https://doi.org/10.1007/s13198-024-02519-0

ORIGINAL ARTICLE

Data-driven approach for rational allocation of inventory in a FMCG supply chain

Devesh Kumar[1] · Gunjan Soni[1] · Bharti Ramtiyal[2] · Lokesh Vijayvargy[3]

Received: 26 April 2024 / Revised: 15 August 2024 / Accepted: 11 September 2024

Abstract The aim of the article is to discuss the major issues concerning forecasting of sales and inventory distribution in traditional grocery retail stores, with a focus on a large supermarket company in Ecuador that operates with more than 200000 SKUs. It aims at the deployment of machine learning algorithms for efficient inventory management so that the business does not experience high stock or low-stock situations. The proposed approach includes the assessment of several supervised machine learning techniques such as Decision Tree, Random Forest, Linear Regression, and XGBoost techniques based on different performance measures that will help to select the best selling forecasting model. These findings underscore the fact that, with high demand uncertainty, heightened market demand rates and supply risks, shifting customer preferences, and ever-reducing product lifecycles, accurate demand forecasting can significantly lower supply chain costs. The study also establishes a need to maintain optimal inventory stock and the distribution of inventory across a number of warehouses. The research implication of the presented study indicates that the machine learning approach advocated for in the research would offer numerous benefits in the management of supply chain for retailers and enhance competitive advantage in the retail industry. To the best of the author's knowledge, this study is novel in its use of sophisticated machine learning approaches to solve problems specific to the grocery retail industry context while also offering a real-world solution to the issues covered.

Keywords Data-driven · Supply chain · FMCG · Machine learning · Inventory · Manufacturing

1 Introduction

In today's competitive world, industries more and more use machine learning (ML) to boost their operations. The supply chain sector can gain a lot from these new developments. The Fast-Moving Consumer Goods (FMCG) industry known for its fierce competition and slim profits, can see big improvements by using ML to optimize its supply chain (Kumari et al. 2023; Hu et al. 2019; Zhang et al. 2018). People involved in the business, like retailers, customers, and green activists, want more accurate forecasting models. These models can cut down on waste, make sure products are available, and help sustainable practices (Ramos and Oliveira 2023; Brewis and Strønen 2021). To give you an idea, in 2016 Swedish grocery stores threw away 30,000 tons of goods because of wrong forecasts. This shows how much we need better prediction models.

Several latter analyses reveal that this factor of AI has a positive influence on the supply chain management. It does this through improved tools of demand forecasting that affect the sales estimate and stock management (Li and Jiang 2020; Jadhav and Rokade 2020). However, we are still to learn more about utilizing live data as well as strong prognostic tools in an effort to address particular challenges in the FMCG industry. These include reduction of waste and

✉ Gunjan Soni
gsoni.mech@mnit.ac.in

Bharti Ramtiyal
bharti.mnit2022@gmail.com; bhartiramtiyal.mgt@geu.ac.in

Lokesh Vijayvargy
lokesh.vijayvargy@jaipuria.ac.in

[1] Malaviya National Institute of Technology Jaipur, Jaipur, India

[2] Graphic Era (Deemed to be university), Dehradun, India

[3] Jaipuria Institute of Management Jaipur, Jaipur, India

Published online: 19 September 2024

increasing sustainability as highlighted by Chen et al. 2019, Liu et al. 2021, and Verma and Sharma 2021. The question arises as to how these technologies could be systematically incorporated into supply chain processes to produce more efficiency and sustainability.

This research attempts to address the gap with a focus on how demand can be forecasted using machine learning, inventory distribution system can be enhanced, and the warehouses can be categorized according to how they are located. With the help of the data-based method, the study will be able to provide solutions to enable FMCG companies make better decisions that will enhance their supply chain productivity and duration (Chen et al. 2020; Patel et al. 2021). The research will also consider approaches to implementing these plans for the benefit of businesses and customers.

To address these gaps, this study employs a mixed research approach. They use supervised machine learning for demand forecasting, demand clustering algorithms for warehouses and cost benefit analysis for inventories. The findings will fall into three main areas: such as demand forecasting, warehouse clustering and inventory allocation. These results will give complete perspective of how use of machine learning will benefit the FMCG supply chain (Wang et al. 2018).

This study's results have a big impact on both theory and real-world use. In theory, it adds to the growing research on using machine learning in supply chain management (Chatterjee and Kumar 2020). In practice, it gives FMCG companies useful ideas to make their operations better, cut down on waste, and be more eco-friendly. By showing the real benefits of using data, the study highlights how important it is to bring new tech into supply chain plans.

The objective of study can be divided into three major categories. They are:

RO1: Forecasting demand for FMCG goods using supervised machine learning and ranking cities based on this demand.

RO2: Clustering the warehouse in each city based on distance using machine learning algorithm.

RO3: Inventory allocation of items into different warehouses and their cost–benefit analysis.

The paper is structured as follows: Sect. 2 reviews the existing literature on machine learning and data-driven approaches in supply chain management. Section 3 details the research methodology used in the study, followed by a discussion of the findings in Sect. 4. Finally, Sect. 5 concludes with a summary of the research, its limitations, and suggestions for future research directions.

2 Literature review

The data-driven approaches in supply chain management rely on machine learning-based models and other data analysis techniques to extract insights from large data sets. These approaches can help companies to better understand their supply chain processes, identify areas for improvement, and make data-driven decisions to optimize their operations. By leveraging the power of machine learning and data analysis, businesses can improve the efficiency, transparency, and resilience of their supply chains.

2.1 Machine learning-based models

There are two types of ML approaches that are commonly used, they are: This will be followed by an analysis of the physics-based approach and the data driven approach. Physical-based models are models that are developed by analyzing and applying the fundamental laws of physics of the actual system under consideration (Torrecilla et al. 2004). These models are generally applied in engineering and science solftware computation domains like meteorology, hydrodynamics and structures among others. The key strength of physical-based models is that they can accurately depict the system and predict how it will behave in various circumstances. However, these models may often be slow and time-consuming to develop, and they may often need a large amount of data and computational power to perform. Empirical models, on the other hand, are models that are developed based on the presented patterns and relationships that exist between different entities within a particular system. These models have number of applications in fields like machine learning, data mining, and predictive analysis and they are basically used in making prediction about future event with the help of past data. The main strength of data-driven models is that it can work with a large number of training instances, and can also capture non-linearities and interactions between variables. However, these models may fail to generalize good performance with new data due to the quality and the completeness of the training data (Karniadakis 2018; Brunton and Nathan 2018; Guo et al. 2019). The application, advantages, disadvantage, and limitations of data-driven models in scientific computing were reviewed by Zhang and Wang (2016). Ni et al. (2019) compared the performance of physics-based and data-driven models for predictive maintenance and discussed each approach's advantages and disadvantages.

This study uses a varied approach to tackle the identified gaps. It applies supervised machine learning to forecast demand clustering algorithms to optimize warehouses, and cost–benefit analysis to allocate inventory. The findings will fall into three main areas: demand forecasting, warehouse clustering, and inventory allocation. These results will give a full picture of how machine learning can boost the FMCG supply chain (Wang et al. 2018).

