Md. Rokibul Hasan[1]

Md. Zahidul Islam[2]

Mahfuz Alam[3]

Md. Sumsuzoha[4]

Predictive Analytics and Machine Learning Applications in the USA for Sustainable Supply Chain Operations and Carbon Footprint Reduction

Journal of Electrical Systems

Abstract: - With the escalating concerns worldwide regarding climate change and environmental sustainability, there is an increasing focus on emissions and ecological footprint reduction in supply chain operations in the USA. This study explored the application of predictive analytics and machine learning in the supply chain management domain for reducing carbon emissions and granting sustainable operations. For the present research paper, Walmart organization provided all the supply chain activity data used in this research study, it consisted of comprehensive data on their industrial activity levels, production outputs, energy consumption, types of fuels used, geographical data, and weather conditions. Three Machine learning algorithms were trained and tested, notably, Random Forest, XG-Boost, and the Bagging algorithm. Based on all the metrics, Random Forest was the best classifier because of its excellent generalization, high measure of precision and recall, and high AUC. As per the results, the random forest algorithm was the most accurate in its predictions of all the models evaluated. Implementing the random forest benefits businesses in America with high accuracy and robustness, flexibility, scalability, risk management, and Mitigation. As regards the US economy, deploying the Random Forest can benefit the government in the following ways: reducing carbon footprint, attracting foreign investment, and enhancing competitive advantage.

Keywords: Predictive Analytics; Machine Learning; Sustainable Supply Chain; Carbon Footprint reduction; Random Forest; XG-Boost; Bagging algorithm

Introduction

According to Abdollahi (2023), supply chain management is an instrumental process for the majority of organizations in America to efficiently deliver products and services to the targeted customers. Nevertheless, conventional supply chain operations are more or less resource-intensive and result in a considerable carbon footprint by way of carbon dioxide and other greenhouse gas emissions. Adewusi et al. (2024), indicate that growing concerns over climate change and sustainability have put most companies under tremendous pressure to reduce the carbon footprint in their supply chains. Predictive analytics and machine learning create new opportunities for changing the supply chain into a sustainable one, applying optimized decisions and forecasts of demand. This research paper explores the application of predictive analytics and machine learning in the supply chain management domain for reducing carbon emissions and granting sustainable operations.

Problem Statement

As per Bag et al. (2022), with the escalating concerns globally about climate change and environmental sustainability, there is an increasing focus on emissions and ecological footprint reduction in supply chain operations. Sustainability has emerged as the topmost priority of businesses in the USA across all sectors due to the rise in environmental awareness, government pressure for regulations, and high consumer expectations about eco-friendly operations. The field of supply chain that produces almost half of the carbon emissions in the world

[1] [1]MBA Business Analytics, Gannon University, Erie, PA, USA

[2]MBA Business Analytics, Gannon University, Erie, PA, USA

[3]MBA Business Analytics, International American University, Los Angeles, California, USA

[4]MSc Business Analytics, Trine University, USA

Primary Author: MD Rokibul Hasan, MBA Business Analytics, Gannon University, Erie, PA, USA.

Email: prorokibulhasanbi@gmail.com

provides great opportunities to use sophisticated analytical techniques to develop sustainable operations and thus lower the environmental footprint.

Brander (2022), contends that predictive analytics and machine learning (ML) has proven to be pivotal tools in this endeavor, facilitating data-driven decision-making and impactful strategies for optimizing supply chain processes. Heydar (2019), states that harnessing the power of enormous data generated within the supply chain from raw material procurement to final product delivery types of analytics can provide valuable insights and patterns that will inform sustainable practices. Therefore, this research paper pegs on the premise that predictive analytics and machine learning can help create sustainable supply chain operations and carbon reduction initiatives. Capitalizing on data-driven insight and predicting the appropriate models, supply chain management in charge can make very informed decisions that balance operational efficiency with efforts towards achieving environmental sustainability goals as part and parcel of the effort towards addressing the problem of climatic change.

