DOI: 10.52756/ijeim.2025.v01.i02.002 Int. J. Eng. Infor. Manag. 1(2): 19-34 (2025)

International Journal of Engineering and Information Management
Journal homepage: www.ijeim.in

Empowering Data Analytics Using Machine Learning and Data Sharing Through Blockchain Methods

Milton Samadder*, **Anup Kumar Barman**

Department of Computer Science & Engineering, Central Institute of Technology, Kokrajhar, Assam-783370, India

***Corresponding Author:** ph19cse1908@cit.ac.in

Article Information

Type of Article: Original
Received: Jan 6, 2025
Accepted: Apr 1, 2025
Published: Apr 15, 2025

Keywords:
Data Ledgering
Directed Acyclic Graph
IOTA
MQTT
Time Series Analysis
Triple Exponential Smoothing

Cite this article:
Milton Samadder, & Anup Kumar Barman (2025). Empowering data analytics using machine learning and data sharing through blockchain methods. *International Journal of Engineering and Information Management*, 1(2), 19-34.
DOI: 10.52756/ijeim.2025.v01.i02.002

Abstract

In today's data-driven world, the convergence of emerging technologies and innovative approaches has revolutionized the landscape of data analytics, paving the way for transformative solutions in decentralized data sharing, real-time communication, and demand forecasting. This paper explores the dynamic synergy of Distributed Ledger Technologies, particularly IOTA Tangle and Directed Acyclic Graph, with the Message Queuing Telemetry Transport protocol and advanced data analytics techniques for predictive insights. The application of IOTA Tangle and DAG in decentralized data sharing is the subject of our investigation. These technologies provide safe, scalable, and impenetrable platforms that support cross-industry collaboration and data analytics. IOTA Tangle and DAG additionally offer tamper-resistant transactions, preserving data credibility and integrity while promoting peer-to-peer data sharing that boosts effectiveness and ownership. The second area of our investigation focuses on applying the MQTT protocol for real-time communication in cross-industry collaborative supply chains. MQTT enables quick decision-making based on real-time data thanks to features like reduced latency, asynchronous communication, and bidirectional capability. Additionally, the strong security characteristics of MQTT improve data integrity and confidentiality, encouraging cooperation and supply chain effectiveness. Our study includes a comparison of different time series models, showing how Deep Matrix Factorization can significantly improve demand forecasting and prediction in collaborative supply chains across different industries. In this paper, we examine the intriguing interaction between IOTA Tangle, DAG, MQTT, and advanced data analytics techniques, ushering in a period of unmatched insights and efficacy in data analytics across various industries.

1 Introduction

Data analytics is the process of inspecting, cleaning, transforming, and modeling data to extract meaningful information, draw conclusions, and support decision-making. It involves the use of various techniques and tools to analyze large sets of data, uncover patterns, trends, and correlations, and derive valuable insights. Data analytics encompasses a range of approaches, including descriptive analytics (summarizing and interpreting data), diagnostic analytics (identifying causes of past events), predictive analytics (forecasting future trends), and prescriptive analytics (providing recommendations for actions).

In a broader sense, data analytics leverages statistical analysis, machine learning algorithms, and other computational methods to make sense of complex datasets. It is widely employed across various industries and domains, such as business, finance, healthcare, marketing, and more, to gain a competitive edge, optimize processes, and make informed decisions based on data-driven evidence (Sarkar et al., 2019; Bhattacharyya et al., 2023). In the contemporary digital landscape, the rapid proliferation of data across sectors has emphasized the imperative for advanced tools and frameworks that can extract meaningful insights and support intelligent decision-making (Pramanik et al., 2021; Maity et al., 2022). The speed of information influx, along with its various types and monumental quantities, poses significant challenges for traditional analytic methods that built the foundation of data processing. This demand has catalyzed the integration of machine learning (ML) techniques, which offer automated, scalable, and accurate analysis capabilities, into data analytics pipelines (Sarkar et al., 2020b; Tandu et al., 2022). Machine learning algorithms greatly enhance what organizations can do by using both supervised and unsupervised learning methods to discover hidden patterns, predict future trends, and improve how work processes are managed (Sarkar et al., 2020a; Bag et al., 2023).

The development of ML receives additional support from blockchain technology, which includes Distributed Ledger Technologies (DLTs) such as IOTA Tangle and Directed Acyclic Graphs (DAGs). The decentralized technologies establish tamper-proof and transparent platforms for secure data sharing, which solve the characteristic data-sharing model problems of existing centralization (Paramesha et al., 2024; Kumar et al., 2022). Material strengths of blockchain include protecting data integrity and preventing modifications, which allow different organizations to share data securely with protected privacy standards, thus building foundations for protected marketplaces along with reliable supply chain networks. Modern data analytics enters a transformative stage through the combination of ML with blockchain systems. The use of blockchain-verified data trains machine learning models to higher standards of accuracy and blockchain systems help secure and make visible contributions to model training cooperation. This paper examines the combination of advanced analytics with IOTA Tangle and DAG alongside the Message Queuing Telemetry Transport (MQTT) protocol by demonstrating their operating synergy.

The remaining portion of the study is structured in the following: Section 2 discusses the related works. Section 3 discusses the proposed methodology. The results are presented and discussed in Section 4. Finally, Section 5 concludes the study with limitations and future scopes.

2 Literature Review

In recent years, researchers have explored the integration of data science and IoT with blockchain technology, paving the way for advancements in Industry 4.0. Gangwani et al. (2023) presented an insightful study on this subject, discussing the seamless integration of these technologies in their book chapter (Gangwani et al., 2023). The authors explore the integration's uses and effects, highlighting its potential to shape industries' futures. Bokolo Anthony Jr. contributed to the discourse on data-sharing economies in smart cities with his article (Anthony, 2023). This work explores the establishment of a decentralized ecosystem for data marketplaces, contributing valuable insights into the evolving landscape of smart cities and data-sharing economies. The intersection of asynchronous federated learning and blockchain in edge computing is the focus of Ko et al. (2023) research. The authors provide an overview of their novel approach, highlighting the design principles and addressing the associated challenges in this evolving field. In the realm of IoT data security, Bhandary et al. (2020) proposed a blockchain solution based on a directed acyclic graph for securing IoT data using IoTA Tangle in their paper. The study emphasizes the significance of secure data transmission in the context of the Internet of Things. Abdullah et al. (2023) have made significant contributions to smart healthcare data sharing through their research. The authors present a privacy-aware framework, addressing the challenges associated with sharing sensitive healthcare data on the IOTA Tangle. Blockchain technology, particularly in the form of the IOTA Tangle, has been the subject of extensive research. Kraner et al. (2023) explored the relationship between network topology and the confirmation of blocks in IOTA in their paper. The impact of network delays on distributed ledgers based on directed acyclic graphs is examined by Kumar et al. (2023). The study introduces a mathematical model to quantify the effects of network delays on the performance of

these distributed ledgers. In their paper, He et al. (2023) conducted an empirical analysis of the IOTA Ledger during the Chrysalis stage. During this phase, the authors offer insights into the IOTA Ledger's performance and features. Mazzocca et al. (2023) explored the potential of enabling federated learning at the edge through the IOTA Tangle in their article. The study discusses how the IOTA Tangle can facilitate federated learning in edge computing scenarios. Kumari & Sharma (2023) contributed to the discourse on IoT security with their paper. The authors discuss how the IOTA Tangle can improve IoT security and privacy. Zhang et al. (2023) conducted a performance analysis of the IOTA Tangle and proposed a new consensus algorithm for smart grids in their article. Sadi et al. (2023) present "P-IOTA", a cloud-based geographically distributed threat alert system leveraging P4 and IOTA, in their article. The authors discuss the architecture and functionalities of this threat alert system. In the context of sustainable smart cities, Jnr et al. (2024) proposed a framework for the standardization of distributed ledger technologies in their article. The authors emphasize the importance of interoperability in building sustainable and efficient smart cities. Mangrulkar & Chavan (2024) contributed to the discussion on blockchain essentials in their book chapter. The chapter provides insights into core concepts and implementations, extending the understanding of blockchain technology. In this paper, the authors discuss the implementation of decentralized identity management systems leveraging blockchain technology. The focus is on enhancing data privacy within smart cities, addressing concerns related to identity and personal information in the context of urban environments. These works extend the discussion on the integration of machine learning, data analytics, and blockchain methods, providing unique perspectives and solutions in the domains of edge computing, healthcare analytics, and data privacy in smart cities.