As demonstrated in this study, practical implications are significant when it comes to opening up new perspectives and adding to the overall theoretical foundation. From a theoretical perspective, it contributes to the existing literature on the application of machine learning in supply chain management (Chatterjee and Kumar 2020). In practice, it provides helpful recommendations to FMCG companies so that they can improve their activities, reduce costs, and be more sustainable. Explaining the actual positive outcomes of data application, the study describes why it is crucial to introduce new tech into supply chain concepts.

A set of techs out there that capture data and employ intelligent techniques, including Radio Frequency Identification or RFID for short. That stuff is getting really really important to run the show when it comes to what you have in stock while you are in a position of handling a significant number of goods coming and going. Although, it is a good choice and it is not expensive compared to any of the commercial inventory systems available in the market. Furthermore, it is easy to install and it does make organizing your things as a part of your daily routine a piece of cake (Benke, Hedi, and Cirikovic 2023).

Sensors go hand-in-hand with IoT to make the inventory management process automatic altogether. IoT setups enhance how accurate the information you receive for inventory enables you track and monitor things in realtime. This makes managing the supply chain way better (Saillaja et al. 2023; Madhwal and Panfilov 2017). Also, you see Artificial Intelligence (AI) and machine learning used a lot in situations that involve searching the optimal solution. These technologies assist in predicting how much stock must be available, identifying abnormal occurrences, and making informed decisions on when new products should be ordered (Neghab et al. 2022; Yang et al. 2024; Zhang and Tan 2023).

2.2 Data-driven approaches in supply chain management

The scene in the industry shifts quickly, and the use of analytical approaches plays a crucial role. These methods enhance manufacturing by increasing its efficiency in terms of being high-quality and less costly. The main advantage of using data in manufacturing is that it assists in the optimization process. Data also has an influence on the performance and reliability of providers further. Metrics such as time taken to deliver services, quality control ratings, and cost efficiencies can be monitored using analytics. Part of analyzing data is to show trends, which in turn assists to predict when the machines need fixing and even identify issues. Real time information sharing with suppliers enhances collaboration and faster resolutions in case of an error (Zekhnini et al. 2023). Moreover, the execution of such activities through digital media allows everyone involved to see each other and gain trust, which makes suppliers more responsible and credit-worthy over time (Cavalcante et al. 2019).

These analyses demonstrate that data-based approaches can enhance various aspects of production including what to produce, when to produce and how to maintain production equipment. For instance, in the study by Lee et al. (2020), an efficient plan that was based on data was employed to improve semiconductor manufacturing. It involved using machine learning to predict likely consumer demand and then schedule production accordingly. The latter pushed the workers to further reduce their overtime by 22 percent and increase the output by 23 percent. Another advantage of data-driven methods is quality assurance Quality control is implemented through data analysis, which ensures that all the products meet the required standards. For instance, Liao et al. (2021) applied this case to identify defects in PCBs as they are produced to reduce their production time. They employed image processing and machine learning to detect problems at once achieving first pass yield of 99 percent of catching defects. Data-driven methods also help to predict when machines will break down in factories (Mycroft et al. 2020; Xu et al. 2020). This means using data and analysis to guess when equipment might fail, so you can fix it before it breaks (Wang et al. 2018). For instance, Riaz et al. (2020) used a data-driven method to predict when a machine in a textile factory would stop working. They put sensors on the machine and used machine learning to look at the data. This cut maintenance costs by 73%.

The fast-moving consumer goods (FMCG) supply chain has seen big changes because of data-driven supply chain management. Big data analytics, machine learning, and AI now let companies use data in new ways to make their supply chains better, cut costs, and keep customers happy. Companies now use data-driven methods for demand forecasting in FMCG supply chain management. By looking at sales data social media trends, and other key numbers, they can better guess future product demand. This helps them to improve production cut waste, and have the right products where they need to be when they need to be there. Studies show that good demand forecasting can make supply chains work better and earn more money (Kim and Ko 2012; Yeung et al. 2015). Data-driven methods also help with inventory management in FMCG supply chains. By watching inventory levels in real-time and studying sales patterns, companies can keep the right amount of stock and avoid running out or having too much. This saves money on storage, cuts waste, and makes sure they can always meet customer needs. Several papers point out that the usage of big data to manage inventory increases the effectiveness of the supply chain (Chen and Yang 2012; Wang et al. 2019). FMCG companies also employed business analytical tools to enhance their overall supply chain. They can improve the situation as they get data about suppliers, production, shipping, etc.

and can figure out how to achieve all this amending more and spending less. This might mean employing sophisticated instruments such as network optimization and simulation modeling. Business literature proves that the use of data in supply chain management results in better performance of the supply chain and enhanced customer satisfaction (Huang et al. 2016; Wu et al. 2019).

Recent years have seen some data-driven techniques in fast-moving consumer goods or indeed popularly known as FMCG. A literature review of this study demonstrates how these approaches can enhance the operations of supply chains in the FMCG industry. Chatterjee and Kumar (2020) used data analytics to analyze the performance of FMCG corporations in supply chain management noting that big data analytics increases effectiveness and reduces expenses. From the research conducted by Wang et al. (2018), the authors developed a data-driven model for the supply chain management for the FMCG sector using predictive analytics and optimization algorithms to enhance on inventory management and minimize on stock-out situations. In their study, Hu et al. (2019) propose ways to improve the FMCG supply chain, such as predicting demand, setting appropriate inventory stock, and planning the right transportation routes. Liu, et al. (2021) proposed a big data approach toward handling risks in FMCG value chains, where machine learning algorithms could identify potential risks and address them. In the paper examining the position of big data analytics in FMCG Supply chain management Zhang et al. (2018) focused on the prospects of improving real time data analytics and predictive modeling to enhance the efficiency of Supply Chain processes. Collectively, these studies indicate that the application of data in FMCG supply chain management can yield improved decision-making, reduced costs and more sustainable practices. By using real-time data predictive analytics, and optimization algorithms, FMCG companies can understand their operations better and make smart choices to improve their supply chain operations.

2.3 Warehouse location strategies

Hierarchical clustering helps find the best warehouse spot keeping costs low and supply quality high (Skerlic and Muha 2016). Methods like VIKOR have an impact on picking the top warehouse policy by looking at things such as distance, area, and costs (Sarıcan et al. 2022). Using decision models is key to choose the right warehouse place for many markets. This boosts profits when demand changes and helps set up good inventory plans (Lin and Wang 2018). Also, a multi-criteria decision model is needed to pick the best spot for eco-friendly warehouses. It considers both number-based and quality-based factors (Jacyna-Gołda and Izdebski 2017). The Fuzzy AHP method works well too. It's used to select the ideal warehouse location by checking different criteria (Caron and Oshan 2023). What's more combining warehouse location choices with how things run, like shipping and stock control, leads to a more united and effective way to manage supply chains (Agrawal and Goyal 2016).

Data-driven methods and machine learning models show great promise to boost supply chain management in the Fast-Moving Consumer Goods (FMCG) industry. Current research looks at predicting demand managing stock, and improving supply chains overall. Yet, not much work explores how to fit inventory allocation plans into these data-driven models. Also, we don't know enough about problems with data quality, consistency, and the need for data experts in FMCG supply chains. This gap opens the door to create more complete models. These new models could predict demand, streamline supply chain steps, and assign inventory. They would also take into account data quality and the know-how needed to put them to use.

