum) purchases a shirt B (Ecohawk).

Rule 10

A customer who purchases jeans O (Revit) purchases a shirt B (Ecohawk).

Rule 12

A customer who purchases trouser D (Sixth Element) purchases a shirt B (Ecohawk).

Rule 13

A customer who purchases jeans N (Mufti) purchases a shirt N (Mufti).

Rule 14

A customer who purchases trouser D (Sixth Element) purchases a shirt N (Mufti).

Rule 15

A customer who purchases jeans F (US Polo) purchases a shirt F (US Polo).

The products with the lift ratio between 1 and 2 should be clubbed and kept together on the same shelf. For example, T-shirts N (Mufti) and shirt N (Mufti), jeans C (Nostrum) and shirt B (Ecohawk), jeans O (Revit) and shirt B (Ecohawk), trouser D (Sixth Element) and shirt B (Ecohawk), trouser D (Sixth Element) and shirt N (Mufti) should be clubbed and kept together.

Similarly, the products with the lift ratio between 2 and 6 should be clubbed and kept together on the same shelf.

Case II

Association rules: fitting parameters	
Method	Apriori
Min support	100
Min confidence	20

Rules:

Rule ID	Antecedent	Consequent	A-support	C-support	Support	Confidence	Lift ratio
Rule 1	[TSHIRTSA]	[SHIRTSB]	660	1994	143	21.67	0.57
Rule 2	[TSHIRTG]	[SHIRTSB]	547	1994	116	21.21	0.56
Rule 3	[SHIRTSN]	[SHIRTSB]	1121	1994	259	23.10	0.61
Rule 4	[JEANSC]	[SHIRTSB]	623	1994	240	38.52	1.01
Rule 5	[JEANSO]	[SHIRTSB]	234	1994	101	43.16	1.13
Rule 6	[TROUSERD]	[SHIRTSB]	438	1994	168	38.36	1.00
Rule 7	[JEANSN]	[SHIRTSN]	401	1121	187	46.63	2.17

In **case II,** the rules having lift ratio more than 1 are rule 4, rule 5, rule 6, rule 7. A brief description of these rules is given below.

Rule 4

A customer who purchases jeans C (Nostrum) purchases a shirt B (Ecohawk).

Rule 5

A customer who purchases jeans O (Revit) purchases a shirt B (Ecohawk).

Rule 6

A customer who purchases trouser D (Sixth Element) purchases a shirt B (Ecohawk).

Rule 7

A customer who purchases jeans N (Mufti) purchases a shirt N (Mufti).

Thus, jeans C (Nostrum) and shirt B (Ecohawk), jeans O (Revit) and shirt B (Ecohawk), trouser D (Sixth Element) and shirt B (Ecohawk), jeans N (Mufti) and shirt N (Mufti) should be clubbed and kept together.

Case III

Association rules: fitting parameters	
Method	Apriori
Min support	150
Min confidence	20

Rules:

Rule ID	Antecedent	Consequent	A-support	C-support	Support	Confidence	Lift ratio
Rule 1	[SHIRTSN]	[SHIRTSB]	1121	1994	259	23.10	0.61
Rule 2	[JEANSC]	[SHIRTSB]	623	1994	240	38.52	1.01
Rule 3	[TROUSERD]	[SHIRTSB]	438	1994	168	38.36	1.00
Rule 4	[JEANSN]	[SHIRTSN]	401	1121	187	46.63	2.17

In **case III,** the rules having lift ratio more than 1 are rule 2, rule 3, rule 4. A brief description of these rules is given below.

Rule 2

A customer who purchases jeans C (Nostrum) purchases a shirt B (Ecohawk).

Rule 3

A customer who purchases trouser D (Sixth Element) purchases a shirt B (Ecohawk).

Rule 4

A customer who purchases jeans N (Mufti) purchases a shirt N (Mufti).

Thus, jeans C (Nostrum) and shirt B (Ecohawk), trouser D (Sixth Element) and shirt B (Ecohawk), jeans N (Mufti) and shirt N (Mufti) should be clubbed and kept together.

From the data, it was observed that shirt B (Ecohawk) and shirt N (Mufti) had a very strong association as they both were sold together for 259 times. Similarly, shirt B (Ecohawk) and jeans C (Nostrum) were sold together for 240 times.

Figure 1 gives a radar chart of what products are purchased together:

Different colors represent different types of products that were available in the store. Each concentric circle represents 50 transactions. The following suggestions can be given to the entrepreneur after analyzing the data.

- On the basis of market basket analysis, products which are sold together with high frequency like shirt B (Ecohawk) and jeans N (Mufti) should be kept near to each other such that it reduces the handling time of the customer by the salesperson. Furthermore, baskets can be developed using the analysis done above.
- Products which have a low frequency should be near to the products that are preferred more by the customer with some dynamic discount pattern so as to increase the sales of the low frequency products.