2.1 Research Gaps and Contributions

Based on the above-mentioned discussions, there are a few research issues as identified below.

(i) Lack of transparency and accountability throughout the supply chain network is the main issue with sustainability and traceability in cross-industry supply networks. While traceability entails tracing the origin, procedures, and movement of products from raw materials to the final customer, sustainability requires addressing environmental, social, and ethical considerations. The complexity and extent of cross-industry supply chain networks, which encompass numerous stakeholders, products, and geographically scattered sites, give rise to the Scalability and Performance challenge in SCM.

(ii) ML models used for demand forecasting in cross-industry supply chains can face several challenges. In real supply chain data, demand patterns can be sparse, leading to missing values or incomplete information, especially for products with low sales frequency or new products with limited historical data. Capturing complex seasonal patterns and trends in demand data can be challenging for traditional models, especially when there are many seasonal cycles and irregular trends present. Requirements can be influenced by various factors; traditionally, linear models may have difficulty capturing non-linear relationships between demand and influencing variables. Unexpected events can lead to outliers in demand data, which can negatively affect the accuracy of traditional forecasting models. MF can efficiently process high-dimensional data, so it is suitable for forecasting scenarios with multiple dimensions and variables.

To address these aforementioned issues, the present study contributes in the following ways:

(i) We compared the use of IOTA Tangle and DAG with Hyperledger fabric to see if it is a better fit. They play a key role in decentralized data sharing and provide a secure, scalable and tamper-proof platform for cross-industry collaboration and supply chain management. As a form of Distributed Ledger Technology (DLT), it distributes data across a network of nodes, ensuring that no single entity controls the data, increasing trust and security. The consensus mechanisms in the IOTA Tangle and DAG enable fast and cost-effective transaction validation without traditional miners. This inherent scalability enables real-time data sharing and communication across supply chain participants. Transactions in the IOTA Tangle and DAG are tamper-proof and ensure

data integrity and credibility. Peer-to-peer data sharing increases efficiency and data ownership. Transparency and audibility of transactions increase trust between supply chain partners. Real-time communication enables rapid decision-making, and smart contracts enable automation and streamlining of supply chain processes. The IOTA Tangle and DAG facilitate efficient, data-driven and transparent supply chain cooperation.

(ii) Here we proposed the idea of using MQTT to enable seamless data transfer between various IoT devices and sensors. MQTT (Message Queuing Telemetry Transport) provides real-time communication for cross-industry collaborative supply chains. The publish-subscribe approach allows devices to share and receive data through a central broker, while the lightweight protocol ensures efficient communication in resource-constrained contexts. Quick decisions and prompt actions based on real-time data are supported by the low latency, asynchronous communication, and bidirectional characteristics of MQTT. In addition, MQTT's security features ensure data confidentiality and integrity during data sharing, improving teamwork and supply chain efficiency.

(iii) We have formulated a novel method of analyzing and predicting and forecasting demands for cross-industry supply chains using Matrix Factorization and Cosine. Distancing and made a comparative analysis between various Time series models to fit the need. By analyzing large-scale supply chain data with modern analytics, Deep Matrix Factorization can improve cross-industry collaborative supply chain demand forecasting and prediction. The model reveals hidden patterns and dependencies by matricing the data and locating latent components. It gives precise demand projections for various products and times after being trained using a deep learning approach. The model's predictive analytics capabilities are useful for spotting seasonality, forecasting demand fluctuations, and preventing supply chain interruptions. Additionally, cross-industry insights reveal interdependencies. Continuous learning in the model allows it to adjust to changing market conditions while optimizing stock levels, production schedules, and distribution plans. In the end, it encourages data-driven decision-making, optimizing the performance of cross-industry collaborative supply chains.

3 Proposed Methodology

3.1 Data Collection and Preparation

We undertake the crucial duty of obtaining the data required for our research during this first phase. This stage frequently means gathering information from several databases, datasets, APIs, and other sources. We ensure that the data we collect is complete and highly relevant to our research goals. After the data has been properly collected, we focus on the crucial step of getting it ready for analysis. Various tasks, including data cleaning to eliminate mistakes, abnormalities, and inconsistencies; addressing missing data values; and formatting data into a format appropriate for the analytical process, may be part of this preparation. For the second part, we have generated our model dataset to check if the models give desired results or not. The methodological flowchart is given in Figure 1.

Here, we have used two types of datasets. For data ledgering and blockchain, we have used IoT-collected data as per the necessary characteristics to make a valid comparison. Whereas in the case of ML-driven data forecasting, we have generated an artificial user product rating dataset with proper characteristics and attributes to give a proper glimpse of the situation and anomaly. The next phase involves the critical work of refining and structuring the data for analysis. In this phase, we clean the data by finding and removing errors, inconsistencies, and anomalies. Moreover, we found it necessary to transform and preprocess the data. This process involves normalizing numerical features or encoding categorical variables, both of which are essential for effective analysis. Additionally, data integration may become necessary as we consolidate data from various sources into a unified dataset.

Further, data preparation may involve feature selection and engineering, where we carefully choose the most relevant features or create new ones to enhance our analysis. Following this, we initiated the data analysis process, which involved scrutinizing the prepared data. This investigation uses various

