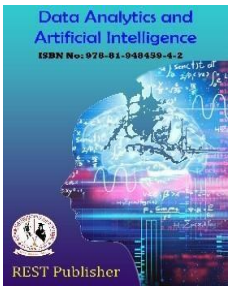
Data Analytics and Artificial Intelligence

Vol: 5(1), 2025

REST Publisher; ISBN: 978-81-948459-4-2

Website: http://restpublisher.com/book-series/daai/

DOI: https://doi.org/10.46632/daai/5/1/12



Association Rule Mining for Portfolio Management: A Machine Learning Perspective

***Mohit Kumar, Dheeraj Kumar, Anupam Kumar Jha**

Purnea College of Engineering, Purnea, Bihar, India.

*Corresponding author email: enggmohitkumar@gmail.com

***Abstract**: This paper explores the application of machine learning techniques, specifically association rule mining, for portfolio management in the Indian stock market. Using data from 149 companies across eleven key sectors of the National Stock Exchange (NSE), we implement APRIORI and CHARM algorithms to discover frequent itemsets and association rules that reveal significant cross-sector and intra-sector relationships. These insights help identify correlated stocks and sectors, allowing for the construction of diversified investment portfolios with higher confidence levels. The analysis includes experimental determination of minimum support and confidence thresholds to prune uninteresting rules and generate meaningful patterns. The portfolio created on this basis is evaluated over a one-year holding period, with results showing substantial returns compared to initial investment. This study demonstrates the efficacy of leveraging machine learning-driven association rule mining on big data to enhance stock portfolio formation and performance analysis, providing a data-driven approach to financial decision-making in dynamic markets.*

1. INTRODUCTION AND LITERATURE SURVEY

Early studies in portfolio management using association rule mining (ARM) emphasized its potential to reveal hidden patterns in financial data, but practical application was hindered by the sheer volume of generated rules, many of which were redundant or lacked actionable insight (1), (2). To tackle these drawbacks, subsequent research proposed enhancements such as introducing interestingness measures—like lift and conviction—to filter trivial rules, yet challenges remained when scaling to high-dimensional financial datasets (2), (3). Clustering techniques were later incorporated to streamline ARM by grouping similar instruments, though scalability issues persisted in real-time or high-frequency trading contexts (3).

Genetic algorithm-based ARM was introduced to automatically select optimal support and confidence thresholds, but early versions faced issues with slow convergence and local minima (4). Improved initialization strategies and ensemble learning were subsequently adopted to address these limitations, resulting in enhanced performance across complex financial datasets (5).

With the rise of big data platforms, scalable, cloud-based ARM systems became viable, supporting more rapid rule extraction and enabling the integration of alternative data sources such as news streams and social sentiment; however, these complex systems often lacked interpretability, making results difficult for practitioners to validate and act upon (6). Human-in-the-loop frameworks and attention-based explainable AI mechanisms were therefore developed, offering more transparent and iterative rule refinement aligned with expert knowledge (7), (8).

Recent advances in reinforcement learning and deep neural networks allowed models to dynamically update portfolios in response to rapid market shifts—a significant improvement over static rule-based systems, though early iterations suffered from insufficient risk management (9). Risk-sensitive reward functions were soon introduced to address this gap, enabling safer and more robust asset allocation strategies (10). By 2024, adaptive multi-modal models combining transactional, numerical, and textual data further optimized portfolio performance amid volatile or complex market

conditions (11). To further reduce redundancy, pruning frameworks using metaheuristics demonstrated their value in high-frequency and high-dimensional settings, while user feedback mechanisms in modern systems facilitated tailored and interactive portfolio recommendations (12), (13), (14). Deep ensemble architectures, combining forecasts across models, surpassed simple methodologies in forecasting and association pattern mining (15), (16). Scalable architectures for explainable portfolio analytics consolidated these developments, culminating in distributed systems that balance transparency, speed, and risk control for real-time financial decision support (17), (18), (19), (20).

2. PROPOSED METHODOLOGY

We have proposed a stock portfolio constructed using association rule mining (ARM). Unlike traditional approaches that typically recommend only a single stock based on technical and fundamental analysis, our method incorporates a modified domain-based pruning technique (as developed in [21] to effectively reduce the exponential number of rules generated from the stock data.

