DOI: 10.54254/2977-3903/2025.20556

Time series data analysis and association rule mining in financial recommendation systems using Hadoop and Spark

*Yaoyu Chen [1], Yichen Xu [2, *]*

[1] The University of Manchester, Manchester, United Kingdom
[2] Australian National University, Canberra, Australia

[*] rara481846778@gmail.com

Abstract. Increasing amounts of financial data demand sophisticated analytics to develop sound recommendation models. This article discusses combining time series analysis and association rule mining for big data in Hadoop and Spark to enrich financial product recommendation engines. The paper is an integrated analysis of two types of prediction algorithms: AutoRegressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) networks to forecast user behavior and demand for financial services in the future from transactional history. The ARIMA model is used as the default while the LSTM model is used to represent non-linear dependencies and give a more dynamic forecast. association rule mining – in particular the Apriori algorithm – is used to find latent patterns and relationships between user transactions and financial products. This article illustrates how time series forecasting and association rule mining can be merged to bring a more useful financial recommendation. The hybrid approach, which combines both approaches, proves to increase user interaction and recommendation accuracy by 20% compared to the previous systems, according to experiments. The paper emphasises the possibilities of using big data in the construction of scalable, individualized financial recommendation systems.

Keywords: Time Series Analysis, Financial Recommendation Systems, Hadoop, Spark, Association Rule Mining

1. Introduction

Over the past couple of years, the amount and complexity of data in the financial services sector have exploded. As there are more and more financial products and services, personal and timely recommendations for users are no longer possible. Predictive systems like collaborative filtering or content-based recommendation models don't always take into account the fluidity and temporal nature of financial transactions. The cold-start issue, the lack of data and the fact that time-dependent behaviour cannot be captured are what makes these systems inapplicable in the real world. Therefore, we're also seeing interest in the application of advanced analytics methods like time series prediction and association rule mining for the optimization of financial recommendation systems. Time series analysis is an important part of financial recommendation systems because it allows forecasting of users' future activity and financial products demand using historical information. Banking transactions show temporal patterns of changing user needs and desires and these patterns need to be factored in when suggesting products. Autoregressive models such as ARIMA are the most commonly employed models for time series prediction but they are impregnable due to non-linear dependence. Deep learning (and more recently, Long Short-Term Memory (LSTM) networks) have shown promise in addressing these drawbacks by capturing long-term dependencies and non-linear dynamical features in time series data. Besides forecasting time series, association rule mining is also another useful technique to uncover the secret connections between user behaviours and financial instruments. Association rule mining (Apriori algorithm), is a popular association rule mining algorithm which has been successfully used to find the common itemsets and produce association rules showing the relationships between products and users. In financial services, it can tell you which products your user might be likely to be interested in given previous behaviours and interactions. In this article, we discuss a hybrid recommendation engine, which fuses time series forecasting (ARIMA, LSTM) with association rule mining (Apriori algorithm) [1]. It is developed to manage huge volumes of financial information as big data and works with Hadoop and Spark models for data processing. The objective of this study is to illustrate how these advanced methods can make the recommendations of financial products more accurate and relevant, and thereby more engaging and happy for users. This research contributes to the increasing knowledge on big data analytics in financial services application.

2. Literature review

Financial advice systems have also attracted a lot of interest in academic and industrial studies. There are research projects that investigate different solutions like collaborative filtering, content filtering, and hybrid filtering. This kind of collaboration has already been used in user item recommendation applications but it is very prone to cold-start and sparse data in dynamic environments such as finance. Content-based filtering, on the other hand, recommends items based on the characteristics of the items and the user's preferences but does not take account of the time dimension of a financial transaction. In order to counter these drawbacks, recent studies introduced time series analysis to recommendation models. We have used time series forecast models such as ARIMA, GARCH and Long Short-Term Memory (LSTM) networks to model financial behaviour using historical information [2]. These approaches take into account temporal dependencies in financial exchanges, giving predictions that are more accurate and user-friendly in financial services. Association rule mining (in particular, the Apriori algorithm) was also employed in financial systems to discover untapped patterns and connections between financial products and users. Identifying frequently traded itemsets enables association rule mining to recommend relevant products to users based on past transaction and creates a recommendation for that user. This merged approach with a big data environment based on Hadoop and Spark for processing is an attractive avenue for building strong financial recommendation engines [3].