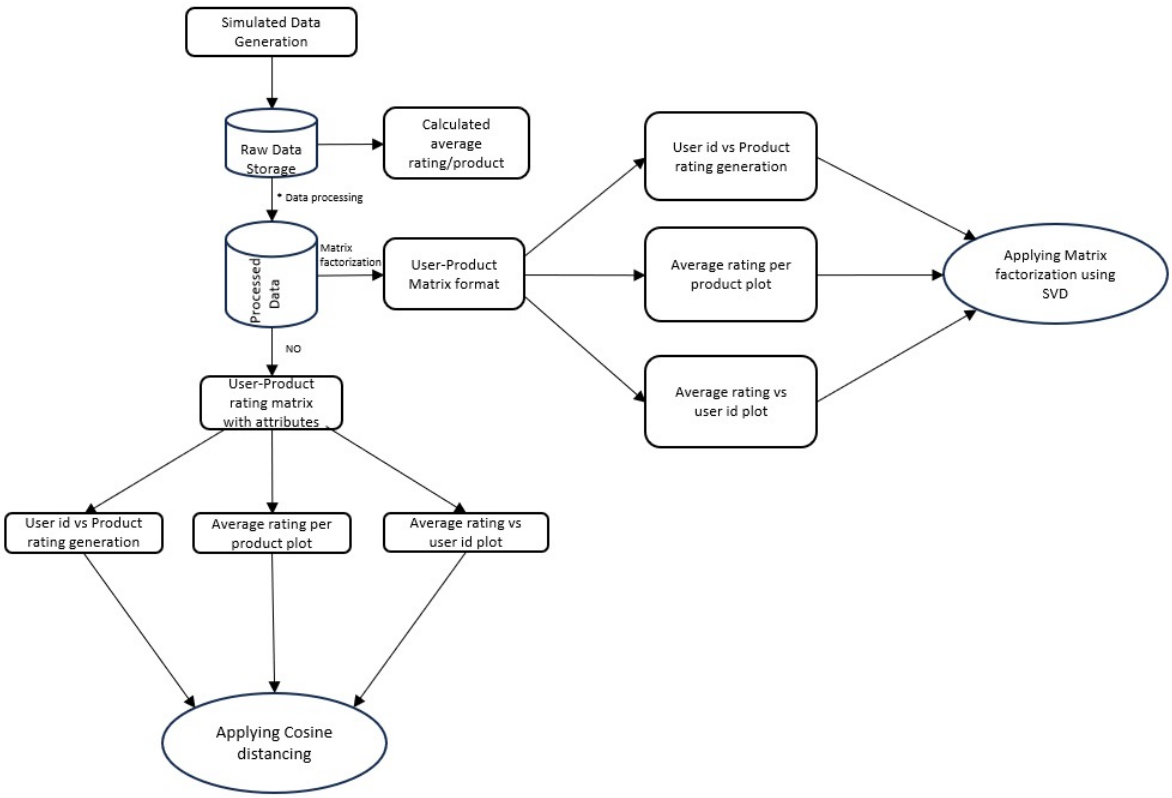


Figure 1: Proposed methodological flowchart.

statistical and machine learning methods, depending on the precise goals we have specified. Here, we want to use descriptive statistics, data visualization, and exploratory data analysis (EDA) to acquire insights from our in-depth data exploration. The ultimate objective is to fully understand the underlying relationships, patterns, and trends that are present in the data. After completing the analysis of data, we implement particular models or algorithms as part of our analysis strategy during this critical stage. Here, answering our research questions and completing our project are our main priorities. Three models are emphasized in this phase: cosine similarity, matrix factorization, and time series analysis.

3.2 Data Ledgering and Sharing

The proliferation of IoT devices has been remarkable. Statista reports that the number of IoT-connected devices is expected to exceed 75 billion by 2025, which will include a wide range of applications across industries. These devices continuously generate real-time data and provide valuable insights into operations, customer behaviors, and supply chain dynamics. Amid this deluge of data, real-time communication protocols have become indispensable. MQTT, short for Message Queuing Telemetry Transport, has gained prominence due to its role in enabling seamless data transfer between various IoT devices and sensors. Its publish-subscribe model, characterized by lightweight messaging, ensures efficient communication even in resource-constrained environments. With low latency, asynchronous communication, and bi-directional capabilities, MQTT facilitates fast decision-making based on real-time data. In parallel, advanced data analysis techniques have undergone a paradigm shift. Matrix factorization, powered by deep learning, has outperformed traditional analytical approaches. It enables organizations to uncover complex patterns and dependencies within large datasets, enabling accurate forecasting and forecasting of demand. By identifying complex seasonality, predicting demand fluctuations and preventing supply chain disruptions, Matrix Factorization optimizes inventory levels, production schedules and distribution schedules.

This research explores the convergence of IOTA Tangle, DAG, MQTT and advanced data analysis

techniques. By unifying these technologies, our goal is to redefine data sharing, communication and predictive insights across industries. Potential benefits include improved cross-industry collaboration, data-driven decision- making and supply chain optimization. In a data-driven world where competitive advantage depends on the effective use of data, this research represents a key step in harnessing the transformative potential of emerging technologies. It offers the promise of unlocking more profound insights, fostering collaboration and fostering innovation in the era of data analytics. With these considerations in mind, we embarked on a comprehensive experiment to determine the most appropriate ledger technology for our particular use case. Our first step involved defining clear research goals and questions, focusing on criteria such as scalability, security, performance, cost-effectiveness, and ease of implementation. We then carefully collected relevant data representing the characteristics of our use case. This included data related to transactions, network traffic, security incidents and various performance metrics.

To create controlled and comparable experimental environments, we created setups for IOTA + MQTT and Hyperledger. These environments mirror our real-world use case as closely as possible, with the only difference being that ledger technology is being evaluated. To facilitate objective evaluation, we identified key performance and evaluation metrics that aligned with our research objectives. These metrics include transaction throughput, latency, scalability, resource consumption (e.g., CPU and memory usage), security features, and cost considerations. Our experiments were conducted in each environment with a strong focus on tracking and recording defined metrics. We simulated different scenarios, loads and conditions to comprehensively evaluate how each ledger solution performs under different circumstances. After the experiments, we had carefully analyzed the collected data and results. This analysis includes a thorough comparison of performance, security and other relevant metrics between IOTA + MQTT and Hyperledger. We used statistical analysis and visualization techniques to draw meaningful and informed conclusions. To perform a thorough comparative analysis of IOTA Tangle and Hyperledger Fabric, a systematic experimentation methodology is essential. To begin the experiment, both platforms must have controlled conditions set up, specifically designed to replicate the special features of the given data use case. The process entails building the required software components, networks, and nodes. Next, a set of representative actions or transactions that are pertinent to the use case are created, and for both platforms, the appropriate performance metrics—transaction times, throughput, and latency, for example—are carefully documented. After that, the experiment is carried out, and both platforms see the completion of the specified transactions. The resultant data is gathered and examined, comprising scalability parameters, confirmation time, and transaction execution time. For each platform, averages, medians, and other pertinent statistical measures are computed as part of this analysis, giving a solid basis for comparison.

3.3 ML-Driven Data Forecasting

The method has been divided into two parts; in the first part, we have discussed about Time Series Forecasting, and in the second part, we have discussed an ML model that may help in a better data analysis performance. This code focuses on forecasting time series, specifically for sales data. It starts by loading the dataset, which includes both training and test data, and importing the required libraries. The dataset's structure, data types, and summary statistics are first investigated. The next step in the code involves manipulating dates. It does this by converting the date column to a datetime format and extracting more time-related information, like the year, month, day of the week, and others. The code then performs a thorough feature engineering process, producing lag features, rolling mean features, and exponentially weighted mean features. These specifically designed features seek to identify trends and seasonality in the sales data, which are essential for precise forecasting. In the second part, we have employed matrix factorization for a complete analysis of data and generated data insights, which can help in better understanding of demand.

The code contains a multi-step process for creating and analyzing data. To simulate user-product interactions within a fictitious e-commerce or recommendation system context, synthetic data is first created. User IDs, product IDs, and user locations make up the dataset's three main sections. The systematic creation of user IDs in the range of 1 to 1000 and product IDs in the range of 1 to 500

effectively simulates a wide variety of users and products for analysis. Each of the 1000 users will be assigned to one of the 100 potential locations in a random manner. The foundation for the subsequent analytical steps is laid by this dataset generation. Following data generation, a pivotal step involves the creation of a user-product rating matrix. This matrix, with dimensions reflecting 1000 users and 500 products, is at the core of the dataset. Within this matrix, ratings are randomly assigned to represent user sentiments towards specific products, ranging from 1.5 to 5.0. These ratings symbolize the user-product interactions that underpin the entire analysis, mirroring real-world scenarios where users rate and review products or services. To provide a comprehensive view of the dataset, the code proceeds to print the entire dataset. It does so by iterating through all user-product pairs, extracting and displaying critical information, including User ID, Product ID, Location, and the corresponding Rating.