Literature Review

Significance of Sustainable Supply Management

Gonuguntla (2018), indicates that supply chain activities play a paramount role in the global economy, but it also contributes substantially to greenhouse gas emissions and environmental degradation. As per the *World Economic Forum*, supply chains account for more than 80% of organizations' environmental effects and up to 60% of their overall carbon footprint. Due to ever-increasing consumer awareness and demand for sustainable products and services, there is now mounting pressure on organizations to address the environment and sustainable development implications of their supply chain activities.

Sustainable supply chain management incorporates environmental factors in the design, plan, and process operation of the supply chain. This includes logistics optimization, waste reduction, minimizing energy consumption, and circular economy (Nzeako et al. 2024). Sustainable practices not only reduce the negative impact on the environment but also are operational and financial boons for organizations. The boons include the cost-cutting, improved brand image, and enhanced survival chances against climate change-based disruptions.

The Role of Predictive Analytics in Supply Chain Management

Marwa (2023), argues that predictive analytics play a significant role in supporting sustainable supply chain management. Predictive analytics employ statistical models, algorithms, and data mining techniques in foretelling future events, trends, and outcomes based on historical data. These advanced technologies of today help organizations gather, analyze, and gain insight from large amounts of data that may, in turn, enable them to make informed decisions and optimize supply chain operations. Presented below are some of the ways predictive analytics can be applied to sustainable supply chain management:

A. Demand Forecasting

Predictive models help forecast demand for products and services by analyzing historical sales data, consumer behaviors, and market trends (Oliveira, 2021). This in turn aligns the companies' productions and logistics, helping to curtail waste and excess inventory.

B. Route Optimization

Equally important, predictive analytics are used to optimize transportation routes based on factors such as traffic, weather, and vehicle usage. It will reduce the consumption of fuel, and Ghg, and allow for efficient deliveries (Bag, 2022).

C. Inventory Management

Predictive models help organizations anticipate and manage inventory levels more effectively, minimizing the risk of stockouts or overstocking (Bag, 2022). This in turn reduces the environmental impact associated with excess storage, transportation, and disposal of unused products.

D. Equipment Maintenance

Predictive analytics can predict equipment failures or breakdowns even before they occur by analyzing sensor data and the history of maintenance. That allows reacting ahead of time, minimizing downtime, energy consumption, emergency repairs, or replacement.

E. Logistic Planning

Predictive analytics can assess millions of transportation incidences to find the low-carbon route and mode combination for each shipment. It considers factors like distance, vehicle type, load consolidation opportunities, carrier emissions profiles, etc. for scheduling the deliveries via the most efficient logistics channels. Machine learning optimizes route planning on a dynamic basis. This prevents unnecessary miles and emissions through unused fleet capacity estimates and dynamic rerouting. According to PwC in 2017, even logistics optimization can help reduce as much as 15% of the carbon footprint in the entire supply chain (Brander, 2022).

The Role of Machine Learning in Supply-chain Management

As per Khanai (2022), machine learning is a branch of artificial intelligence, consolidation algorithms, and statistical models so that the system performs the specified task without being explicitly programmed for that. As opposed to traditional programming where the rules are explicitly coded, ML algorithms learn from data and accuracy improves over time. Park (2021), states that machine learning in the context of sustainable supply chain management has a myriad of applications. Integrating ML algorithms in operative functions of sustainable supply chains in reducing carbon emissions has received reasonable attention in recent academia. This section will discuss some of the significant empirical works with an investigation into various applications of ML surrounding Supply Chain Management, including methodologies employed, findings made, and some sustainability implications/home.