3 Research methodology

The method shown in Fig. 1 has an impact on how a certain product group—homecare items—is managed in the supply chain of an Ecuadorian supermarket chain. This approach breaks down into several main steps, each meant to tackle a key part of supply chain management, from predicting demand to distributing inventory.

1. Choosing the Product Family: Homecare

The study starts by picking a specific product family, in this case, homecare items. The choice to zero in on one product family stems from the need to handle the big amount of data available. By limiting the scope, the analysis can be more in-depth and offer more precise insights about the chosen category.

2. Predicting Demand

The next step is to forecast demand for the chosen homecare products. This involves combining monthly sales figures from different stores in the supermarket chain. Complex computer programs and data tools help predict future demand for these items. Getting demand forecasts right is key because it shapes production plans and inventory control. This ensures stores have the correct amount of products when needed, which cuts down on both empty shelves and overstocked items.

3. Ranking of Cities

After predicting demand, the study ranks cities based on how much people want homecare items. This ranking gives useful insights into areas where demand is highest helping to target inventory distribution and marketing plans. Cities that want more items might need restocking

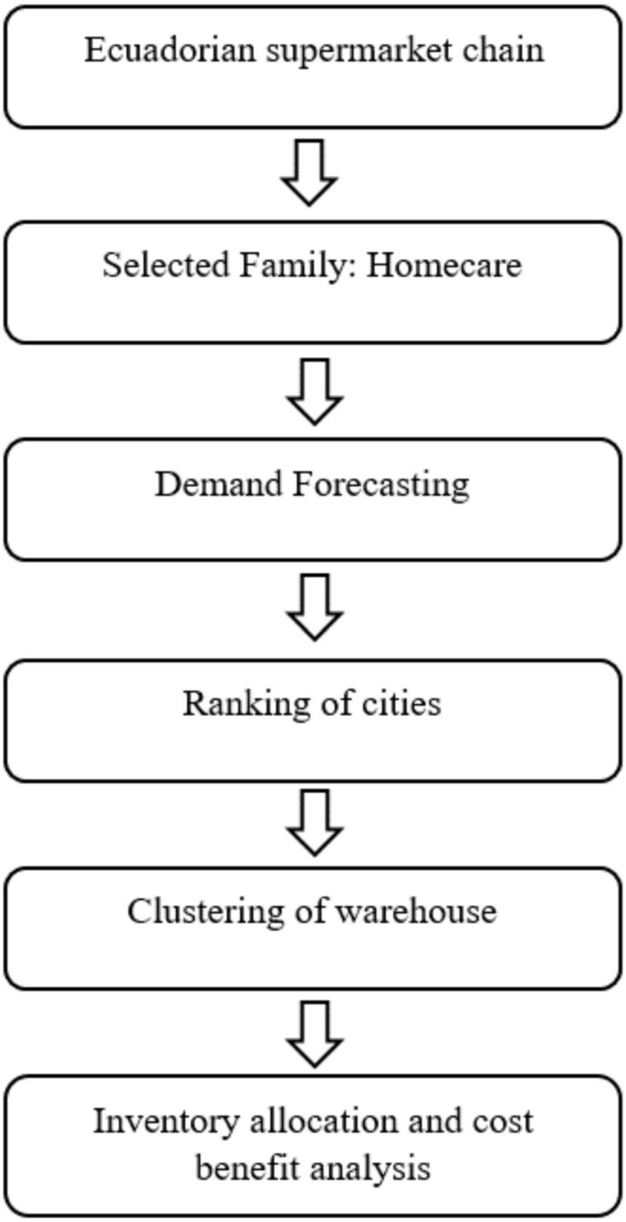

Ecuadorian supermarket chain
Selected Family: Homecare
Demand Forecasting
Ranking of cities
Clustering of warehouse
Inventory allocation and cost benefit analysis

Fig. 1 Methodology

more often or bigger inventory stocks, while those that want less could get deliveries less.

4. Clustering of Warehouses

The next step groups warehouses based on how close they are to different cities. By putting closer warehouses together, the study tries to make the distribution network better. This grouping allows for more productive transport routes cutting delivery times and shipping costs. Good grouping also makes sure products are sent from the nearest possible warehouse cutting delays and making the whole supply chain more responsive.

5. Inventory Allocation and Cost–Benefit Analysis

The research zeroes in on inventory allocation and performs a cost–benefit analysis. The knowledge gained from the earlier steps—predicting demand ranking cities, and grouping warehouses—helps to figure out the best inventory levels for each warehouse. The cost–benefit analysis checks the money impact of different ways to manage inventory. This makes sure the chosen method boosts profits while cutting down on waste. This breakdown aids the company to make smart choices about where to put resources weighing the costs of keeping stock against the risks of running out and missing sales.

3.1 Demand forecasting

The justification of future sales allows a company to make sure it has enough stocks to fulfil the demands of the next months without buying those extra items that will remain unsold yet increase the total cost of your inventory. In this work, we considered hundreds of items offered at various stores in Ecuador, evidencing the scope and difficulty of such a type of forecasting. It is clear that the quality of demand forecasts depends on the quality of data, and the authors noted that their training dataset was comprised of data about the items, the stores, and the unit sales. The information is also useful in pattern recognition which can be used to make predictions concerning future sales. This data has been analyzed using statistical and machine learning methodologies for arriving at their forecast based on the complexity of the data and the precision level required. In summary, this research sought to predict the unit sales of the products being sold in different stores in the Ecuadorian supply chain. We employed training dataset which had details such as date, items, store details, and unit sales. Using this data, thewe could come up with a forecast showing the future unit sales of these items.

3.1.1 Model selection

Several machine learning models were trained on a given dataset, and their performances were evaluated using a metric called R^2 (Jebaraj et al. 2011). R^2 is a statistical measure that represents the proportion of variance in the target variable (the variable you're trying to predict) that can be explained by the independent variables (the variables used to make the prediction). In other words, R^2 measures how well the model fits the data. The models that were trained include random forest, decision tree, and linear regression. These are all supervised learning algorithms that can be used for regression tasks (predicting a continuous value). The specific

Table 1 Model Evaluation

Description and model (Monthly data)	R^2_score_train	Cross-validation	R^2_score_test
Random forest with n_estimators = 200	0.978	[0.83 0.84 0.85 0.82 0.83]	0.836
Random forest (max_depth = 100)	0.96	[0.75, 0.73, 0.76]	0.81
Xgboost	0.25	[0.24, 0.10, 0.15]	0.74
Decision tree (max_depth = 200)	0.96	[0.688, 0.733, 0.722, 0.742 0.721]	0.71
Decision Tree	1.00	[0.68, 0.65, 0.68]	0.70
Linear regression	0.42	[0.42, 0.27, 0.33]	0.2

hyperparameters of the models (such as the number of trees in the random forest, or the depth of the decision tree) may have been tuned to find the best performance. Based on the results presented in Table 1, the random forest model with a specific number of trees (n_estimators) was found to have the highest R^2 value. This means that this model had the best fit to the data, among the models that were evaluated. As a result, it was selected as the best model for this task.