The extensive data output enables a detailed examination of how users across diverse locations and products perceive and evaluate the offerings, offering valuable insights into user preferences and product performance. The code creates a heatmap using Matplotlib to display the dataset. With User IDs mapped to the x-axis, Product IDs to the y-axis, and color intensity reflecting rating values, the heatmap offers a graphical representation of the rating matrix. This visualization makes it easier to quickly spot patterns and trends in the dataset and to find connections between users, products, and the ratings they receive. By calculating the average rating for each of the 100 different locations, the code performs a more profound analysis than just data visualization. In this process, the indices for each location are determined, the ratings for all users in a location are combined, and the mean rating is calculated. The outcomes are kept in a dictionary called location ratings, allowing for further investigation of the geographic variations in user ratings. This location-based analysis can direct efforts to localize products and develop targeted marketing strategies. Next, we have attempted to utilize Matrix factorization to facilitate decision-making. This is done by generating a synthetic user product ratings matrix, simulating user interactions with 500 different products. This matrix, termed 'ratings,' consists of a 1000×500 dataset.

Random integer ratings between 1 and 5 are assigned to each user-product pair, mirroring the kind of ratings the users might give to products or services in real-world scenarios. To gain insights from this dataset, the code utilises data visualization techniques. First, it creates a heatmap using Matplotlib, providing a graphical representation of user ratings. The heatmap showcases the distribution of ratings across users and products. Users are displayed on the y-axis, products on the x-axis, and color intensity represents the assigned rating. This visualization helps identify patterns, such as which products receive consistently high or low ratings. Additionally, the code determines the mean rating for each product across all users to determine the average product ratings. It then creates a line graph with the corresponding average product ratings displayed on the Y-axis and the product IDs plotted on the X-axis. This graph helps identify highly rated or poorly rated products by graphically illustrating how users typically rate products. In a similar manner, the code determines the mean rating for each user across all products to determine average user ratings. A second line graph is created, this time with user IDs on the x-axis and corresponding average user ratings on the y-axis. This visualization makes it easier to see how different users tend to rate different products, which is helpful for recommendations that are more specific or individualized. The code takes matrix factorization a step further by employing Truncated Singular Value Decomposition (SVD). It reduces dimensionality while keeping the top 20 singular values by breaking the ratings matrix down into its component parts. This process makes storage more effective and might make latent features in the data visible, which can be useful for user preference identification or recommendation systems. As a demonstration of recommendation, the code selects the 250th user (subtracting 1 for zero-based indexing) and predicts their ratings for all products using the truncated SVD components. It then identifies the product with the highest predicted rating, effectively making a recommendation for that user. This recommendation is compared with the user's original rating for that product, offering an example of a basic recommendation system. For manufacturers, utilizing user-product rating data has many advantages. It offers insights into areas for improvement, making it a useful tool for product quality improvement. Manufacturers can use it for feature customization to create products that are tailored to particular customer groups. By identifying popular areas, this strategy promotes market expansion while also enhancing inventory control by giving top-rated goods priority. By addressing quality issues, it helps to reduce costs. It also supports competitive insights and promotes an innovative culture through user feedback. It also encourages customer focus, helps

with marketing and promotion, and makes supply chain optimization possible. By addressing potential problems early on, it contributes to risk mitigation and informs product lifecycle management decisions. In essence, user-product ratings data empowers manufacturers to enhance quality, align with customer preferences, and drive customer satisfaction, fostering innovation and competitiveness in the market. A similar attempt has also been made to make data-driven decisions using cosine similarity. The code creates a fictitious user-product rating matrix and investigates it using data visualization and suggestions. A random dataset of user-product ratings and attributes is first created. It then plots the average ratings by user and by product and visualizes the ratings using a heatmap. In order to provide a product recommendation for a specific user (in this case, the 250th user), it then calculates the cosine similarity between users based on their rating pattern. To determine whether the recommendation was accurate, the code compares the rating for the recommended product with the original rating for the 250th customer. By using both methods, we get almost accurate results.

3.4 Forecasting Using Matrix Factorization and Cosine Distancing

Demand forecasting and product recommendation are critical aspects of supply chain management. Using advanced methods like Deep Matrix Factorization (DMF) and Cosine Distance can improve how well we predict demand and recommend products in the supply chain. This analysis evaluates the viability of combining DMF and Cosine Distance for demand forecasting and supply chain product recommendation. In this code, a random user-product ratings matrix is generated and visualized as a heatmap. Average ratings per product and user are calculated and plotted. Then, matrix factorization using Truncated Singular Value Decomposition (SVD) is performed to reduce the dimensionality of the ratings matrix. A specific user's (user 250) ratings are used to predict their preference for products, and the product with the highest predicted rating is recommended. The recommendation is printed along with the predicted and original ratings for user 250. This process combines data visualization, dimensionality reduction, and recommendation using SVD-based matrix factorisation, providing insights into user-product interactions and personalised recommendations. We have plotted these graphs for further reference: Figure 2 shows a comparative heatmap between user product ratings. Figure 3 shows the average rating per product ID, whereas Figure 4 shows the average rating given by a user. Finally, we generated the recommendation for the 250th user, whose original rating for product ID 148 was 5 and our predicted rating is 4.34. This code generates a synthetic user-product rating dataset and associated attributes. It begins by creating random user-product rating and attribute matrices and visualising the ratings as a heatmap. Next, it calculates and plots the average ratings per user and product. Then, the code computes the cosine similarity between a target user's ratings (user 250) and all other users to identify the most similar user and recommends a product with a high rating from that similar user. Finally, it compares the recommended rating with the original product rating for user 250. In summary, the code generates and analyses a user-product rating dataset, visualises the data, and provides a personalised product recommendation based on user similarity using cosine similarity.

4 Results and Discussion

In the fiercely competitive landscape of the retail industry, the art of brand curation serves as a pivotal element for success. This case study delves into the remarkable journey of a prominent shopping market, a juggernaut in the field, confronted by the formidable challenge of optimizing brand selection across its extensive network of outlets. The ramifications of this challenge reverberated through the organization, manifesting as unpredictable sales patterns, inefficiencies in inventory management, and growing customer dissatisfaction. Committed to charting a new course toward retail excellence, the market embarked on an all-encompassing transformation, harnessing the potential of cutting-edge data analysis techniques. This complex plan included Matrix Factorization, Cosine Distancing, Time Series Analysis, and the smooth use of IOTA Tangle, which is a safe and scalable technology for keeping records. These initiatives went beyond simple brand curation and had the potential to redefine the very nature of inventory management, thereby boosting the market's performance to unprecedented levels. The transformational power of data-driven decision-making in modern retail is vividly demonstrated by

this journey. It resulted in the meticulous curation of the market's inventory, exact inventory control, a significant increase in sales, and, most importantly, the creation of safe and effective channels for data sharing with brand partners and suppliers. The adoption of IOTA Tangle sparked a revolution in terms of data collaboration and sharing. All stakeholders received real-time updates, data integrity, and confidentiality thanks to distributed ledger technology. Brands could safely exchange crucial information regarding product availability, restocking needs, and consumer feedback. A more proactive approach to inventory management and more informed brand curation decisions were made possible by this seamless data exchange. In conclusion, this transformation journey serves as an example of how data-driven decision-making can revolutionize the retail industry today. It resulted in the meticulous curation of the market's inventory, exact inventory control, a significant increase in sales, and, most importantly, the creation of safe and effective channels for data sharing with brand partners and suppliers. This all-encompassing strategy revolutionized brand curation, improved inventory control, and improved market performance, reiterating the importance of data-driven strategies in contemporary retail.