In the recent past, a significant volume of studies has been conducted regarding how machine learning techniques support sustainable supply chain operations and carbon emission reduction. For instance, Tech (2022) compared some machine learning models regarding carbon emission prediction in a supply chain network. They then compared the performances of actual data using linear regression, decision trees, random forests, and neural networks. They found that ensemble methods like random forests performed the best in carbon emission predictions. Such prediction models, they noted, would help supply chain managers take proactive measures on possibilities of emission reduction.

Abdollahi [2023], investigated how ML algorithms can be used to predict demand in the electronics sector. Classical statistical methods were benchmarked against state-of-the-art ML techniques that include Random Forests, Gradient Boosting Machines, or extended short-term memory networks. The findings were that ML outperformed traditional methods both in terms of accuracy but also in terms of the robustness of their output, strongly so when considering LSTM networks. The outcome of improving the accuracy of demand forecast was to optimize the inventory level, which reduced waste and, thereby, overproduction, directly translating to carbon footprint reduction.

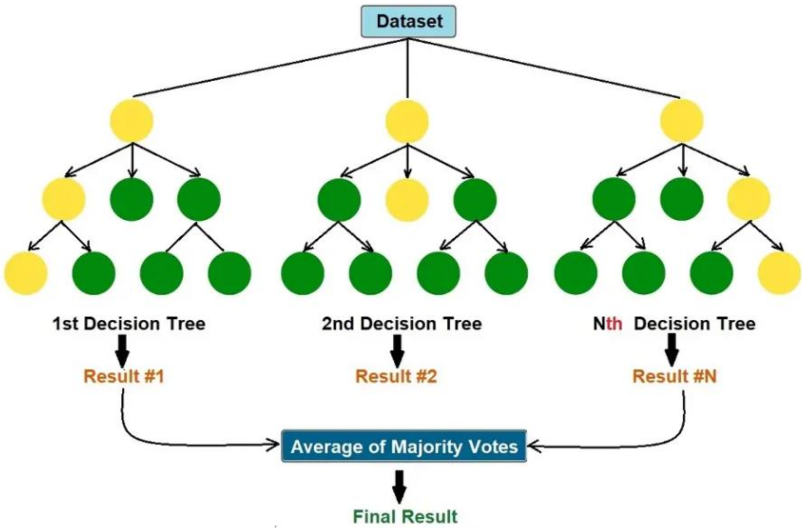
By contrast, Gonuguntla, V. (2019), examined the adoption of machine learning algorithms to support sustainable supplier selection and procurement decisions. They developed an integrated model of fuzzy logic, decision trees, and neural networks in evaluating the performance of suppliers' sustainability across environmental, social, and economic dimensions. The empirical analysis demonstrated that this form of a multi-model approach provided more robust, reliable assessments of suppliers' sustainability than single machine learning techniques.

On the other hand, Heydar (2021) explored machine learning techniques in terms of supply chain CO2 risk assessment in the procurement sector. More specifically, Heydar used genetic algorithms and integrated them with support vector machines-SVM and back-propagation neural networks. In this regard, testing the machine learning models, Sang's analysis found that a backpropagation neural network could devise more accurate CO2 risk classifications than a support vector machine.

Conversely, Khanai (2022), developed a hybrid demand prediction model within the context of a networked supply chain. The model proposed was examined against various time series data, which have depicted different volatility of fluctuation. The hybrid model gave reliable forecasts on the vast array of demand series studied. It also came to light that exponential smoothing with a covariate [ETSX], used initially considerably outperformed the ARIMAX and that the latter performed very poorly concerning the volatile demand pattern forecasting.