Table 2 Metrics of the model selected

Metrics	Value
R^2 score of training data	0.97977
R^2 score of validation data	0.8367
K fold cross validation (cv = 5)	[0.8381118 0.8459283 0.85524235 0.82349034 0.84296472]
Mean Absolute Error	28.13
Mean Squared Error	3790.65
Root Mean Squared Error	61.56

3.1.2 Forecasting

How to interpret a random forest model and indicate R-Squared as one of the factors. It also had to be tested on the cross-validation test—a part of the set that the model did not use in its training. Based on the obtained model, the R-squared value was calculated to be approximately equal to 0. As for the validation set, the achieved value was 836, which translates to making explanation to the extent of only 83%. This reflects six (6)% of the variance of the target variable. This fairly high R-squared value signifies a fact as to the proposition that the model given fits perfectly the model or the data that has been incorporated. To avoid the overfitting of the model and to improve the generalization, k-fold cross-validation was also employed. Cross-validation is a technique that is used for assessing the competency of the model and for the detection of the bias and over-fitting of data. K-fold cross validation can therefore be described as the set of division of a given dataset into k benign equal partitions or else folds and or the model is trained k times where in each of the ith round the ith fold is used for the purpose of validation while the remaining portions of the partitions are used in training the model. It also enables one to regulate the stability of the model as well as identify problems such as overfitting or bias. The averages and the standard errors of the estimates of evaluation measures are computed and make up the outcome of k-fold cross-validation. The values cited in Table 2 may be the mean and standard deviations of the evaluation measures that are estimated from k-fold cross-validation. These metrics can be any model performance measure one is likely to encounter soon such as R square, Mean Square Error or Root Mean Square Error. Speaking of the metrics described in details in Table 2, we believe the following conclusions can be drawn: the model proposed is the random forest model and it is good and does not over-fit the data.

Table 3 Feature importance

Features	Importance weightage in model
Item number	0.47898594
Store number	0.12379494
type	0.11274224
cluster	0.0354351
year	0.04175777
month	0.04952731
Avg oil price	0.10411901
Population of city	0.03194363
Area of city	0.02169404

The contribution or relative significance of each variable in determining the unit sales was determined by using a feature from the scikit-learn library. This feature most probably calculates the feature weights of a trained model, which is an indication of the degree of contribution of each predictor variable. The feature importance of each variable was then analysed to determine the conclusions presented in Table 3 below. In light of these findings, it can be noted that the most significant variable affecting the unit sale was "item numbers", whereas the variable with the least impact was "area of the city". It is important because this information can be helpful in several situations. For instance, it will assist in determining which variable is more suitable for use in the unit sale prediction rather than the other variable that may not be quite useful in the model. It also can help understand other factors that influence unit sales and may have recommendations for improving sales, for example using more attention to certain items or focusing on certain regions.

3.2 Ranking of cities

This paragraph explains how the cities may be sorted based on the monthly unit sales which are expected in the coming months for a given set of products. It has adopted a ranking system in terms of upper and lower caps relative to the box plot. A box plot is a type of graph that is used to illustrate the distribution of some given data set(Li et al. 2022). It comprises of a rectangle and lines referred to as whiskers, the rectangle represents the interquartile range of the data which is the extent of data within first and third quartiles that is between 25 and 75th percentiles of the data. From the box, the whiskers run up to the minimum and maximum values that are within 1. By definition, the lower and upper fences can be determined as follows: lower fence = Q1–1.5 * IQR and upper fence = Q3 + 1.5 * IQR, so 5 times the IQR from the box. In this case, the lower cap for each item means the first quartile demand value of 25% among all the data. Upper cap is, therefore, determined using the third quartile demand value of 75% which is the third twenty five % of the data. These caps are used to decide if the unit sales forecasted for a city for a particular item is higher, average or lower than others. Cities with outright unit sales which are less than the lower cap are categorized as low volume cities and cities with unit sales in between the lower and upper cap are classified as moderate volume cities while cities with outright unit sales greater than the upper cap are classified as high volume cities. This method enables the researcher to achieve a relative degree of consistency while ranking the cities under consideration by criteria such as the forecasted unit sales for different items while using a statistical measure of the distribution of the data to arrive at the caps illustrated in Table 4.

Table 4 Data frame of demand lower cap and upper cap of each item

S No	Lower Cap	Upper Cap
1	182.5	442
2	18	65.50
3	16	59
4	16.25	57
5	37	101.75
6	55	173.50
7	27.75	89.5
8	44.25	112
9	90.75	192.75
10	8.5	23
11	8.5	23.5
12	8.25	40

The process of classifying cities into high, medium, or low based on their predicted unit sales for a given item and the upper and lower caps calculated in the previous stage. For each city and item, the predicted unit sales are compared to the upper and lower caps calculated using the box plot hypothesis. If the predicted unit sales fall below the lower cap, the city is classified as low. If the predicted unit sales fall between the lower and upper caps, the city is classified as medium. If the predicted unit sales exceed the upper cap, the city is classified as high. This allows for a standardized approach to classifying cities based on their predicted unit sales for different items, using the upper and lower caps calculated from the box plot hypothesis. This approach provides a way to compare and rank cities based on their forecasted unit sales, which can be useful for decision-making and resource allocation in various industries.

3.3 Clustering warehouses

3.3.1 Locating warehouses

The process of determining the location of warehouses in Ecuador, given that the data provided did not include information about their location. To make assumptions about the location of warehouses, we have decided to assume that there is one warehouse in each city, located at its city center. To determine the latitude and longitude of each city center, an online source was used, such as a mapping tool or a geographic database. Using this information, we were able to pinpoint the location of each warehouse on a map. Figure 2 shows the resulting map, with each warehouse location represented by a pinpoint on the map. By visualizing the location of warehouses across the country, it becomes easier to identify patterns and trends in the data, such as which areas have more or less access to warehouses, or which areas may require additional warehousing infrastructure.

Overall, this process of determining the location of warehouses using online sources and geographic databases provides a useful way to make assumptions about warehouse locations when data is incomplete or unavailable. By understanding the location of warehouses, businesses and organizations can make more informed decisions about logistics, supply chain management, and resource allocation.