3. Experimental methodology

3.1. Data collection and preprocessing

The experimental approach starts with the data-gathering and preprocessing. Data for transactions came from one of the major banks: user history, purchase, demographic and behavioral data. The data is five years long with 500,000 users and 1 million transactions. Before processing, there were a number of processes that were carried out in order to check that the data were acceptable and fit for analysis. For one, missing values were imputational (by the median of data). This was critical for making sure the time series data (which measured the investment activity of users over time) was comprehensive and uninterrupted. In addition, outliers were identified using the Interquartile Range (IQR) and removed so that no model bias could be applied. Transaction data were also normalized for uniformity in all features. Transaction sizes were, for instance, scaled from 0 to 1. From there, time series conversion was used to deconstruct meaningful temporal trends of user transaction. Preprocessed data was sliced up into time series by month of transaction for time series forecasting algorithms [4]. The table 1 below represents the most important features of the raw data after preprocessing and the spread of user transactions and their behavior by product category. As you can see from the table, after preprocessing, the dataset was nearly 100% correct, with very few missing values (mainly product type-specific) and outliers were effectively eliminated [5].

Table 1. Summary Statistics of Preprocessed Financial Data

Feature	Description	Total Records	Missing Values (%)	Outliers Removed (%)
User ID	Unique identifier for each user	500,000	0.02%	1.5%
Transaction Amount	Monetary value of each transaction	1,000,000	0.05%	0.8%
Product Category	Category of financial products	1,000,000	0.00%	1.2%
Transaction Date	Date of the transaction	1,000,000	0.00%	0.0%

3.2. Time series analysis and forecasting models.

It was at the heart of the time series analysis to anticipate user's and financial product demand in the future. We used two prediction models for this, ARIMA (AutoRegressive Integrated Moving Average) and LSTM (Long Short-Term Memory). We used the ARIMA model as a benchmark because it was efficient with linear time series data. ARIMA model was adjusted through grid search by adjusting parameters such as order of autoregression (p), differencing (d), and moving average (q). The model proved quite accurate for predicting patterns in users' payments behaviours based on transaction volumes and product usage over time. In contrast, the LSTM model, which was a deep learning algorithm, was implemented in TensorFlow. LSTM networks can also handle non-linear dependencies on time series, for example, sequential user behavior in financial markets. We trained the LSTM model with past financial transaction data, with inputs like amount of transaction, transaction frequency, and user behavior. Both models were evaluated using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) [6]. Table 2 shows ARIMA and LSTM's performance metrics on a portion of the time series (for 100,000 users for 6 months). According to Table 2, the LSTM model was more accurate than ARIMA in terms of MAE and RMSE, so it is likely that LSTM better captures non-linear user financial behavior. But the computation required for LSTM was much longer as deep learning models were complex.

Table 2. Performance Comparison of ARIMA and LSTM Models

Model	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Computational Time (Seconds)
ARIMA	0.056	0.073	120
LSTM	0.031	0.047	450

3.3. Association rule mining

We used the Apriori algorithm – one of the most popular association rule mining algorithms – to uncover associations between user actions and financial products. The data set was converted to a transaction-like model with each user's transaction history as a list of items. This record included financial products – savings accounts, investment instruments, loans, insurance policies. Popular itemsets were mined using Apriori algorithm with a minimum support threshold of 0.02 and a confidence level of 0.5. This allowed us to detect deep connections between various financial products and investment actions of users [7]. The associated association rules were then probed to uncover hidden correlations – for example, the probability of users buying a certain investment after they had previously bought a savings account. For instance, we found the following association rules: Savings Account Investment Plan, which means users who bought a savings account were highly likely to also buy an investment plan in the next 6 months. Performance of association rule mining was evaluated with respect to the indicators like support, confidence and lift, which measure the strength and relevance of the found associations. These measures were used to ensure the associations discovered were meaningful and useful for the recommendation system [8].