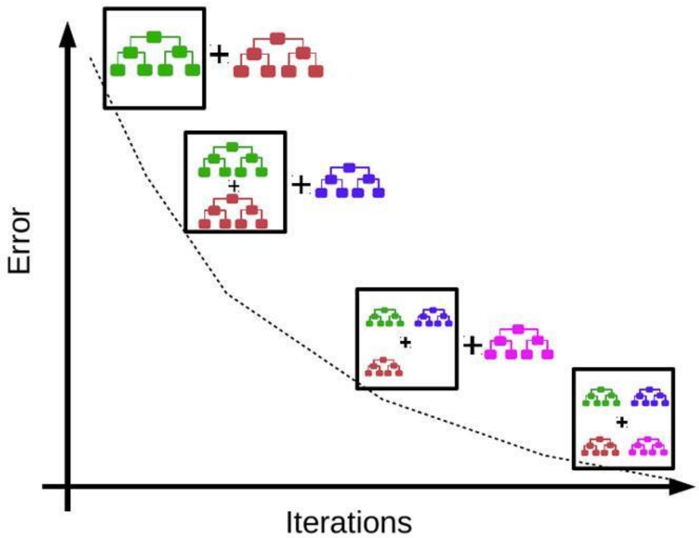
Types of Machine Learning Algorithms

Random Forest Algorithm



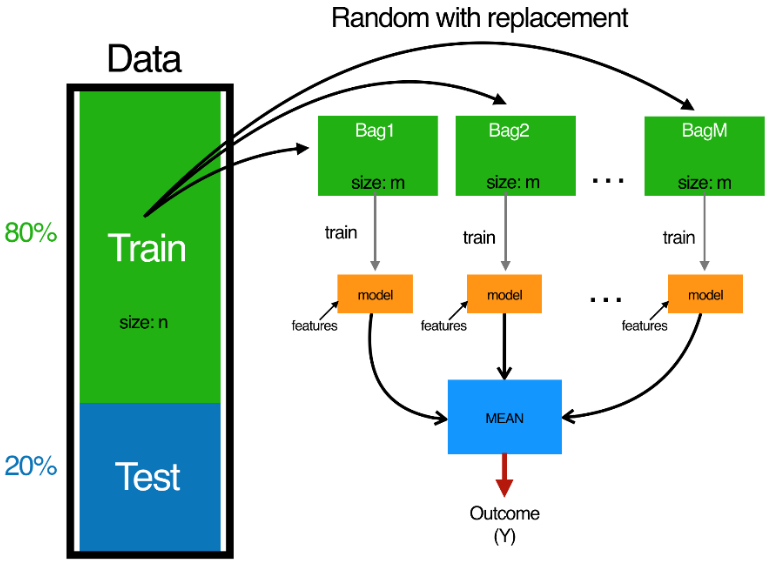
The first technique to be applied will be the random forest technique, which offers prediction based on the average of the predictions made by multiple decision trees trained on random samples drawn with replacement from the original training data, that is the ensemble method. Mandala (2024), stated that the RF algorithm consists of decision trees, where a selected input variable is in a top-down tree-like fashion, recurrently split into two or more categories based on metrics such as entropy or information gain until a stopping criterion is reached, such as no significant improvement of the splitting metrics or a maximum tree depth. Due to the reduced variance from averaging, the averaging of predictions from this "forest" of decision trees results in more accurate forecasts than those of a single decision tree.

XG-Boost Algorithm



The XG-Boost objective function is composed of two functions, most notably, regularization and training error. The algorithm keeps calculating the loss of each leaf node during the training process, chooses the leaf that maximally reduces the gain, and grows new decision trees adaptively and iteratively. Specifically, XG-Boost starts with the first tree, fitting the residual left over from the previous tree, and repeats this process such that each succeeding tree fitting is boosted by focusing on the mis-predictions of the previous stages, to yield a total of K trees upon training where the leaf nodes of each tree correspond to scores that represent sample characteristics (Marta, 2023). Then, to generate predictions for new data, these matched scores from each tree are averaged with the model, summing up the contributions of all weak learners to output the final prediction, as shown in the machine learning framework of XG-Boost.