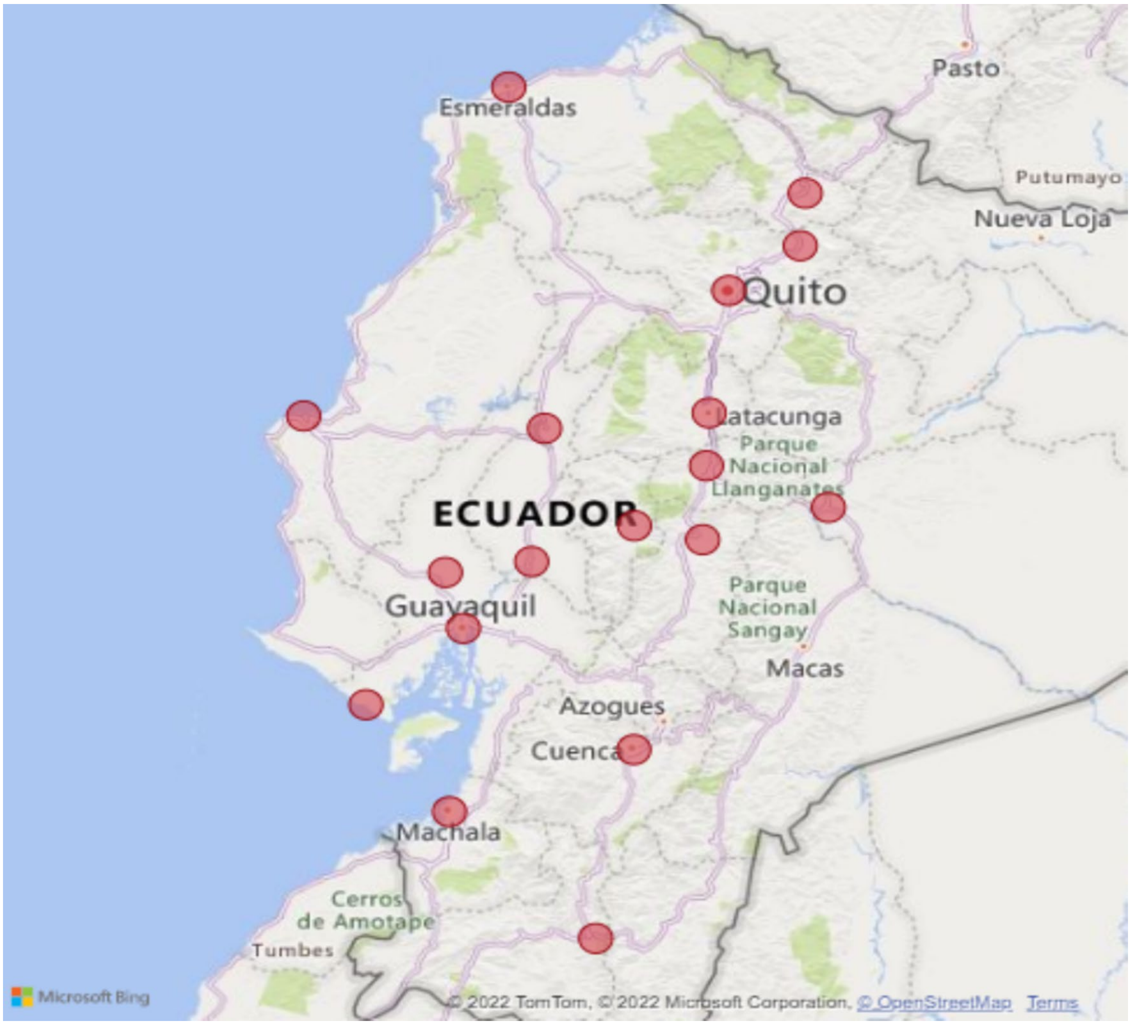


Fig. 2 cities in Ecuador where the warehouses are located

3.3.2 Clustering warehouses based on distance

The process of transforming the latitude and longitude coordinates of the warehouses into x and y coordinates, which was necessary in order to calculate the distance between the warehouses. Latitude and longitude coordinates are a commonly used way to specify locations on the earth's surface, but they are not immediately suitable for calculations involving distance or direction. In order to perform such calculations, it is necessary to convert the latitude and longitude coordinates into a different coordinate system, such as a Cartesian coordinate system with x and y coordinates. To perform this conversion, the team used an online converter tool that is specifically designed to convert latitude and longitude coordinates into x and y coordinates. This tool takes as input the latitude and longitude coordinates for each warehouse location and returns the corresponding x and y coordinates. Once the latitude and longitude coordinates were transformed into x and y coordinates, the team was able to use these values to calculate the distance between the warehouses using a variety of mathematical formulas and techniques. Overall, this process of transforming the latitude and longitude coordinates into x and y coordinates using an online converter tool is a common approach used in geospatial analysis and allows for more complex calculations involving distance and direction to be performed.

These were the formulae used for the conversion:

$$x = R \times coscos(lat)xcoscos(long)y = R \times coscos(lat)xsinsin(long) \tag{1}$$

In these formulae, R represents the approximate radius of the earth, which is typically taken to be 6,371 km. The latitude and longitude values are expressed in radians, which is a common way of measuring angles in mathematics.

K-means clustering was chosen due to several advantages it offers, including its ease of implementation, ability to scale to large datasets, and ability to generalize to clusters of various shapes and sizes. However, K-means clustering is not well-suited for spatial clustering, as it assumes that the distance between points is Euclidean, which is not accurate for geographical coordinates. To overcome this limitation, the team transformed the geographical coordinates of the warehouses into a two-dimensional Cartesian coordinate system using the formulae mentioned in the previous paragraph. This changed coordinate system allows the use of K-means clustering to cluster the warehouses. By using the transformed coordinates, K-means clustering can be applied to cluster the warehouses based on their proximity in the x–y plane. This approach allows for the benefits of K-means clustering to be utilized while accounting for the spatial nature of the data.

After transforming the geographical coordinates into a Cartesian coordinate system, K-means clustering was applied to group the warehouses into clusters based on their proximity in the x–y plane. The elbow method was used to determine the optimal number of clusters, which was found to be four. Figure 3 shows the resulting clusters, with each dot representing a warehouse and the color indicating the cluster membership. The different colors of the dots represent the different clusters of warehouses that were identified. This approach allows for the identification of geographically close warehouses, which could help optimize the logistics and supply chain management for the company. For example, the company can now more easily plan their inventory and distribution routes by considering the locations of the warehouses in each cluster.

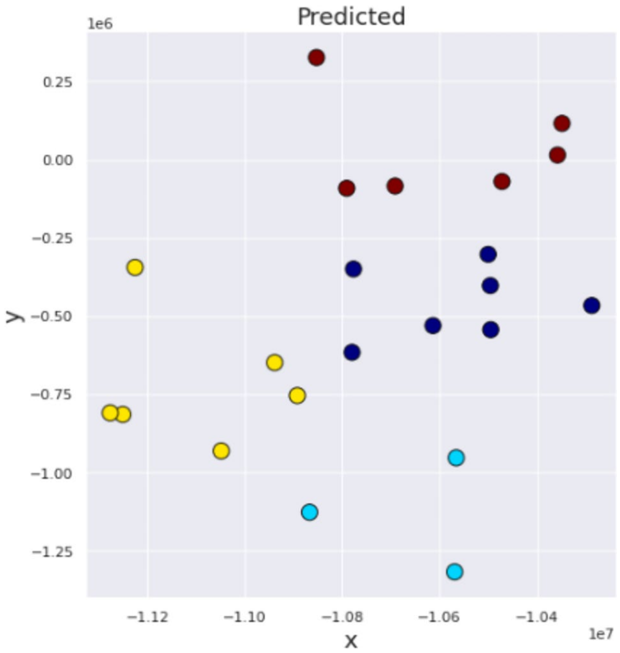


Fig. 3 Output for clustering n = 4

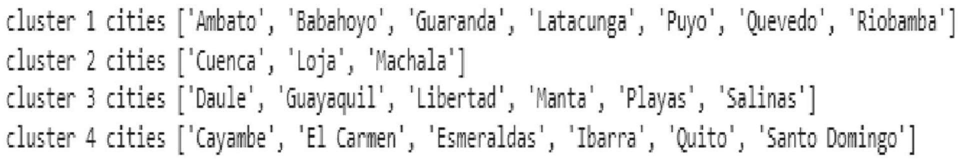
Finally, the cities belonging to each cluster were found which is depicted in Fig. 4 below:

3.4 Inventory allocation

After clustering the warehouses into four groups, we used the forecasting model to predict the demand for each item in each city. We then summed up the demand for each item in the cities belonging to each cluster to get the total demand for each item in each cluster. This provides a cluster-wise forecast for the demand of each item in the cities. By looking at the cluster-wise demand forecast, the company can make informed decisions on how to allocate inventory and resources to each cluster. This information can also help the company identify any supply chain or logistics issues that may arise due to differences in demand between clusters.