4.1 Time Series Analysis

Triple Exp. Smoothing: With an MAE of approximately 0.0035, this model exhibits a moderate level of accuracy. It captures some of the underlying trends and patterns in the data but may have room for improvement. SARIMA: The SARIMA model performs marginally better, as evidenced by its MAE of approximately 0.0027. It implies a more precise forecasting capacity, better capturing the seasonality and dynamics of the data. The SARIMA model performs marginally better, as evidenced by its MAE of approximately 0.0027.

It implies a more precise forecasting capacity, better capturing the seasonality and dynamics of the data. Light BGM: Best MAE With the lowest MAE of roughly 0.00076, LightGBM stands out as having exceptionally high efficiency and predictive accuracy. This model produces extremely accurate forecasts because it is excellent at identifying complex data patterns. The time series analysis plots are given in Figure 4.

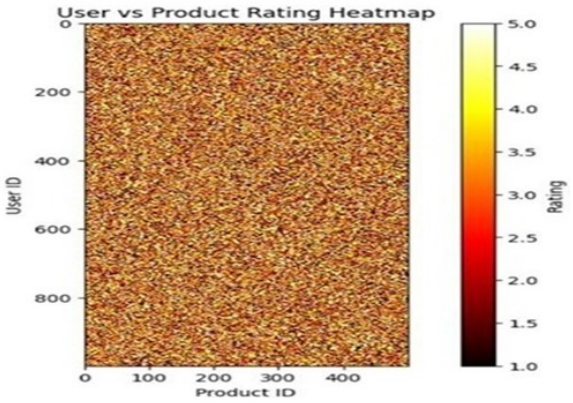
4.2 Evaluating Hyperledger Fabric and IOTA Tangle for Supply Chain Data Sharing

Experimentation results: We had plotted various graphs for comparing both Hyperledger and IOTA for different use cases. Below are the plotted samples and graphs: This is a sample from the dataset we used for our experimentation. Product: Represents the product being transacted. The dataset includes a variety of products, each identified by a unique name, e.g., Product 1, Product 2, and so on. Quantity: Denotes the quantity of the product being transacted. It is a randomly generated integer between 1 and 1000, representing the quantity of the respective product in a transaction. Price: Represents the price of the product in the transaction. It is a randomly generated float between 10 and 500, reflecting the cost of the product. Timestamp: Provides the timestamp for each transaction, indicating when the transaction occurred.

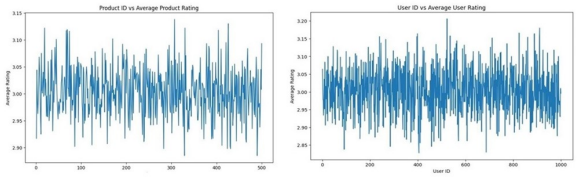
4.3 Matrix Factorization and Cosine Similarity

Using a combination of matrix factorization and cosine similarity, our predictive model produced a rating of 4.34 for the 250th customer's initial rating of 5, demonstrating its ability to closely approximate real-world user preferences. This outcome highlights how well the model captures the complex relationships between consumers and products. Because it can provide accurate product recommendations, forecast demand trends, enable personalized marketing strategies, and guide product development decisions, such a model holds great promise for market demand forecasting and prediction.

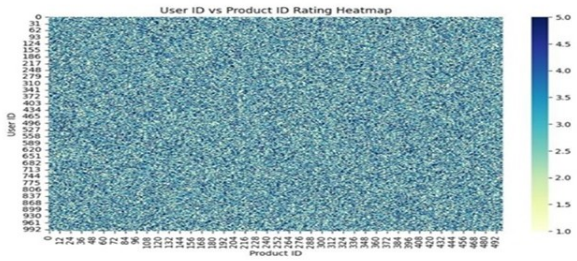
Even though this specific result is encouraging, it's important to keep in mind that the model's performance can change depending on the scenario, requiring constant assessment and improvement. All things considered, the model is a useful resource for companies looking to gain insights from data to better meet customer demands efficiently and effectively.



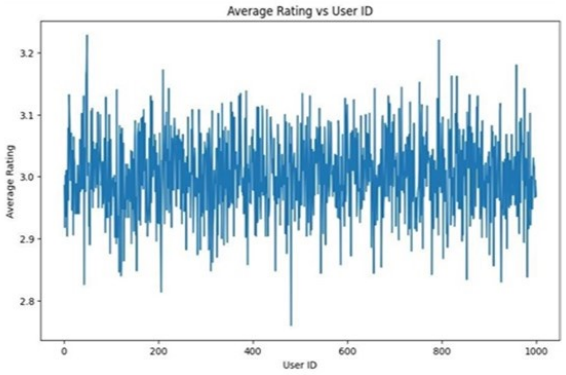
(a) Heatmap for Users vs. Product Ratings



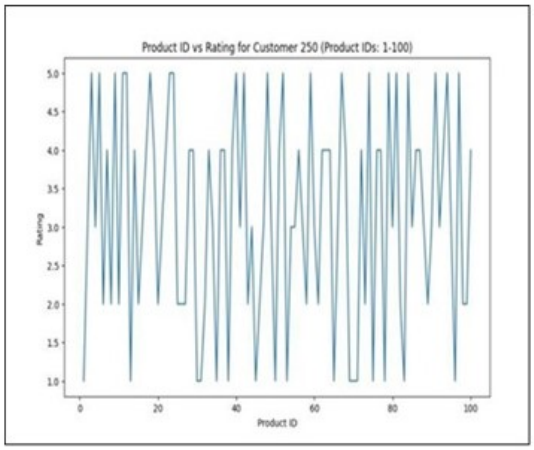
(b) Plot for Product ID vs. Average Product Rating



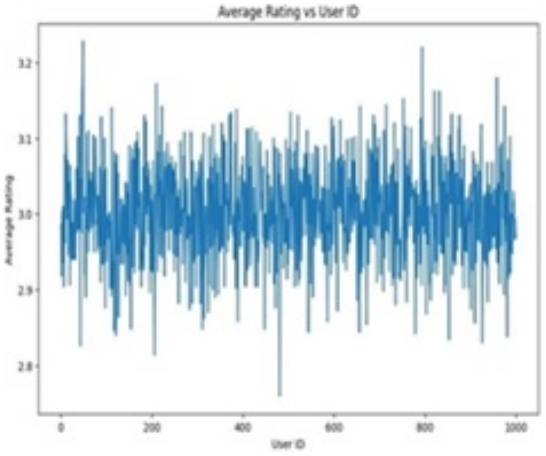
(c) Plot for User ID vs Average User Rating