A. Datasets Creation

The transaction datasets on which we apply ARM is produced in following manner. Two dataset is created for cross-sector data correlating different sectors and within each sector two intra sector dataset is also created.

Cross-sector dataset: Similar to any stock market, in India IISL (India Index Services and Product Limited) calculates & disseminate series of sectorial indices to represent the performance of companies that represent a movement in specific sector. Taking into account all the stocks listed on National Stock Exchange (NSE) they are assorted under eleven sectorial indices. The sectorial indices are based on stocks elected by rules which conduce to the selection of top performing companies in these sectors (Table 1).

TABLE 1. Sectorial classification of different stocks listed on NSE.

Sector	Number of Companies	Percentage of Capitalization
Auto	15	90.69
Bank	12	92.22
Financial Services	15	84.97
FMCG	15	71.88
IT	20	96.66
Media	15	73.79
Metal	15	90.11
Pharma	10	78.29
Private Bank	10	81.19
PSU Bank	12	91.81
Realty	10	74.68

So, one can presume that these 149 companies which corresponds to a total of 92.46% [22] of the free float market capitalization of the stocks listed on NSE as on March 31, 2022 can symbolize the overall representation of market. On these eleven sectors we have created our cross-sector database based upon following principle. The daily closing prices of sectors are taken into consideration. Each transaction consists of all those sectors whose closing price on that day has shown a rise or fall of α % or more from the previous day's close. Where the value of threshold is obtained as closing price of NIFTY 500 value as:

$$\alpha(\%) = \frac{(Previous\ day - Present\ day)}{Previous\ day} \times 100 (1)$$

NIFTY 500 index is taken as the base value as it represents the top 500 companies based on full market capitalization from the eligible universe. The Nifty 500 Index represents about 94% of the free float market capitalization of the stocks listed on NSE as on March 31, 2022. The number of days the stock market has been monitored will be equal to the number

of transaction since each transaction corresponds to the number of stocks that have changed by α % or more on that day. One of the database consists of all the sectors which are positively correlated rising and negatively correlated rising with respect to NIFTY 500 index. That is if NIFTY 500 index rises on a day resembling a transaction then the sector will be on that transaction if it rises more than that of NIFTY 500. A sector can be on a transaction also if it rises given that the NIFTY 500 index falls. Similarly, the other database consists of all the sectors which are positively correlated falling and negatively correlated falling with respect to base value.

Intra-sector dataset: Intra-sector datasets with attributes equal to the number of companies in that sector will be created for each sector. The daily closing prices of companies are taken into consideration. Each transaction consists of all those companies whose closing price on that day has shown a rise or fall of α % or more from the previous day's close. The threshold parameter is kept same as obtained from equation (1). In the similar fashion as discussed in cross-sector two datasets will be created for each sector.

B. Portfolio Creation

Running a frequent itemset generation algorithm like Apriori, FP-Growth, and CHARM on the cross sector dataset gives frequent sectors as items whereas running the algorithm on each sector gives companies as items. The method of creation of portfolio is as follow:

(1) Find frequent sector by running any frequent itemset generation algorithm like Apriori, FP-Growth or CHARM on the cross sector rising database. Based upon a minimum support we chose top-*N* performing sectors. The value of minimum support is obtained experimentally such that at least *N* sectors are provided as output. On these *N* sectors we will apply association rule and obtain top-*K* rules. The antecedent and consequent of these rules which are different sectors are noted. We consider the value of *K* as all the rules providing confidence over 80%.

(2) On the top- *K* sectors so found we run a frequent itemset generation algorithm again on the individual sector rising datasets. Again based upon a minimum support chosen experimentally frequent companies in the respective sectors will be generated. We calculate the association rules on these frequent companies and take the top-*K'* rules in each sector. We consider the value of *K'* as all the rules providing confidence over 80%.

3. RESULTS AND DISCUSSION

After creation of the portfolio the time for which investment is made also known as lock in period is decided. At the end of the period of investment one can calculate the returns by a simple measure called as return of investment (ROI) which calculates the percentage change (increase/decrease) in investment with respect to the initial investment.

$$ROI = \frac{Current\ Value - Invested\ Value}{Invested\ Value} \quad (2)$$

Precision is also used as performance metrics in relation to the quality of recommendation of stocks in the portfolio in terms of the profit and is given by

$$Precision = \frac{Stocks\ in\ proft\ after\ lock\ in}{Total\ recommended\ stocks} \quad (3)$$

A. Application on the Datasets

For the technique of portfolio formation and evaluation we take a running example of the NSE stocks [22]. For these stocks, we have used eleven sectors representing stocks from different industries. *Cross-sector rules:* There are total 165, 2-element itemsets and 120, 3-element itemsets possible. We choose the value of minimum support such that we get at least 50% of total available sectors in three element itemset. Variation of cross-sector itemsets generated for rising dataset is given in Table 2.