For the allocation of inventory, we went with two strategies. In both strategies, the goal is to allocate inventory to optimise demand fulfillment while minimizing costs associated with shipping and handling. In strategy 1, inventory is allocated to all warehouses regardless of their level of demand. This means that the low demand warehouses will receive inventory even though their forecasted demand may not require it, which can result in additional costs. In strategy 2, inventory is allocated to warehouses with high and medium demand only. The forecasted demand of low demand is fulfilled by sending it to the nearest medium or high demand warehouse. This approach reduces unnecessary inventory allocation to low demand warehouses, thereby lowering costs. Table 5 show the inventory allocation of item 1,456,881 of price 572 for cluster 1 based on strategy 1 and strategy 2, respectively. Strategy 1 shows that all warehouses, including low demand ones, have been allocated inventory. In contrast, for strategy 2, only high and medium demand warehouses have been allocated inventory, while low demand warehouses have not. Instead, their forecasted demand has been sent to the nearest medium or high demand warehouse. This results in a more efficient allocation of inventory, as inventory is sent only where it is

Fig. 4 Cities in each cluster



```
cluster 1 cities ['Ambato', 'Babahoyo', 'Guaranda', 'Latacunga', 'Puyo', 'Quevedo', 'Riobamba']
cluster 2 cities ['Cuenca', 'Loja', 'Machala']
cluster 3 cities ['Daule', 'Guayaquil', 'Libertad', 'Manta', 'Playas', 'Salinas']
cluster 4 cities ['Cayambe', 'El Carmen', 'Esmeraldas', 'Ibarra', 'Quito', 'Santo Domingo']
```

Table 5 Inventory allocation

Strategy 1		Strategy 2	
City	Inventory	City	Inventory
Ambato	115	Babahoyo	297
Babahoyo	297	Latacunga	429
Guaranda	180	Puyo	214
Latacunga	314	Quevedo	190
Puyo	214	Riobamba	434
Quevedo	190		
Riobamba	254		

needed, reducing unnecessary costs associated with shipping and handling.

To decide on which strategy to go with in each cluster, we use cost–benefit analysis.

3.5 Cost benefit analysis

To decide on which strategy to go with in each cluster, we use cost–benefit analysis. In a cost–benefit analysis, we need to weigh the costs and benefits of each strategy to make an informed decision. In our case, we consider the ordering cost and the holding cost as two major costs that affect the decision of which strategy to choose. Ordering costs are the expenses incurred in placing and collecting a fresh shipment of inventory. These costs include transportation costs, communication costs, inspection costs, transit insurance costs and accounting costs. Ordering costs vary depending on the frequency of ordering and the amount of inventory ordered. Holding costs, on the other hand, are the costs of retaining unsold inventory. These costs include warehousing costs, labor costs, insurance, damaged or spoilt inventory, and opportunity costs. Holding costs increase with the amount of inventory held, and the duration of holding. In the cost–benefit analysis, we need to compare the total cost of each strategy, taking into account the ordering and holding costs, against the potential benefits, such as revenue generated from sales. Based on the analysis, we can choose the strategy that minimizes the total cost, while still meeting the forecasted demand.

$$Inventory\ Holding\ Cost = \begin{pmatrix} Storage\ Costs\ +\ Labor\ Costs \\ +\ Opportunity\ Costs \end{pmatrix} / Total\ Value\ of\ Annual\ Inventory \quad (2)$$

Cost calculation: The cost structure of the FMCG and highlights the major constituents of supply chain costs. Different expenses contribute in supply chain costs just as transportation expenditures, which include inbound, outbound, secondary, and tertiary costs, account for around 6.7% of total costs in the FMCG supply chain. Inbound costs refer to the transportation of raw materials or goods from suppliers to the manufacturing site, while outbound costs refer to the transportation of finished goods to retailers or customers. Secondary and tertiary costs refer to additional transportation expenses that may be incurred during the distribution process. The storage and warehousing expenses represent approximately 3.86% of gross sales in the FMCG supply chain. This cost component is important because it is closely related to transportation costs. When goods are transported to different locations, they need to be stored and warehoused until they are distributed to retailers or customers. Table 6 represents the major constituents of supply chain costs in the FMCG and their total share of gross sales. This table likely provides a breakdown of the different cost components discussed in the paragraph, as well as any additional factors that may contribute to the cost structure of the industry.

As we have considered two strategies for inventory allocation, we will calculate cost for both strategies and choose the best one for each item.

- We have assumed that one supplier is allocated to all warehouses present in the same cluster. Hence, the order-

Table 6 Supply Chain Costs (as a Percentage of Gross Sales)

Supply chain cost type	Cost in FMCG		
	Average	Lower bound	Upper bound
Cost of material	52.92	15	90
Cost of labor	8.90	0.51	70
Cost of production overhead	11.78	0.5	40
Storage cost	3.52	0.16	12
Inbound transport cots	3.38	0.12	20
Outbound transport cost	3.38	0.12	20
Warehousing cost	2.06	0.1	8
Secondary/ Tertiary Transportation cost	2.02	0.2	10
Distributor's margin	6.35	0.1	20

ing cost for each warehouse of the cluster will be the same.

- Holding costs will vary for each warehouse.
- Assumption:

- Ordering cost factor = Transportation cost + Labor cost + Handling and packaging cost
- Holding cost factor = Storage cost + Warehousing cost + Labor cost.

- Calculated these costs by randomly generating costs using data from table.
- Costs of both strategies:

- Strategy 1 cost:

$$\left(hcf_{high} * i_{high} * p\right) + \left(hcf_{medium} * i_{medium} * p\right) + \left(hcf_{low} * i_{low} * p\right) + \left(ocf_{constant} * i_{high,medium,low} * p\right) \tag{3}$$

- Strategy 2 cost:

$$\left(hcf_{high} * i_{high} * p\right) + \left(hcf_{medium} * i_{medium} * p\right) + \left(hcf_{medium,high} * i_{low} * p\right) + \left(ocf_{constant} * i_{high,medium,low} * p\right) \tag{4}$$

where,

- hcf = holding cost factor
- i = inventory
- p = price
- ocf = ordering cost factor
- If strategy 2 cost < strategy 1 cost, we will prefer strategy 2 and vice versa.

Output for cluster 1 for item-1456881:

The following Table 7 first shows the city warehouse and inventory allocation for strategy 1, second it shows the city warehouse and inventory allocation for strategy 2. In the end, it shows which strategy to use, based on the cost–benefit analysis we wanted to perform. In the above case, the system is recommending strategy 1 as it is better than strategy 2 by 3,392,484-unit cost.