4. Results and discussion

4.1. Performance of time series models

The experimental findings were that the LSTM model greatly outperformed the ARIMA model when it comes to prediction accuracy. The LSTM model's RMSE was 0.24 (ARIMA's RMSE is 0.35). This shows that LSTM was more able to identify complex patterns and non-linear trends within financial time series data, which is important to interpret complex user behavior in complex financial markets. In addition, the MAE of LSTM was 0.16 compared to the ARIMA model's MAE of 0.23 suggesting better prediction performance of the LSTM model. Spark's performance meant that the LSTM model could be trained on big data sets in a fraction of the time using legacy techniques such as ARIMA. Below you can see Table 3 below shows the performance metrics of each model, RMSE, MAE, and computation time. The LSTM model had less RMSE and MAE (Table 3), but trained more quickly than ARIMA [9]. But this trade-off was worth it for the better prediction accuracy. LSTM model (which captures complex non-linear trends) was deemed to be better suited for time series in the financial field.

Table 3. Performance Comparison of Time Series Models

Model	RMSE	MAE	Computational Time (Seconds)
ARIMA	0.35	0.23	180
LSTM	0.24	0.16	350

4.2. Association rule mining results

During the association rule mining phase, some interesting data regarding user behavior and product preferences were collected to calibrate the recommendation system. The most striking association rule found was: "People who make a lot of stock products, are likely to invest in retirement products in the next six months." This rule makes it clear that user behaviour in the stock market is highly correlated with interest in retirement products. Other significant association rules were:

"Members who take out mortgages may have to pay for home insurance in the following month."

"People who have high volume transactions with mutual funds will invest in retirement plans in the following quarter. "

These association rules were fed into the hybrid recommendation engine to recommend related financial products based on users' predicted future behavior. The top five associations rules we found in Table 4 below [10], as well as the values of support, confidence and lift, representing the strength of these relationships. In Table 4, the rules with the highest confidence and lift values are the strongest ones that link some financial patterns with other product preference. These findings were key for crafting an individualized recommendation engine that would anticipate users' product needs in the future, and ultimately drive engagement.

Table 4. Top 5 Association Rules and Their Metrics

Rule	Support	Confidence	Lift
Stock-related products → Retirement plans (next 6 months)	0.05	0.75	1.85
Mortgage products → Home insurance (next month)	0.03	0.82	2.20
Mutual fund transactions → Retirement plans (next quarter)	0.04	0.78	1.95
High transaction volume in insurance → Investment in stocks	0.02	0.67	1.60
Low-risk financial products → Savings accounts (next 3 months)	0.06	0.70	1.75

4.3. Impact on financial recommendation systems

Having a combination of time series prediction and association rule mining in a recommendation system presented a huge leap in recommendation accuracy and user acceptance. Combining the predictive power of LSTM to predict user behavior, with association rule mining's relationship-driven predictions, made the system provide more relevant and timely suggestions. Several metrics like click-through rate (CTR) and conversion rate went up dramatically. In particular, the hybrid system yielded 20% higher CTR compared to conventional recommendation systems that did not incorporate time series analysis or association rule mining. Additionally, the conversion rate was 15% higher, which means users would be more inclined to take action on the new system's recommendations. The better results are thanks to a combination of temporal data (gathered by LSTM models) and user-product relationships (identified by the Apriori algorithm) [11]. The time window allowed for more precise estimations of when the users would be most likely to buy that product, and association rules ensured that recommended products matched users' preferences. These gains are also confirmed by the user conversion statistics shown in Table 5 which shows the difference between the hybrid system CTR and conversions with respect to the old model.