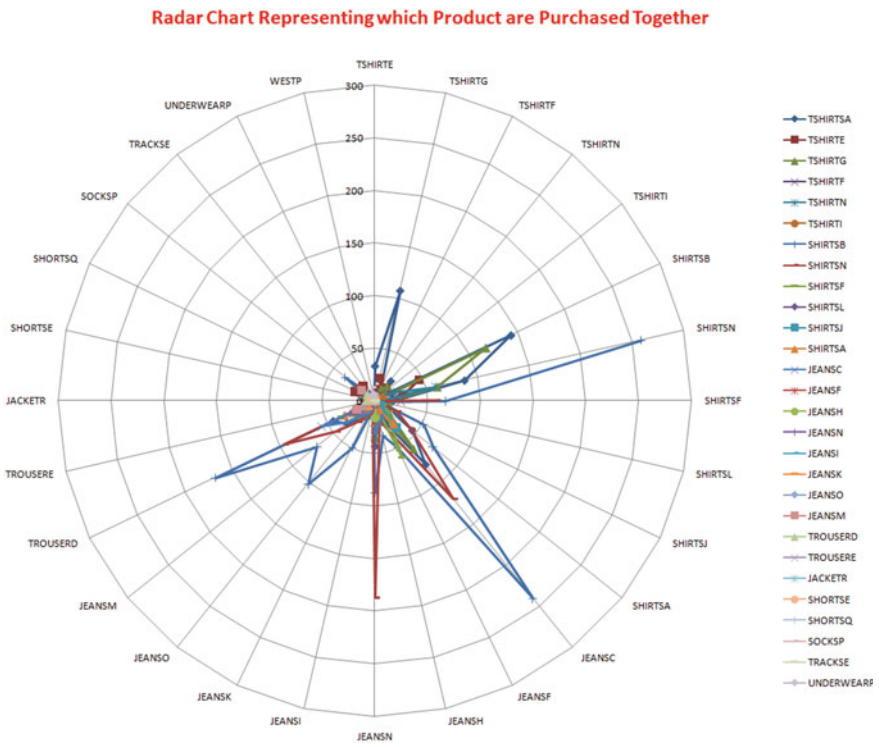

Fig. 1 Radar chart

4 Conclusion

Applying market basket analysis, some meaningful patterns were obtained which would help the entrepreneur in planning the discount pattern of the products and also making various combos according to the rules generated in order to increase the sales. It will also help in preparing the shelf design for the new store so as to minimize the product searching time by the customer.

Appendix

Details containing brand name and code used in the research.

Serial No.	Code	Brand
1	A	Cookyss
2	B	Ecohawk

(continued)

(continued)

Serial No.	Code	Brand
3	C	Nostrum
4	D	Sixth Element
5	E	Status Que
6	F	US Polo
7	G	Stride
8	H	Killer
9	I	Ed Hardy
10	J	Yankee
11	K	Vogue Raw
12	L	Delmont
13	M	Rookies
14	N	Mufti
15	O	Revit
16	P	UCB
17	Q	Beevee
18	R	Silver Surfer
19	T	Status Quo
20	U	M Square
21	W	Got It
22	X	Borgoforte
23	Y	Fort Collins
24	Z	Okane

References

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Association of Computing Machinery (ACM) SIGMOD Record*, *22*(2). Newyork, USA. https://doi.org/10.1145/170036. 170072. ISSN-0163-5808.

Avcilar, M. S., & Yakut, E. (2014). Association rules in data mining: An application on a clothing and accessory specialty store. *Canadian Social Science, 10*(3), 75–83.

Ballard, A. (2018, August 06). Pricing intelligence: What it is and why it matters. Retreived from https://www.mytotalreatil.com/article/pricing-intelligence-what-it-is-and-why-it-matters/. Date of downloading January 31, 2019.

Cochran, C., Ohlmann, F., Williams, A. S. (2015). *Essentials of business analytics*, pp. 323–324. ISBN-13: 978-81-315-2765-8.

Kurniawan, F., Umayah, B., Hammad, J., Mardi, S., Nugroho, S., & Hariadi, M. (2018). Market basket analysis to identify customer behaviors by way of transaction data. *Knowledge Engineering and Data Science KEDS, 1*(1), 20–25.

Roodpishi, M. V., & Nashtaei, R. (2015). Market basket analysis in insurance industry. *Management Science Letters, 5*, 393–400.

Sagin, A. N., & Ayvaz, B. (2018). Determination of association rule with market basket analysis: An application of the retail store. *Southeast Europe Journal of Soft Computing, 7*(1), 10–19.

Santarcangelo, V., Farinella, G. M., Furnari, A., & Battiato, S. (2018). Market basket analysis from egocentric videos. *Pattern Recognition Letters, 112*, 83–90.

Srinivasa Kumar, V., Renganathan, R., VijayBanu, C., & Ramya, I. (2018). Consumer buying pattern analysis using apriori association rule. *International Journal of Pure and Applied Mathematics, 119*(7).

Surjandari, I., & Seruni, A. C. (2005). Design of product layout in retail shop using market basket analysis. *MakaraTeknolgi, 9*(2), 43–47.

Szymkowiak, M., Klimanek, T., & Jozefowski, T. (2018). Applying market basket analysis to official statistical data. *Econometrics Ekonometria Advances in Applied Data Science, 22*(1), 39–57.

Tan, P.-N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*, pp. 2–3. ISBN 978-93-3257-140-2.

Tatiana, K., & Mikhail, M. (2018). Market basket analysis of heterogeneous data sources for recommendation system improvement. *Procedia Computer Science*, 246–254. ISSN 1877-0509.