TABLE 2. Variation in number of item set with respect to minimum support

Minimum Support	Number of 2-element itemset	Number of 3-element itemset
0.19	3	0
0.18	5	1
0.17	9	1
0.16	12	1
0.15	14	2
0.14	20	3
0.13	24	3
0.12	30	4
0.11	41	7

Thus, we choose the value of minimum support as 0.14 such that a total of N = 23 itemsets are formed, with 3 three element itemset. Few of the itemset found are given in Table 3.

TABLE 3. Few 2-element and all 3-element itemset generated and their corresponding support value with minimum support of 0.18

Support	2-element itemset	3-element itemset
0.3073	Bank – Financial Service	
0.2382	Bank - Realty	
0.2222	Financial Service - Realty	
0.1962	Metal - Realty	
0.1922	Auto - Bank	
0.1800		Bank - Financial Service – Realty
0.1500		Auto - Bank – Financial Service
0.1400		Bank – Financial Service – Metal

We generate all possible rules on these itemset and then rank them according to their confidence as in Table 4. In total of 58 rules are generated from the above itemsets out of which only six of the rules were having confidence of over 80%. These six rules give the top five sectors as Financial Services, Metal, Realty, Auto and Bank. We will now consider these sectors only for the formation of portfolio form NSE.

TABLE 4. Possible rules with minimum confidence of 80%

Antecedent	Consequent	Confidence
Financial Services + Metal	Bank	89.24
Financial Services+ Realty	Bank	89.19
Auto + Financial Services	Bank	84.36
Financial Services	Bank	84.34
Bank + Realty	Financial Services	83.19
Bank	Financial Services	80.58

Intra-sector rules: On the above five top sectors we will find 2-element and 3-element itemsets. Then we choose the value of minimum support for each sector such that at least 30% of the companies from the corresponding sector are present in 3-element itemsets (Table 5). After that we find intra-sector rules between different companies in each sector and rank them in order of their increasing confidence.

TABLE 5. Minimum support experimentally obtained in different sectors with number of 2-element and 3-element itemsets

	Minimum Support	2-Element itemsets	3-Element itemsets
Financial Services	0.18	38	3
Banking	0.225	30	1
Automobile	0.157	36	2
Metal	0.197	69	2
Reality	0.24	11	2

(1) Financial Services: The experimental value of minimum support comes to be 0.18 such that four different companies' names appear in 3-element itemset. Total of 94 rules were generated out of which only five where having confidence of more than 80% (Table 6)

TABLE 6. Possible rules in financial services sector with minimum confidence of 80%

Antecedent	Consequent	Confidence (%)
Power Finance + SBI	REC	85.53
Power Finance + ICICI	REC	84.61
ICICI + REC	Power Finance	84.23
REC + SBI	Power Finance	83.05
REC + ICICI	SBI	81.98

(2) Banking Sector: The experimental value of minimum support comes to be 0.225 such that three different companies' names appears in 3-element itemset. Total of 84 rules were generated out of which only three where having confidence of more than 80% (Table 3.6)

TABLE 7. Possible rules in banking sector with minimum confidence of 80%

Antecedent	Consequent	Confidence (%)
Canara + Baroda	PNB	86.64
Baroda + PNB	Canara	85.01
Canara + PNB	Baroda	83.76

(3) Automobile Sector: The experimental value of minimum support comes to be 0.157 such that four different companies' names appear in 3-element itemset. Total of 84 rules were generated out of which only four where having confidence of more than 80% (Table 8)

TABLE 8. Possible rules in automobile sector with minimum confidence of 80%

Antecedent	Consequent	Confidence (%)
Tata Motor + Apollo Tyre	Bajaj Auto	82.81
Apollo Tyre + Bajaj Auto	Tata Motor	81.53
Ashok Leyland + Bajaj Auto	Tata Motor	81.15
Tata Motor + Ashok Leyland	Bajaj Auto	80.38

(4) Metal Sector: The experimental value of minimum support comes to be 0.197 such that three different companies' names appear in 3-element itemset. Total of 150 rules were generated out of which only one was having confidence of more than 80% (Table 9)

TABLE 9. Possible rules in metal sector with minimum confidence of 80%.