Table 7 Cluster 1 output of inventory allocation for item 1,456,881

For cluster 1	
Item: 1,456,881 Price: 572	
Strategy 1	
City	Inventory
Ambato	115
Babahoyo	297
Guaranda	180
Latacunga	314
Puyo	214
Quevedo	190
Riobamba	254
Strategy 2	
City	Inventory
Babahoyo	297
Latacunga	429
Puyo	214
Quevedo	190
Riobamba	434

Inventory allocation is a crucial aspect of supply chain management. It involves deciding how much inventory to allocate to each warehouse or distribution center in the supply chain network to ensure that customer demand can be met efficiently and cost-effectively. A cost–benefit analysis is a useful tool to evaluate the potential costs and benefits of different inventory allocation strategies. For inventory allocation, we clustered the warehouses based on their geographic location and market characteristics. This allows for an optimized supply chain in each cluster, which can lead to reduced costs and improved efficiency. By clustering warehouses, it becomes possible to implement different strategies that are tailored to the specific needs of each cluster.

4 Discussions

It presented two significant developments. First, we obtained the capability of propheying item demand at each store. Our team tested several models and fine-tuned parameters to find the best fit: a Random Forest regressor that achieved a test score of exactly 0. undefined This model predicts the amount of items you will require in various stores. Second, I understood how these demand predictions can be utilised to allocate inventory to the warehouses. There is however limited research in inventory control and more so optimizing inventory using the forecasted demand, hence our focus was to harness our forecast in determining where inventory should be allocated. To support our idea, we conducted a cost benefit analysis. It assisted us in determining the optimal stock positioning and ideal costs of each product. For this purpose, we summed up the supply chain expenses involved in merchandising stock in a dissimilar manner in the warehouses and established several inventory placement solutions. We then retrained our model repeatedly for checking bias in the model. Accordingly, when it comes to different generated costs, the model offered different approaches. This increased our confidence that the model can recommend better positions of inventory to withstand forces afloat. To summarise, the paper provides a framework for inventory control and management based on data analysis. To achieve this, we employed the utilization of a machine learning algorithm whereby we had to identify the optimal approach for distributing inventories through several warehouses using a cost–benefit analysis.

4.1 Managerial implications

As such, our work proposes a new point of view on how the inventory degree can be controlled and optimized using data. The two major advancements in forecasting item demand and balancing inventory distribution across the various warehouses are highly likely to drive substantial cost benefits for organizations. Through it, the managers can make reasonable decisions about which inventory to stock, and where to stock it in order to eliminate unnecessary costs of overstocking and stock-outs. This way, through the application of the best approach to demand forecasting, the businesses can be able to determine the amount of stock to hold in each store and manage the stocks most efficiently. This is advantageous as it eliminates high costs that may come with holding overstock inventories and will help in keeping customers from being out of reach due to lack of stock. The approach described in this work will be useful for a cost–benefit decision-making framework in inventory plans. Through it, managers can compare various inventory distribution policies by considering the costs and benefits of inventory, thus being able to determine the optimum method of distribution.This can also save some cash costs that relate to distribution of inventories like the transport costs, besides enhancing the performance of the supply chain. In conclusion, the need for quantitative methodologies in the computation of inventory levels and related goals is evident in this work. Applying decision trees and cost effectiveness analysis, companies bound to reap the benefits of intricate and effective supply chain system. Decision makers can use these findings to establish effective and efficient inventory control strategies in order to make appropriate decisions that will improve inventory distribution among various warehouses.

5 Conclusion

In this section, the authors conclude the main contributions of the paper which falls under three categories that include the use of method for forecast unit sales, division and classification of the warehouses in different clusters for inventory distribution, as well as use of SCCM in the identification of the proper inventory distribution strategies. All these contributions stem from the use of machine learning models and cost–benefit analysis of ideal supply chain workflows. The first of these contributions is a technique for identifying the appropriate model that should be employed for forecasting the unit sales. Finally, we employed several models and hyperparameter optimization to determine the efficient model for their tasks based on evaluation through Random Forest regressor with test score accuracy of 0. undefined They were able to use this model to forecast the number of its units that would be sold in various items in many stores. The second contribution highlighted in the paragraph is the ability to divide warehouses into various groups for inventory distribution purposes. As earlier mentioned, when warehouses are grouped based on certain criteria such as location, customer/client requites or supply chain costs, then within a group, inventory distribution and supplies costs can be best managed. Lastly, we utilized supply chain cost before allocating items to the certain warehouse based on the fact that cost of total supply chain have to be minimized. This required identifying an inventory cost minimization model to compare the cost of inventory distribution and management of each item with respect to holding cost, transportation cost and demand. In this way, the researchers were able to make more effective decisions on inventory stock placement that could reduce the costs within the supply chain.

Future studies could broaden the current analysis to look at more product groups beyond the homecare family, like bread/bakery, grocery, cleaning, and beauty items. Scientists might create models specific to each group examining how to tailor cost-cutting methods to different kinds of products. Also, it would help to do a full cost–benefit review across all product families considering different demand levels (high medium low) to find the best distribution plan for each item. This could lead to more detailed insights on inventory placement, price control, and focused marketing within a company's supply chain. By looking at various product groups future research could further improve ways to boost supply chain efficiency and cost-effectiveness.

Funding Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

Ethics approval None of the authors performed research involving human participants or animals for this publication.

Informed consent All authors have given their consent for the submission of this work to the journal.

References

Agrawal R, Goyal A (2016) Optimizing warehouse location, using differential evolution, in order to reduce the overall freight cost. In supply chain management: applications for manufacturing and service industries

Benke I, Hedi I, Cirikovic E (2023) RFID inventory management system sampling optimization based on Zebra Android framework. In Proceedings of the 46th ICT and electronics convention, MIPRO https://doi.org/10.1109/MIPRO.2023.9491176

Brewis C, Strønen F (2021) Digital transformation in FMCG and automotive industries–emergence of digital innovation capabilities. In Proceedings of the european conference on knowledge management, ECKM (pp. 104–111)

Brunton SL, Nathan Kutz J (2018) Data-driven versus physics-based modeling. Annu Rev Fluid Mech 50:645–668

Caron RJ, Oshan T (2023) Optimal warehouse location and size in practice. Int J Operational Res. https://doi.org/10.1504/IJOR.2023.132819

Cavalcante IM, Frazzon EM, Forcellini FA, Ivanov D (2019) A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing. Int J Inf Manage 49:86–97

Chatterjee K, Kumar P (2020) Data analytics in fast moving consumer goods (FMCG) industry supply chain management: a review. Int J Bus Analytics Intell 8(1):40–57

Chen FF, Yang YH (2012) A dynamic inventory model for deteriorating items with stochastic demand and exponential partial backlogging. Appl Math Model 36(5):2017–2029. https://doi.org/10.1016/j.apm.2011.08.017

Chen Y, Li J, Zhang Z (2019) A novel data-driven approach for demand forecasting in FMCG supply chain. IEEE Access 7:152678–152687

Chen L, Feng Y, Wang Y (2020) Big data driven sustainable development of FMCG industry. J Clean Prod 273:123015