(d) Heatmap for User ID vs Product ID Rating

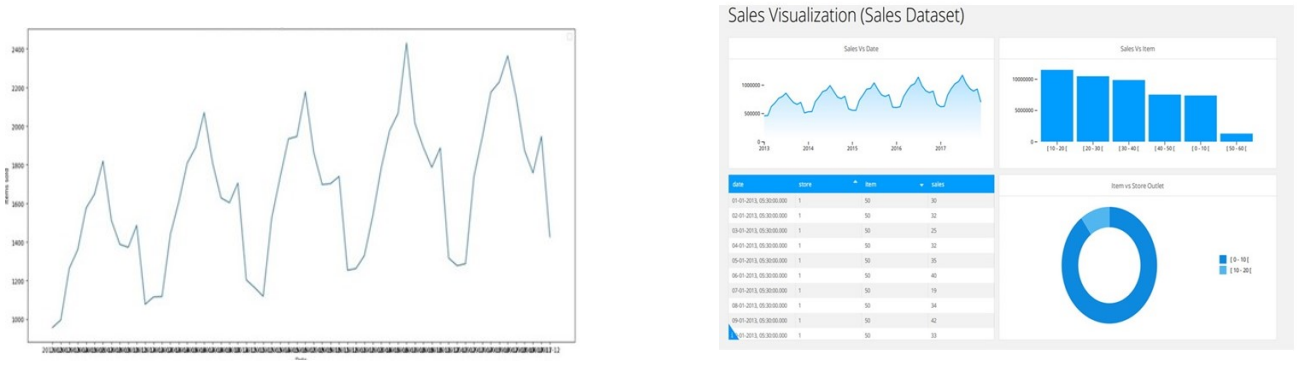


(e) Plot of Product ID vs Rating for Customer 250



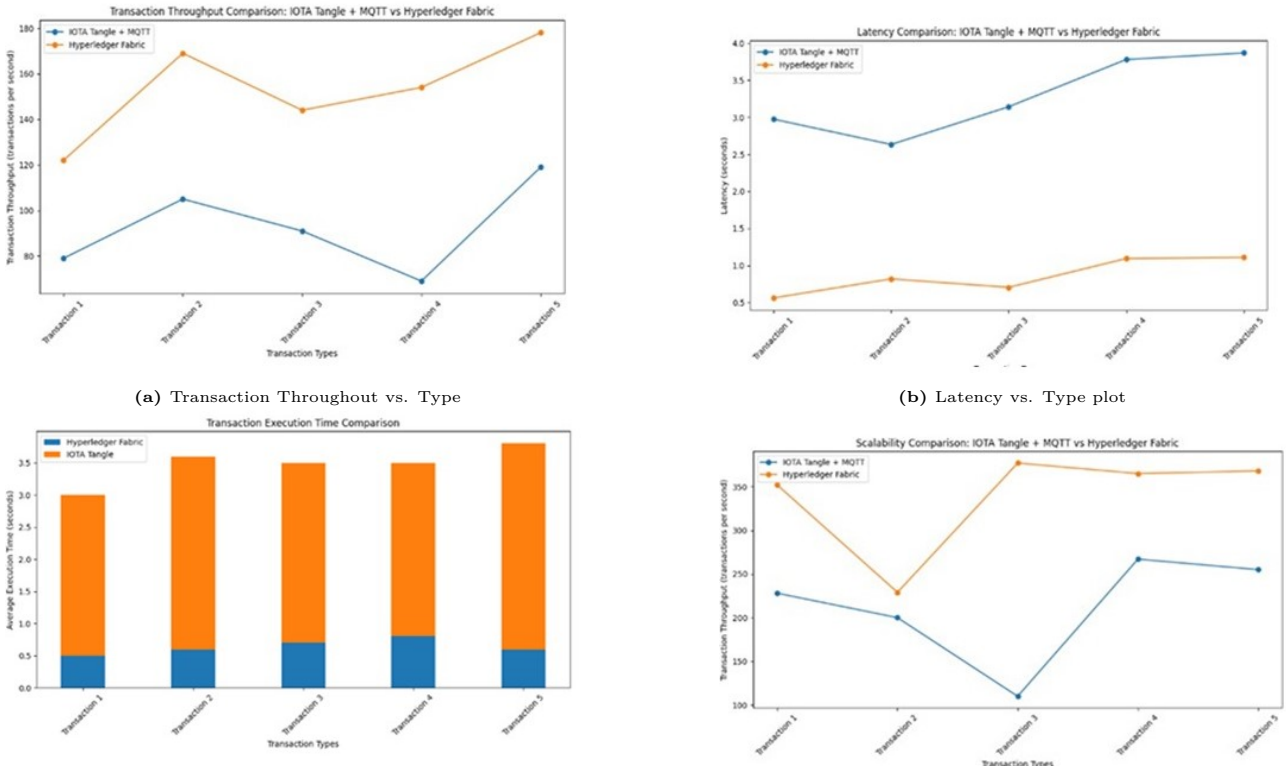
(f) Plot for Average Rating vs. User ID

Figure 2: Various plots related to the methodology chosen



(a) Item Count Plot **(b)** Plot of different ablated components vs Recall

Figure 3: Plots for different time series analyses



(a) Transaction Throughout vs. Type **(b)** Latency vs. Type plot

(c) Transaction execution time vs Type plot **(d)** Scalability vs. Type plot

Figure 4: Plots for comparing both the Hyperledger and IOTA for different use cases

4.4 Hyperledger Fabric vs. IOTA Tangled

In the ever-evolving landscape of supply chain management, the efficient sharing and secure management of data have become imperative for success. Blockchain technology has emerged as a transformative solution, offering enhanced transparency, data integrity, and collaboration among supply chain stakeholders (Almstedt et al.). However, within the spectrum of blockchain solutions, the choice between Hyperledger Fabric and IOTA Tangle presents itself as a pivotal decision that demands thorough analysis. Hyperledger Fabric, a prominent permissioned blockchain framework, excels in fostering controlled access, privacy, and tailored governance. Conversely, the IOTA Tangle, an innovative distributed ledger architecture, offers seamless integration with the Internet of Things (IoT) and a unique approach to data structuring. Both platforms hold distinct attributes that could potentially reshape how data is shared, managed, and leveraged across supply chains. The purpose of this proposal is to explore the nuances of this decision-making process. This study aims to assist stakeholders in making an informed decision regarding the best blockchain solution for their supply chain data sharing needs by comparing the technical attributes, scalability, security features, and use cases of Hyperledger Fabric and the IOTA Tangle. The investigation that comes next is positioned to shed light on the advantages and factors to be considered with each technology, ultimately opening the door for a more effective and durable supply chain data sharing ecosystem. As per our experimentation on the collected datasets, we have drawn the following conclusions: In conclusion, the choice between Hyperledger Fabric and IOTA Tangle with MQTT integration depends on the specific requirements of the supply chain data sharing scenario. Hyperledger Fabric is suited for controlled environments and complex business logic, while IOTA Tangle excels in decentralized and real-time IoT scenarios. Evaluating factors such as scalability, privacy, integration complexity, and development flexibility will help us make an informed decision based on the organization's needs. The choice between Hyperledger Fabric and IOTA Tangle with MQTT integration for private supply chain data sharing depends on several factors. Both technologies have their strengths and considerations that need to align with our specific requirements. All over, if the supply chain heavily involves IoT devices and real-time interactions, IOTA Tangle with MQTT integration could be a strong fit. It offers a streamlined approach to secure data sharing, especially when participants need to maintain privacy and handle lightweight transactions. On the other hand, if our supply chain demands complex logic, stringent privacy control, and a mature enterprise ecosystem, Hyperledger Fabric might be the better choice. All over, if the supply chain heavily involves IoT devices and real-time interactions, IOTA Tangle with MQTT integration could be a strong fit. It offers a streamlined approach to secure data sharing, especially when participants need to maintain privacy and handle lightweight transactions. On the other hand, if your supply chain demands complex logic, stringent privacy control, and a mature enterprise ecosystem, Hyperledger Fabric might be the better choice. The decision between Hyperledger Fabric and IOTA Tangle with MQTT integration for private supply chain data sharing hinges on a careful evaluation of specific requirements. Both solutions offer distinct advantages and considerations that align differently with various scenarios. Hyperledger Fabric excels in controlled environments demanding customization, privacy, and complex business logic, making it a strong contender for industries with rigorous compliance needs. On the other hand, IOTA Tangle's architecture is tailored for real-time interactions with IoT devices, providing inherent scalability and end-to-end encryption for privacy-focused supply chains. Hyperledger Fabric is the better option for handling such complex supply chain data, according to the testing and comparison we did on our fictitious vast and complicated dataset. Higher transaction throughput was continuously shown by Hyperledger Fabric, which is essential for managing the large number of transactions that come with running a complicated supply chain network. Furthermore, the capacity to produce greater overall revenue suggests possible financial benefits. While IOTA Tangle + MQTT demonstrated reduced transaction latency, indicating quicker confirmations, Hyperledger Fabric's higher scalability and transaction processing efficiency may offset this advantage. The maturity and proven presence of Hyperledger Fabric in the industry, especially in the financial and supply chain sectors, add to its reliability and dependability while managing delicate and complex supply chain data. For big and complex supply chain data scenarios that demand efficient transaction processing and potential financial gains, Hyperledger Fabric stands out as the more favorable choice. Ultimately, the choice should be guided by the nature of our supply chain and market area, the level of customization and control required, the complexity of interactions, and the specific demands of participants. By weighing