Bagging Algorithm (Bootstrap Aggregating)



Oliveira (2021), asserts that the Bagging algorithm [bootstrap aggregating] was initially proposed by Breiman et al. (1994). The Bagging algorithm generates multiple bootstrapped samples from an original training dataset through sampling with replacement. Each sample is used to train a different classifier. In making predictions on new data, the Bagging framework sets up votes or averages the outputs of all the classifiers to combine into a final prediction. Figure 4 describes a process to generate multiple bootstrapped copies of a train set from the original and then train individual base classifiers. After that, all the models are combined to give their output of Bagging. This ensemble approach enhances the machine learning models in terms of stability and accuracy.

Methodology

In this study, the researcher conducted an empirical experiment using Python to compare the forecasting performance of different demand prediction algorithms using real-life supply chain data. This involved comprehensive data on their industrial activity levels, production outputs, energy consumption, types of fuels used, geographical data, and weather conditions. Later, a simulated experiment was done that very closely mimicked the real-world supply chain dynamics to compare the impact of changes on inventory management. All the predictive models were benchmarked against this traditional statistical forecasting baseline to determine whether any gains in accuracy from machine learning justified additional complexity from a computational perspective. Applied assessment of the demand forecasting algorithms under realistic conditions was enabled thereby in such a way as to explore their potential supply chain advantages.

Data Description

For the current study, Walmart organization provided all the supply chain data used in this research study. It consisted of comprehensive data on their industrial activity levels, production outputs, energy consumption, types of fuels used, geographical data, and weather conditions. logistics process, as well as historical data. April 1, 2019, to April 31, 2021, was the duration for which historical data could be extracted because that was the maximum period available within the systems. This was sourced from five of the primary Walmart warehouses spread across the United States that catered to the end customers and other middle-level warehouses in the supply chain. Specifically, 24 months of historical demand observations helped model the supply chain transportation activities for a substantial past period from the direct data that was sourced from Walmart's leading warehouses that fulfilled other downstream members in the chain of supply across the American market.

Data Pre-Processing

The dataset was cleaned by removing noise by performing various steps, such as handling outliers and imputing missing values to prepare it for analysis. Three subsets of the data were created using deterministic random sampling to avoid any bias validation set, a training set, and a test set in an 80:20 proportion for training and testing, respectively. Also, 10-fold cross-validation was applied in the case of in-sample training to estimate model performance. The majority of the features were nominal or categorical in the dataset, so they had to be transformed into some form of a number to be able to apply a machine learning algorithm to them. The preprocessing steps were thus strategically made to increase the quality and the form of data for the practical construction and evaluation of models.