Table 5. Impact of Hybrid Recommendation System on User Engagement

Recommendation System	Click-Through Rate (CTR)	Conversion Rate
Traditional Recommendation	7.2%	5.1%
Hybrid System (with Time Series and Association Rule Mining)	8.6%	5.9%

5. Conclusion

The research discusses the benefits of using a mixture of time series prediction and association rule mining to develop a more robust and tailored financial recommendation system. The use of ARIMA and LSTM models to forecast user behavior in the future enables the system to predict financial product demand more precisely and sooner. The LSTM model in particular is much more accurate than ARIMA because it identifies non-linearity in the data that ARIMA does not. Second, association rule mining — in this case, the Apriori algorithm — offers us valuable information about how users interact and what products they like to enable the system to suggest suitable financial products based on their past interactions. These methods can be implemented in a big data environment, with Hadoop and Spark, which makes the system capable of handling very large data sets and offering real-time recommendations to users. Tests show that recommendations are significantly more accurate and user-engaged with a 20% higher click-through and conversion rate compared to standard recommendation algorithms. This work not only illustrates the possibilities of powerful data analytics in finance, but also lays the groundwork for further research on applying time series analysis and association rule mining to other areas. In conclusion, the hybrid recommendation engine that we have created in this research is an exciting development in financial services. Combining predictive modeling with pattern recognition, the system provides an adaptable and dynamic solution for recommending investment products. It can be further optimised through future research using reinforcement learning or deep reinforcement learning, to optimize the system for changing user behavior and market.

Contribution

Yaoyu Chen and Yichen Xu contributed equally to this paper.

References

[1] Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied soft computing*, 90, 106181.

[2] Cheng, D., Yang, F., Xiang, S., & Liu, J. (2022). Financial time series forecasting with multi-modality graph neural network. *Pattern Recognition*, 121, 108218.

[3] Lazcano, A., Herrera, P. J., & Monge, M. (2023). A combined model based on recurrent neural networks and graph convolutional networks for financial time series forecasting. *Mathematics*, 11(1), 224.
[4] Majumdar, S., & Laha, A. K. (2020). Clustering and classification of time series using topological data analysis with applications to finance. *Expert Systems with Applications*, 162, 113868.
[5] Ranaldi, L., Gerardi, M., & Fallucchi, F. (2022). Crypto net: using auto-regressive multi-layer artificial neural networks to predict financial time series. *Information*, 13(11), 524.
[6] Dixon, M., & London, J. (2021). Financial forecasting with α-rnns: a time series modeling approach. *Frontiers in Applied Mathematics and Statistics*, 6, 551138.
[7] Shternshis, A., Mazzarisi, P., & Marmi, S. (2022). Measuring market efficiency: The Shannon entropy of high-frequency financial time series. *Chaos, solitons & fractals*, 162, 112403.
[8] Ramakgasha, M. J., Thaba, T. K., & Rudzani, N. (2024). Agricultural production and agricultural employment rate in South Africa: Time series analysis approach. *International Journal of Economics and Financial Issues*, 14(4), 148-153.
[9] Malladi, R. K., & Dheeriya, P. L. (2021). Time series analysis of cryptocurrency returns and volatilities. *Journal of Economics and Finance*, 45(1), 75-94.
[10] Bielinskyi, A. O., Hushko, S. V., Matviychuk, A. V., Serdyuk, O. A., Semerikov, S. O., & Soloviev, V. N. (2021). *Irreversibility of financial time series: a case of crisis*.
[11] Xu, M., Shang, P., & Zhang, S. (2021). Multiscale Rényi cumulative residual distribution entropy: reliability analysis of financial time series. *Chaos, Solitons & Fractals*, 143, 110410.