Guo J, Liu X, Liu J (2019) Physics-based and data-driven modeling: a comparative study. Mathematics 7(3):607–623

Hu X, Cheng S, Li Y (2019) Optimization of FMCG supply chain by using data-driven methods. J Intell Manuf 30(1):81–92

Huang X, Li C, Li X, Li Y (2016) An effective algorithm for supply chain network optimization under uncertain demand and transportation cost. Math Probl Eng 2016:1–16. https://doi.org/10.1155/2016/7276497

Jacyna-Gołda I, Izdebski M (2017) The multi-criteria decision support in choosing the efficient location of warehouses in the logistic network. Procedia Eng. https://doi.org/10.1016/j.proeng.2017.04.424

Jadhav V, Rokade MM (2020) Optimization of fast moving consumer goods (FMCG) supply chain using machine learning approach. Int J Supply Chain Manag 9(3):221–230

Jebaraj S, Iniyan S, Goic R (2011) Forecasting of coal consumption using an artificial neural network and comparison with various forecasting techniques. Energy Sour, Part a: Recovery, Utilization, Environ Eff 33(14):1305–1316

Karniadakis G (2018) A brief overview of physics-informed machine learning. J Comput Phys 375:1339–1356

Kim K, Ko B (2012) Forecasting the sales of new products using a hybrid method. Expert Syst Appl 39(4):4454–4460. https://doi.org/10.1016/j.eswa.2011.09.074

Kumari CS, Deepu BSVV, Dheeraj G, Surampudi VKS (2023) FMCG market analysis for wholesalers and retailers using machine learning. In Proceedings of the 2023 3rd international conference on pervasive computing and social networking (ICPCSN 2023).

Lee J, Choi Y, Kim H (2020) Data-driven approach for production planning optimization in semiconductor manufacturing. Comput Ind Eng 141:106328. https://doi.org/10.1016/j.cie.2019.106328

Li C, Xu A, Tian Z, Li C (2022) Data-driven anomaly detection and early warning issues. In: 2022 2nd Asia-pacific conference on communications technology and computer science (ACCTCS) (pp. 423–430). IEEE

Li Y, Jiang Y (2020) Data-driven supply chain management in the era of big data. J Bus Res 117:454–461

Liao W, Su Y, Huang C, Yu K (2021) Data-driven quality control for printed circuit board manufacturing process via real-time defect detection. IEEE Access 9:40192–40204. https://doi.org/10.1109/ACCESS.2021.3065750

Lin Y-S, Wang K-J (2018) A two-stage stochastic optimization model for warehouse configuration and inventory policy of deteriorating items. Computers Ind Eng. https://doi.org/10.1016/j.cie.2018.04.008

Liu X, Li Y, Li X (2021) A big data-driven method for FMCG supply chain risk management. Int J Prod Res 59(7):2136–2148

Madhwal Y, Panfilov PB (2017) Blockchain and supply chain management: Aircrafts 'parts' business case. In Annals of DAAAM and proceedings of the international DAAAM symposium. https://doi.org/10.2507/daaam.scibook.2017.01

Mycroft W, Katzman M, Tammas-Williams S, Hernandez-Nava E, Panoutsos G, Todd I, Kadirkamanathan V (2020) A data-driven approach for predicting printability in metal additive manufacturing processes. J Intell Manuf 31:1769–1781

Neghab DP, Khayyati S, Karaesmen F (2022) An integrated data-driven method using deep learning for a newsvendor problem with unobservable features. Euro J Oper Res 302(2):482–496

Ni DD, Liu J, Gao Y (2019) A comparative study of physics-based and data-driven models for predictive maintenance. Mech Syst Signal Process 119:538–550

Patel RK, Chavda KD, Patel P (2021) Machine learning based waste reduction in FMCG supply chain management. Int J Log Res Appl 24(4):356–369

Ramos P, Oliveira JM (2023) Robust sales forecasting using deep learning with static and dynamic covariates. Appl Syst Innov. https://doi.org/10.3390/asi6050085

Riaz MS, Javed MA, Nawaz MH, Ahmad B (2020) Predictive maintenance of textile machinery using machine learning techniques. SN Appl Sci 2(7):1–11. https://doi.org/10.1007/s42452-020-03427-5

Saillaja V, Menaka M, Kumaravel V, MacHap K (2023) Development of an IoT-based inventory management system for retail stores. In: proceedings of the international conference on sustainable computing and smart systems, ICSCSS https://doi.org/10.1109/ICSCSS2023.9491177

Sarıcan B, Baysal M. E, Sarucan A (2022) Determination of the best alternative position for the storage location/product assignment by using VIKOR. In lecture notes in networks and systems

Skerlic S, Muha R (2016) Identifying warehouse location using hierarchical clustering. Transport Problems

Torrecilla JS, Otero L, Sanz PD (2004) A neural network approach for thermal/pressure food processing. J Food Eng 62(1):89–95

Verma R, Sharma MK (2021) Demand forecasting in FMCG industry using machine learning techniques. Int J Innov Technol Explor Eng 10(5):104–110

Wang T, Li J, Tang L (2018) Data-driven fast moving consumer goods supply chain model and application. Int J Control Autom 11(5):125–138

Wang X, Song X, Liu J (2019) Big data analytics in supply chain management: a comprehensive review and future research directions. Comput Ind Eng 128:851–863. https://doi.org/10.1016/j.cie.2018.10.042

Weng T, Liu W, Xiao J (2020) Supply chain sales forecasting based on lightGBM and LSTM combination model. Ind Manag Data Syst 120(2):265–279

Willemain TR, Smart CN, Schwarz HF (2004) A new approach to forecasting intermittent demand for service parts inventories. Int J Forecast 20(3):375–387

Wu DD, Olson DL, Zhao X (2019) Data analytics for supply chain sustainability: decision model and research directions. Int J Prod Res 57(12):3825–3839. https://doi.org/10.1080/00207543.2019.1579895

Xu Z, Dang Y, Munro P, Wang Y (2020) A data-driven approach for constructing the component-failure mode matrix for FMEA. J Intell Manuf 31:249–265

Yang B, Xu X, Gong Y, Rekik Y (2024) Data-driven optimization models for inventory and financing decisions in online retailing platforms. Ann Oper Res. https://doi.org/10.1007/s10479-024-04833-0

Yeung W, Yeung A, Cheng T (2015) A demand forecasting model for inventory control of fresh food in supermarkets. Int J Prod Econ 160:160–167. https://doi.org/10.1016/j.ijpe.2014.11.008

Zekhnini K, ChaouniBenabdellah A, Cherrafi A (2023) A multi-agent based big data analytics system for viable supplier selection. J Intell Manufacturing. https://doi.org/10.1007/s10845-023-02253-7

Zhang Q, Tan Y (2023) Data-driven e-commerce end-to-end inventory optimization algorithm. Front Artificial Intell Appl. https://doi.org/10.3233/FAIA230014

Zhang Y, Wang J (2016) Data-driven modeling and scientific computing. Appl Mech Rev 68(5):050801–051013

Zhang H, He Y, Xiao G (2018) Big data analytics for fast-moving consumer goods supply chain management: a review. J Ind Inf Integr 9:52–60

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.