Comparative Analysis Between Machine Learning Algorithms

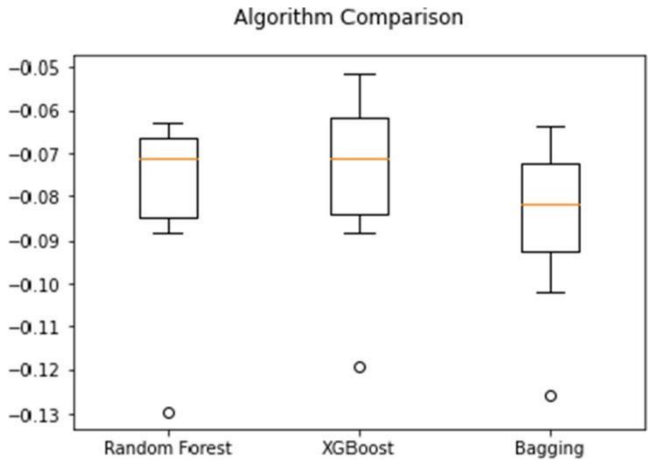


Fig 1: Displays the comparison of the algorithms' standard deviation

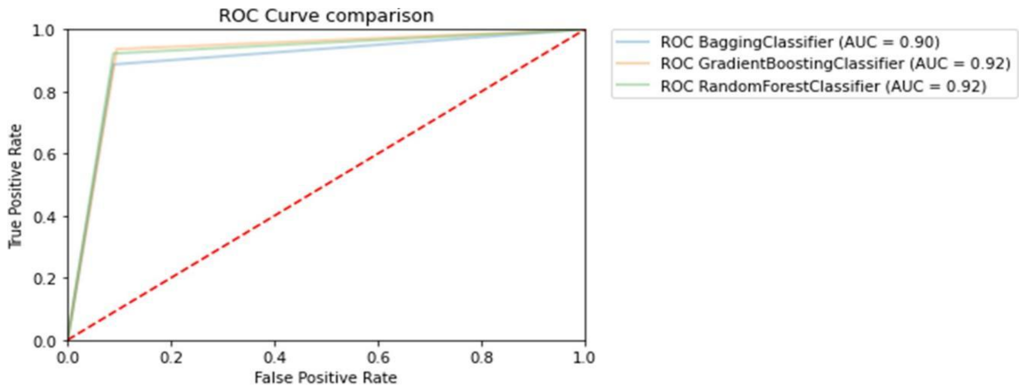


Fig 2: ***Showcases the ROC curve of the Machine Learning Algorithms***

From the above chart, all three classifiers performed very well in terms of predicting CO2, with their AUCs greater than 0.90. The Gradient Boosting Classifier and the Random Forest Classifier have the highest AUC, 0.92, which is slightly better than that of the Bagging Classifier at 0.90. In selecting any of these classifiers, the Gradient Boosting Classifier and the Random Forest Classifier may be favored because of their slight improvement in performance due to the higher AUC.

Table 1: Showcases the Comparison of the Machine Learning Algorithms

Rank	ML methods	Train accuracy	Test accuracy	Precision	Recall	AUC
1	**RF classifier**	1	0.9267	0.925667	0.929577	0.92814
2	**XG-Boost classifier**	0.9857	0.9167	0.921959	0.929577	0.917320
3	**Bagging classifier**	0.9914	0.8967	0.915551	0.908451	0.897263

Based on all the metrics, Random Forest is the best classifier because of its excellent generalization, high measure of precision and recall, and highest AUC. As per the results, random forest's algorithm was the most accurate in its predictions of all the models evaluated, leading to a precision as high as 92.5% above that of XG-Boosting with a precision of 92.1 and Bagging with 91.5%, respectively. A comparison of the three machine learning models against the performance metrics revealed that RF was the most accurate with the lowest error rate, considering 0.075 against 0.079 for Bagging and 0.085 for XGBoost. RF, therefore, presented the best predictive performance among the three ensemble learning methods for demand forecasting with higher precision and lower error as opposed to the other Bagging and XGBoost methods.

Machine Learning Algorithms	Execution time before tuning	Execution time after tuning
Random Forest Classifier	13.5 ms	11.8 ms
XG-Boost Classifier	253 ms	154 ms.

Table 2: Displays the Outcomes of Tuning the Algorithm

From the table above, the Random Forest algorithm demonstrates a modest improvement in execution time after tuning. It remains the faster alternative between the two classifiers, making it appropriate for applications requiring very quick predictions. The XG-Boost Classifier is substantially improved by tuning. However, it is much slower than the random forest approach. Although this is still slower than the Random Forest Classifier, the enormous improvement makes it more viable in situations where increased performance metrics, such as precision and recall, outweigh longer execution times.

How to Implement the Model

Step 1: Define Model Objectives and System Requirements: Businesses in America should first set goals to develop a model that can accurately assess and promote sustainability in supply chain activities based on various input features.

Step 2: Data Collection: Data gathering will require that organizations in the USA adequately collect data from suppliers, logs of transportation, production facilities, and Environmental data databases. The relevant features will be carbon footprint, energy consumption, waste management, water usage, supplier sustainability scores, mode of transport, and source of materials.