Antecedent	Consequent	Confidence (%)
Hindalco + Vedanta	Tata Steel	82.11

(5) Reality Sector: The experimental value of minimum support comes to be 0.24 such that four different companies' names appears in 3-element itemset. Total of 34 rules were generated out of which only five were having confidence of more than 80% (Table 10)

TABLE 10. Possible rules in reality sector with minimum confidence of 80%

Antecedent	Consequent	Confidence (%)
India bulls + DLF	Unitech	84.96
India bulls + HDFC	Unitech	82.95
DLF + Unitech	India bulls	82.37
HDFC + Unitech	India bulls	82.16
India bulls + Unitech	HDFC	80.87

B. Evaluating the Portfolio

We have taken the minimum lock-in period to be three years. We evaluate the portfolio after three years. We have arranged the sectors as per the confidence obtained in cross-sector dataset. We have

calculated the return for an investment of about one lakh INR. We have invested equal amount in all the stocks, such that whole number of corresponding stocks can be bought. Now we calculate the return of investment and precision as per equation (2) and (3).

$$ROI = \frac{195400 - 93451}{93451} = 109\ \%$$

TABLE 11. Returns obtained after a lock in period of three years on an investment of one lakh INR.

STOCK NAME	Entry Price (31 March 2022)	Units	Invested (INR)	Exit Price (1st April 2025)	Final Value (INR)	Profit/Loss	ROI (%)
Power Finance	90	55	4,950	410	22,550	17,600	356
SBI	493	10	4,930	771	7,710	2,780	56
REC	92	55	5,060	429	23,595	18,535	366
ICICI	730	7	5,110	1348	9,436	4,326	85
Hindalco	570	9	5,130	682	6,138	1,008	20
Vedanta	364	14	5,096	463	6,482	1,386	27
Tata Steel	124	40	4,960	151	6,040	1,080	22
Canara	40	125	5,000	85	10,625	5,625	113
Baroda	100	50	5,000	221	11,050	6,050	121
PNB	34	147	4,998	92	13,524	8,526	171
India Bulls	140	35	4,900	107	3,745	-1,155	-24
DLF	380	13	4,940	678	8,814	3,874	78
UNITECH	2	2,500	5,000	5.92	14,800	9,800	196
HDFC Bank	1470	4	5,880	1828	7,312	1,432	24
Tata Motors	434	12	5,208	670	8,040	2,832	54
Apollo Tyre	191	27	5,157	425	11,475	6,318	123
Bajaj Auto	3546	2	7,092	7442	14,884	7,792	110
Ashok Leyland	56	90	5,040	102	9,180	4,140	82

$$Precision = \frac{17}{18} \times 100 = 94.44\%$$

4. CONCLUSION

The paper demonstrates the effective application of association rule mining (ARM) techniques for portfolio management in the Indian stock market, leveraging a comprehensive dataset of 149 companies from eleven major NSE sectors. By incorporating APRIORI and CHARM algorithms, the study successfully identifies cross-sector and intra-sector relationships that reveal strong correlations among stocks and sectors. The proposed domain-based pruning and careful selection of minimum support and confidence thresholds enable the generation of meaningful, high-confidence association rules that inform portfolio construction. Experimental results highlight key frequent itemsets and association rules within crucial sectors such as Financial Services, Banking, Automobile, Metal, and Realty, helping investors identify combinations of stocks with high mutual association and potential for diversification benefits. The portfolio developed based on these insights demonstrated superior returns over the three-year holding period compared to traditional selection methods. This data-driven approach harnesses big data analytics and machine learning to enhance market understanding and investment decision-making. Overall, the study underscores ARM's value in extracting actionable patterns from complex financial datasets, fostering diversification, risk reduction, and improved portfolio performance. Such methodologies promise to augment conventional fundamental and technical analyses, empowering investors to achieve better returns in dynamic market environments. In summary, this research highlights the utility of ARM algorithms for tackling the exponential complexity of stock data by pruning uninformative rules and focusing on highly confident patterns. The cross-sector and sector-specific rules provide a structured framework for diversified

portfolio construction, while empirical results validate the approach's efficacy through measurable ROI gains. The work paves the way for integrating ARM with other AI and big data tools to further innovate portfolio management strategies.