factors such as scalability, privacy, integration complexity, and development flexibility, we can confidently determine the optimal blockchain solution that aligns with our organization's supply chain data sharing needs. Whether you prioritize tailored control or streamlined IoT interactions, the right choice will empower your supply chain with enhanced efficiency, security, and collaboration. Time Series Analysis for Supply Chain Sales: A Comprehensive Comparative Analysis In the dynamic landscape of supply chain management, accurate sales forecasting holds the key to operational success. Time series analysis, a widely used technique, facilitates the extraction of valuable insights from historical sales data. This comprehensive comparative analysis dives into three prominent methods: Triple Exponential Smoothing, SARIMA (Seasonal Autoregressive Integrated Moving Average), and LightGBM (Light Gradient Boosting Machine). We delve deeper into their strengths, considerations, and suitability for diverse types of supply chain sales data. Triple Exponential Smoothing, popularly known as Holt-Winters, offers a sophisticated approach to capturing temporal patterns inherent in supply chain sales data. By integrating trends, seasonality, and smoothing factors, this method excels in accommodating data with recurrent patterns, making it well-suited for seasonal sales. Triple Exponential Smoothing has various strengths. Triple Exponential Smoothing effectively captures seasonal trends inherent in supply chain sales, enabling accurate prediction of recurring cycles. For instance, in the retail sector, where sales surge during festive seasons, this method can capture the periodicity and predict the surge accurately. Noise Reduction: The method inherently smooths out noise and fluctuations in data, minimizing the impact of outliers. This results in enhanced forecasting accuracy by providing a clearer picture of underlying trends. Triple Exponential Smoothing is relatively straightforward to comprehend and implement. This accessibility makes it a viable choice for analysts with foundational knowledge of time series data analysis. This method has many benefits, but there are also some considerations. While proficient at capturing trends and seasonality, Triple Exponential Smoothing may struggle when confronted with complex relationships or abrupt changes in data patterns. Unlike more advanced machine learning models, Triple Exponential Smoothing lacks the adaptability to dynamically adjust to evolving data patterns. Triple Exponential Smoothing finds its utility in scenarios where historical sales data displays regular cycles, making it valuable for forecasting demand in consistent product categories such as food and beverage staples.

SARIMA (Seasonal ARIMA): SARIMA, an extension of the ARIMA model, emerges as a powerful tool for supply chain sales data characterized by both long-term trends and seasonal fluctuations. It addresses a broader spectrum of data characteristics. It has many strengths. SARIMA adeptly tackles data with both seasonality and long-term trends, making it suitable for industries where products experience cyclical patterns while undergoing gradual shifts. For example, in the automobile sector, where sales trends align with both seasonal variations and gradual industry changes, SARIMA can accurately predict future sales. SARIMA provides a significant level of flexibility, enabling practitioners to refine the model by modifying the autoregressive (AR) and moving average (MA) components, and integrating differencing orders. Apart from these above-mentioned strengths, there are some considerations as well. Estimating parameters for SARIMA can be intricate and may require specialized expertise. This model demands a higher level of understanding and skill compared to simpler methods. SARIMA assumes that data is stationary or can be transformed into a stationary form through differencing. This assumption may not hold for real-world supply chain sales data, necessitating careful preprocessing. SARIMA is a robust choice for industries where supply chain sales data exhibit both long-term trends and seasonal variations.

LightGBM (Light Gradient Boosting Machine): LightGBM, a machine learning algorithm, represents a paradigm shift in time series analysis. It is well-suited for complex data relationships and nonlinear patterns, offering strengths that extend beyond traditional methods. LightGBM's ability to capture intricate data patterns, irrespective of linearity, provides a competitive edge over traditional methods. This is particularly useful in industries with products influenced by multiple interrelated factors. LightGBM provides insights into the importance of various features, aiding in the identification of key drivers influencing supply chain sales. For instance, in electronics, where sales are influenced by numerous product attributes, LightGBM can reveal the critical factors driving demand. Machine learning models, including LightGBM, demonstrate the capability to adapt quickly to changing data patterns. This adaptability suits volatile sales scenarios where sudden shifts in demand are common. Apart from these strengths, there are some considerations as well. LightGBM's efficacy is maximized

with a substantial amount of high-quality data. Insufficient or inaccurate data may result in suboptimal outcomes, emphasizing the importance of data quality. While LightGBM excels in capturing intricate patterns, overfitting is a concern. Prudent model tuning, cross-validation, and proper feature engineering are essential to mitigate this risk. As per our experimentation with the datasets and the Mean Absolute error calculation, LightGBM is seen to be particularly well-suited for supply chain sales data characterized by intricate relationships and dynamic patterns. It finds its application in sectors where sales are influenced by a multitude of interrelated factors, making linear models inadequate. The choice between Triple Exponential Smoothing, SARIMA, and LightGBM depends on various factors, including the nature of supply chain sales data, the expertise of the analytical team, and the requirement for adaptability. Triple Exponential Smoothing suits cyclic patterns; SARIMA excels with both trends and seasonality, while LightGBM shines in intricate and dynamic scenarios. A tailored approach, leveraging the strengths of each method, is recommended to suit our supply chain's unique sales dynamics and optimize forecasting accuracy.

5 Conclusions

In the contemporary data-driven landscape, marked by the convergence of cutting-edge technologies, this research delves into an enthralling fusion of innovations. We look at how new technologies like IOTA Tangle and Directed Acyclic Graph (DAG) work together with the Message Queuing Telemetry Transport (MQTT) protocol and advanced data analysis methods. The consequence is a profound shift in decentralized data sharing, real-time communication, and demand forecasting. IOTA Tangle and DAG emerge as robust, secure, and scalable platforms, enabling seamless cross-industry collaboration while preserving data integrity and authenticity in peer-to-peer data exchange. At the same time, using the MQTT protocol allows for quick real-time decisions in working together across supply chains because it has lower delays and better security. Additionally, our detailed study shows that Deep Matrix factorization can greatly improve how different industries predict demand in collaborative supply chains. This research ushers in an era of unparalleled data analytics efficiency, offering invaluable insights and agility across diverse industries, where data-driven decision-making attains unprecedented potency.