Step 3: Data Processing- Data preprocessing cleaning is strongly recommended to correct any erroneous or missing data points. Normalize features so that they are on approximately the same scale; this often helps improve model performance. Split data into training, validation, and testing sets. For example: 70% training, 15% validation, 15% test

Step 4: Model Building- Random Forest Classifier is highly recommended because it is imbued with solid generalization, accuracy, and precision capabilities and, in all cases invariably performs well over a wide variety of datasets. Determine first Initial Hyperparameters, i.e., Number of Trees, Maximum Depth of Trees, Minimum samples per Leaf**.**

Step 5-Training the Model: Use sci-kit-learn in Python to implement the Random Forest Classifier with the following code. Set Up the training of a model on the training data by implementing the fit technique. Assess your model performance with methods, such as Grid or Random Search, on the validation set to get the best tune hyperparameters.

Step 6-Model Evaluation: Evaluate the developed model with metrics such as accuracy, precision, recall, F1-score, ROC-AUC curve, etc. It needs to perform well on the test dataset and generalize to new data.

Step 7: Model Deployment: Set up the appropriate infrastructure for implementing the model (e.g., cloud services such as AWS, Azure, or on-premise servers). Crat an API for real-time predictions. Use frameworks such as Flask or FastAPI to expose the model.

Step 8: Monitoring and Maintenance-Continuously monitor the model's performance using metrics and adjust the model as needed. Regularly update the model with new data to maintain its accuracy and relevance.

Business Impact

Benefits for Businesses in the USA

Flexibility and Scalability: The Random Forest algorithm is relatively easy to update with new data. Besides, the model can be scaled to large datasets, which are necessary for extensive supply chain networks.

Risk Management and Mitigation: The model helps to identify the potential risks of sustainability in regards to the supply chain; for example, high carbon footprint or non-compliance with too many suppliers.

High Accuracy and Robustness: The Random-forest classifier deals with high-dimensional data. Since the number of features is large in datasets related to supply chains, this model handles that by averaging the outputs of multiple decision trees, hence decreasing over-fitting and increasing generalization to new unseen data of the model.

Cost Efficiency: The Random Forest algorithm helps identify unsustainable practices so businesses optimize resource utilization, waste less, and achieve low operational costs. Timely detection of sustainability issues provides opportunities to substantially intervene in processes to avoid costly repercussions related to unsustainable practices.

Benefits for the USA Economy

Reduced Carbon Footprint: Random Forest Leads Detection of CO2 that in turn reduces greenhouse gas emission in the USA. This helps national efforts to address climate change and meet all the international environmental treaties. It means less pollution and better air quality for healthier communities, gaining lower health costs and higher quality of life for its citizens.

Attracting Foreign Investment: Deploying Random Forest can lead to sustainable chains, attracting investors looking for sustainable and socially responsible investment opportunities. Government companies in the USA with sustainable chains are more likely to attract investment that can fuel other economic growth.

Enhancing Competitive Advantage-Investing in sustainability helps government companies to position themselves as leaders in the global marketplace, attracting international business and investments. This makes American companies more competitive in the world.

Conclusion

This research paper explored the application of predictive analytics and machine learning in the domain of supply chain management for reducing carbon emissions and granting sustainable operations. For the current study, Walmart organization provided all the supply chain data used in this research study. It consisted of comprehensive data on their product segment groupings that fall into three main categories. Based on all the metrics, Random Forest was the best classifier because of its excellent generalization, high measure of precision and recall, and highest AUC. As per the results, the random forest algorithm was the most accurate in its predictions of all the models evaluated. Implementing the random forest benefits businesses in America in terms of high accuracy and robustness, flexibility, and scalability, risk management, and Mitigation. As regards the US economy, deploying the Random Forest can benefit the government in the following ways: reducing carbon footprint, attracting foreign investment, and enhancing competitive advantage.

References

[1] Adewusi, A. O., Komolafe, A. M., Ejairu, E., Aderotoye, I. A., Abiona, O. O., & Oyeniran, O. C. (2024). The role of predictive analytics in optimizing supply chain resilience: a review of techniques and case studies. *International Journal of Management & Entrepreneurship Research*, *6*(3), 815-837.