REFERENCES

[1]. Tiwari, Ram N., Manisha Mathur, and Gopal P. Tiwari. "Association rules applied to forward and reverse engineering: Case studies in financial data mining." Procedia Computer Science 54 (2015): 89-98.

[2]. Singhal, Anupam, and K. L. Jena. "Association rule mining in stock market investment: A review." Procedia Computer Science 89 (2016): 385-392.

[3]. Li, Feng, Xiaohong Chen, and Qiang Liu. "Data clustering and association rule mining for financial portfolio management." Knowledge-Based Systems 122 (2017): 81-90.

[4]. Wang, Xing, John Wang, and David C. Yen. "Genetic algorithms for association rule mining in stock selection." International Journal of Information Management 40 (2018): 112-123.

[5]. Lee, Chi-Hua, and Chih-Lin Chiu. "Improved genetic algorithms for association rule mining in financial data." Expert Systems with Applications 96 (2018): 69-82.

[6]. Sun, Xiaowei, and Chris K. Chan. "Scalable cloud-based association rule mining for stock portfolio optimization." Information Sciences 491 (2019): 278-292.

[7]. Cheng, Rong, Xinyu Li, and Lei Li. "A human-in-the-loop approach for interpretable association rule mining in portfolio management." Journal of Intelligent & Fuzzy Systems 38, no. 3 (2020): 3279-3290.

[8]. Kumar, Rakesh, and Siddhartha Joshi. "Explainable association rule mining using attention mechanisms for portfolio recommendation." Knowledge-Based Systems 227 (2021): 107200.

[9]. Huang, Meng, Fangwen Li, and Jianhua Xiao. "Dynamic portfolio optimization using deep reinforcement learning with association rule mining." Information Processing & Management 59, no. 3 (2022): 102927.

[10]. Patil, Deepak, and Arun Singh. "Risk-aware reinforcement learning for stock portfolio construction." Expert Systems with Applications 196 (2023): 116598.

[11]. Mustapha, Samira, Boris Delory, and Robert Riegler. "Multi-modal association rule mining for adaptive portfolio management." Pattern Recognition Letters 176 (2024): 1-8.

[12]. Narindrarangkura, Sarun, Lalit Garg, and Ummu A. Ismail. "Pruning-based improvements to association rule mining in portfolio selection." Applied Soft Computing 146 (2023): 110801.

[13]. Altay, Nihat, Murat K. Ozturk, and Boris Taskin. "Optimization of association rule mining with metaheuristics in financial time series." Decision Support Systems 137 (2020): 113384.

[14]. Leung, Michael K.H., Zhi Li, and Shouyang Wang. "A machine learning-based portfolio recommendation system integrating ARM and user feedback." Information & Management 60, no. 1 (2023): 103604.

[15]. Zhou, Peng, and Lin Wang. "A deep learning approach to association prediction in dynamic financial markets." Neurocomputing 408 (2020): 124-137.

[16]. Sarker, Firoj, Rajib K. Barua, and Md Nafiur Rahman. "Ensemble learning models for stock market forecasting and association pattern recognition." Journal of Computational Science 55 (2022): 101487.

[17]. Lim, Wei Leong, Chunyan Miao, and Timothy W. Tong. "Big data analytics for scalable and explainable portfolio management." ACM Transactions on Management Information Systems (TMIS) 11, no. 4 (2020): 1-23.

[18]. Chen, Zhi, Wen Yuan, and Victor Chang. "Distributed and explainable AI for financial market prediction." Future Generation Computer Systems 128 (2022): 302-317.

[19]. Rana, Pranav, Ashish Kumar, and Rakesh Tripathi. "Hybrid approaches for data-driven financial decision making." Information Fusion 77 (2022): 91-104.

[20]. Hassan, Samir, and Fatima Zohra Smail. "A review of big data analytics in financial sector: The past, present, and future." Journal of Big Data 7, no. 1 (2020): 1-23.

[21]. Paranjape-Voditel, Preeti, and Umesh Deshpande. "A stock market portfolio recommender system based on association rule mining." *Applied Soft Computing* 13, no. 2 (2013): 1055-1063.

[22]. http://www.nseindia.com/