Apart from the various advantages of these models, there are a few limitations that may pose a hindrance to proper data analysis. For example, a very common problem that can arise is the cold start problem when we deal with new products or customers with no historical records. This can be primarily avoided by clustering the products into a similar category. The paper's datasets were simulated, but in real time we may get much larger-scale and sparse data, which can lead to impractical decisions. Furthermore, many other external factors like region, topography, and weather may also affect the datasets, which were not considered in this paper. Some other problems may also be visible due to overfitting and the assumption of linearity, which might not hold true in the case of complex supply chain systems. In the cutting-edge realm of demand prediction and supply chain analytics, intricate algorithms and computational paradigms may converge to create advanced ecosystems.

Conflict of Interest

The author declares that there is no conflict of interest in this work.

Data Availability

Data may be available on request.

References

Abdullah, S., Arshad, J., Khan, M. M., Alazab, M., & Salah, K. (2023). PRISED tangle: A privacy-aware framework for smart healthcare data sharing using IOTA tangle. *Complex and Intelligent Systems*, 9(3), 3023–3041. https://doi.org/10.1007/s40747-021-00610-8.

Anthony, B. (2023). Decentralized brokered enabled ecosystem for data marketplace in smart cities towards a data sharing economy. *Environment Systems Decisions*, *43*(3), 453–471. https://doi.org/10.1007/s10669-023-09907-0.

Bag, S., Golder, R., Sarkar, S., & Maity, S. (2023). Sene: A novel manifold learning approach for distracted driving analysis with spatio-temporal and driver praxeological features. *Engineering Applications of Artificial Intelligence*, 123(Part C), 106332. https://doi.org/10.1016/j.engappai.2023.106332.

Bhandary, M., Parmar, M., & Ambawade, D. (2020). A blockchain solution based on directed acyclic graph for iot data security. In *2020 International Conference on Computing, Communication, and Electronics (ICCCE)*. https://doi.org/10.1109/icces48766.2020.9137858.

Bhattacharyya, S., Sarkar, S., Sarkar, B., & Manatkar, R. (2023). Risk modeling framework for strategic and operational intervention to enhance the effectiveness of a closed-loop supply chain. *IEEE Transactions on Engineering Management*, 71, 7015–7028. https://doi.org/10.1109/TEM.2023.3261323.

Gangwani, P., Perez-Pons, A., Joshi, S., Upadhyay, H., & Lagos, L. (2023). Integration of data science and iot with blockchain for industry 4.0. In *Studies in big data* (pp. 139–177). https://doi.org/10.1007/978-981-19-8730-4_6.

He, P., Yan, T., Huang, C., Vallarano, N., & Tessone, C. J. (2023). Welcome to the tangle: An empirical analysis of the iota ledger in the chrysalis stage. In *Lecture notes in networks and systems* (pp. 419–431). https://doi.org/10.1007/978-3-031-45155-3_40.

Jnr, B. A., Sylva, W., Watat, J. K., & Misra, S. (2024). A framework for standardization of distributed ledger technologies for interoperable data integration and alignment in sustainable smart cities. *Journal of the Knowledge Economy*, *15*(3), 12053–12096. https://doi.org/10.1007/s13132-023-01554-9.

Ko, S., Lee, K., Cho, H., Hwang, Y., & Jang, H. (2023). Asynchronous federated learning with directed acyclic graph-based blockchain in edge computing: Overview, design, and challenges. *Expert Systems With Applications*, *223*, 119896. https://doi.org/10.1016/j.eswa.2023.119896.

Kraner, B., Vallarano, N., & Tessone, C. J. (2023). Topology and the tangle: How the underlying network topology influences the confirmation of blocks in iota. In *Lecture notes in networks and systems* (pp. 449–458). https://doi.org/10.1007/978-3-031-45155-3_43.

Kumar, N., Reiffers-Masson, A., Amigo, I., & Rincón, S. R. (2023). The effect of network delays on distributed ledgers based on directed acyclic graphs: A mathematical model. *Performance Evaluation*, *163*, 102392. https://doi.org/10.1016/j.peva.2023.102392.

Kumar, R., Arjunaditya, N., Singh, D., Srinivasan, K., & Hu, Y. (2022). AI-powered blockchain technology for public health: A contemporary review, open challenges, and future research directions. *Healthcare*, *11*(1), 81. https://doi.org/10.3390/healthcare11010081.

Kumari, A. & Sharma, I. (2023). Augmentation of internet of things security and privacy with iota tangle network. In *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)* (pp. 1–5).: IEEE. https://ieeexplore.ieee.org/abstract/document/10353394.

Maity, S., Rastogi, A., Djeddi, C., Sarkar, S., & Maiti, J. (2022). A novel optimized method for feature selection using non-linear kernel-free twin quadratic surface support vector machine. In *Pattern Recognition and Artificial Intelligence*, volume 1543 (pp. 339–353). Springer Nature. https://doi.org/10.1007/978-3-031-04112-9_26.

Mangrulkar, R. S. & Chavan, P. V. (2024). Beyond blockchain. In *Apress eBooks* (pp. 229–248). https://doi.org/10.1007/978-1-4842-9975-3_7.

Mazzocca, C., Romandini, N., Montanari, R., & Bellavista, P. (2023). Enabling federated learning at the edge through the iota tangle. *Future Generation Computer Systems*, *152*, 17–29. https://doi.org/10.1016/j.future.2023.10.014.

Paramesha, M., Rane, N., & Rane, J. (2024). Artificial intelligence, machine learning, deep learning, and blockchain in financial and banking services: a comprehensive review. *SSRN Electronic Journal*, *1*(2), 51–67. https://doi.org/10.2139/ssrn.4855893.

Pramanik, A., Sarkar, S., & Maiti, J. (2021). A real-time video surveillance system for traffic pre-events detection. *Accident Analysis & Prevention*, 154(5), 106019. https://doi.org/10.1016/j.aap.2021.106019.

Sadi, A. A., Mazzocca, C., Melis, A., Montanari, R., Prandini, M., & Romandini, N. (2023). P-IOTA: a cloud-based geographically distributed threat alert system that leverages P4 and IOTA. *Sensors*, *23*(6), 2955. https://doi.org/10.3390/s23062955.

Sarkar, S., Ejaz, N., Promod, C., & Maiti, J. (2020a). Pattern extraction using proactive and reactive data: A case study of contractors' safety in a steel plant. In *Proceedings of ICETIT 2019: Emerging Trends in Information Technology* (pp. 731–742).: Springer International Publishing. https://doi.org/10.1007/978-3-030-30577-2_65.

Sarkar, S., Pramanik, A., Khatedi, N., Balu, A., & Maiti, J. (2020b). GSEL: A genetic stacking-based ensemble learning approach for incident classification. In *Proceedings of ICETIT 2019: Emerging Trends in Information Technology* (pp. 719–730).: Springer International Publishing. https://doi.org/10.1007/978-3-030-30577-2_64.

Sarkar, S., Sammangi, V., Raj, R., Maiti, J., & Mitra, P. (2019). Application of optimized machine learning techniques for prediction of occupational accidents. *Computers & Operations Research*, 106, 210–224. https://doi.org/10.1016/j.cor.2018.02.021.

Tandu, C., Kosuri, M., Sarkar, S., & Maiti, J. (2022). A two-fold multi-objective multi-verse optimization-based time series forecasting. In *Proceedings of the 7th International Conference on Mathematics and Computing* (pp. 743–754).: Springer. https://doi.org/10.1007/978-981-16-6890-6_55.

Zhang, X., Zhu, X., & Ali, I. (2023). Performance analysis of the iota tangle and a new consensus algorithm for smart grids. *IEEE Internet of Things Journal*, *11*(4), 6396–6411. https://doi.org/10.1109/jiot.2023.3311103.