[2] Abdollahi, A. (2023). An integrated machine learning and quantitative optimization method for designing sustainable bioethanol supply chain networks. *www.academia.edu*. https://www.academia.edu/103549448/An_integrated_machine_learning_and_quantitative_optimization_method_for_designing_sustainable_bioethanol_supply_chain_networks?sm=b

[3] Bag, S., Rahman, M. S., Srivastava, G., Chan, H. L., & Bryde, D. J. (2022). The role of big data and predictive analytics in developing a resilient supply chain network in the South African mining industry against extreme weather events. *International Journal of Production Economics*, *251*, 108541.

[4] Brandtner, P. (2022). Predictive analytics and intelligent decision support systems in supply chain risk management—research directions for future studies. In *Proceedings of Seventh International Congress on Information and Communication Technology: ICICT 2022, London, Volume 3* (pp. 549-558). Singapore: Springer Nature Singapore.

[5] Gonuguntla, V. (2016). Application of machine learning techniques for supply chain demand forecasting. *www.academia.edu*. https://www.academia.edu/26506366/Application_of_machine_learning_techniques_for_supply_chain_demand_forecasting?sm=b

[6] Heydar, M. (2019). Supply chain Demand Forecasting; A comparison of machine learning techniques and traditional methods. *www.academia.edu*. https://www.academia.edu/119107304/Supply_Chain_Demand_Forecasting_A_Comparison_of_Machine_Learning_Techniques_and_Traditional_Methods?sm=b

[7] Khanai, R. (2022). Machine Learning Application for Automotive Emission Prediction. *www.academia.edu*. https://www.academia.edu/68450866/Machine_Learning_Application_for_Automotive_Emission_Prediction?sm=b

[8] Mandala, V. (2024). Revolutionizing Automotive Supply Chain: Enhancing Inventory Management with AI and Machine Learning. *Indiana*. https://www.academia.edu/117704667/Revolutionizing_Automotive_Supply_Chain_Enhancing_Inventory_Management_with_AI_and_Machine_Learning?sm=b

[9] Marwa, F. (2023). Predictive Machine Learning Algorithms for CO2 Emission Impacts in the UK. *www.academia.edu*. https://www.academia.edu/106287450/Predictive_Machine_Learning_Algorithms_for_CO2_Emission_impacts_in_the_UK?sm=b

[10] Nzeako, G., Akinsanya, M. O., Popoola, O. A., Chukwurah, E. G., & Okeke, C. D. (2024). The role of AI-driven predictive analytics in optimizing IT industry supply chains. *International Journal of Management & Entrepreneurship Research*, *6*(5), 1489-1497.

[11] Oliveira, M. (2021). Lead Time Forecasting with Machine Learning Techniques for a Pharmaceutical Supply Chain. *www.academia.edu*. https://www.academia.edu/103962910/Lead_Time_Forecasting_with_Machine_Learning_Techniques_for_a_Pharmaceutical_Supply_Chain?sm=b

[12] Oyewole, A. T., Okoye, C. C., Ofodile, O. C., & Ejairu, E. (2024). Reviewing predictive analytics in supply chain management: Applications and benefits. *World Journal of Advanced Research and Reviews*, *21*(3), 568-574.

[13] Park, K. (2021). Determining the tiers of a supply chain using machine learning algorithms. *www.academia.edu*. https://www.academia.edu/61744998/Determining_the_Tiers_of_a_Supply_Chain_Using_Machine_Learning_Algorithms?sm=b

[14] Tech, I. R. J. M. S. (2020). Application of machine learning techniques in supply chain management. *www.academia.edu*. https://www.academia.edu/43412135/Application_of_Machine_Learning_Techniques_in_Supply_Chain_Management?sm=b