Scalable Retail Demand Sensing with Feature-Engineered Time Series using Random Forest and LightGBM

Tejaswini K[1] , Chiranth RD[2] , Utsav Kumar[3] , Devansh Shringi[4] , Gurvir Singh[5]

[1,2,3,4,5]*Computer Science and Business Systems, B.M.S. College of Engineering, Karnataka, India*
tejaswinik.cbs@bmsce.ac.in[1], chiranthrd.bs23@bmsce.ac.in[2], utsavkumar.bs23@bmsce.ac.in[3], devansh.bs23@bmsce.ac.in[4], gurvirsingh.bs23@*bmsce.ac.in[5]*

Abstract

Disruptions that affect supply chains can halt the growth of their individual networks. Early and accurate diagnosis of systemic risks can reduce the likelihood of further damage to the network. The approach described here requires more time, focus, and specialized skill. Real-time ERP and TMS data are used to identify vulnerabilities within the supply chain. Research in artificial intelligence shows strong potential for improving accuracy in risk detection. With major advancements in machine learning, there is now a greater opportunity to enhance coordination and precision in systems that identify and interpret supply-chain risks. This study proposes an AI-based framework for supply chain risk detection and classification that combines reinforcement learning with an LSTM classifier. Reinforcement learning is used to explore and improve risk diagnosis decisions, while the LSTM model extracts patterns related to demand volatility, lead-time variability, and supplier reliability from real-time operational data. Several performance metrics used in the analysis demonstrate that the proposed method outperforms existing techniques, achieving higher accuracy. The phases involved in the risk-detection process include: data acquisition, anomaly detection, noise removal, and prescriptive orchestration.

Keywords- Supply chain risk management, Risk detection and classification, Long Short-Term Memory (LSTM), Anomaly Detection, Perspective Analysis, Gradient Boosting

1. Introduction

Artificial Intelligence (AI) is becoming one of the most important facilitators of enhancing supply chain risk management in terms of circumstances marked by frequent disruptions, geopolitical shocks, and climate-related events. The modern supply chains cross various levels and across regions, thus being incredibly prone to the domino effect when a single node or lane is corrupted. These complex, interdependent risks cannot be predicted and managed using traditional, deterministic methods that use fixed risk registers and periodic evaluations.

The capabilities of AI-based approaches, such as machine learning, optimization, and simulation enable the ability to detect risks and forecast their occurrence and prescribe proactive measures to mitigate risks at the end-to-end networks. AI can enhance visibility, point out important weak points, and aid decision-making about rerouting shipments, rebalancing inventory, or changing suppliers before a disruption grows by combining the information of the internal systems with external sources. Such capabilities transform organizations where they are reactive in fighting fires to one where they are on a continuous data-driven orchestration of risk.

The further development of AI in the context of risk management concerning the transition of simple resilience into the realm of the antifragile supply chains is helpful since organizations also learn and become more effective in managing stresses as they grow more mature in their risk-management practices. AI-enabled cognitive control towers and digital twins allow supply chains to recover more quickly and respond in real-time using scenario analysis and automated responses, as well as flex their structures, policies, and partnerships after every disruption. Studying the potential of AI to be applied systematically in minimizing supply chain

risks and developing antifragile capabilities is thus a relevant and effective research opportunity.

The use of AI-based supply chains is based on the continuous stream of data provided by enterprise systems, logistic partners, IoT devices, and external risk feeds to observe the operations in real-time and produce early-warning signals. Machine learning models process these data and identify anomalies in demand, lead times, capacity utilization or shipment performances and send alerts before small deviations can become significant disruptions. Besides the detection of risks, AI can put itself to imagined cases of disruption and analyze alternative responses, e.g., a production re-location, freight diversion, or a redistribution of safety stock throughout the network.

One of the main themes of the AI-driven risk reduction is the shift between descriptive dashboards to prescriptive and autonomous decision support in some scenarios. The AI systems have the ability to prioritize risk-reduction choices by considering constraints such as cost, service levels and carbon footprint and allow the decision-makers to trade-off resilience and efficiency instead of one off against the other. These models are reinforced over time, as the responses to the implemented strategies can inform the supply chain on the most efficient strategies to apply in particular circumstances, which makes the system antifragile.

Non-trivial challenges however exist in the deployment of AI in the area of supply chain risk management in the form of data, technology, and governance. The inaccuracy of risk predictions might be caused by the fragmented data in ERPs, TMSs, WMSs, and partner systems, and black-box models might fail to earn the planners, auditors, and regulators trust. The issue of cybersecurity, bias in algorithms, and the development of AI policies only exacerbates large-scale implementation and explains the necessity to have explainable models, strong data management, and explicit accountability structures.

With such obstacles, empirical research demonstrates that organizations that effectively incorporate AI into supply chain risk procedures report quicker response times, less stock outs and better supply continuity in case of disruptions. With the growing volatility in the global context, the capacity to integrate human knowledge with AI-driven information will become one of the defining features between the supply chains that only survive the shock and those that evolve and become stronger due to the shock. The study thus makes AI not only a supplementary technology but also a movement power of risk-conscious antifragile supply chain design and operation.

2. Literature Review

The AI-based demand sensing has become a unique branch in the demand forecasting due to the growing volatility of the supply chain, the reduction of the product lifecycle, and the possibility of the high-frequency data provided by the digital channels. As opposed to the traditional forecasting that uses historical sales and low-frequency updates as its primary sources of information, demand sensing attempts to predict the near-term demand based on real time information like point-of-sale (POS), social media, promotions, weather and macroeconomic information and hence demand responsiveness and inventory performance is enhanced.

The preliminary studies of AI-driven demand sensing prove that the machine-learned and deep learning models can significantly enhance the performance in terms of short-term predictions and operational results when fed with a variety of information streams and being regularly updated and replenished with the new data. Recent publications on AI-based demand sensing in supply chains demonstrate that time-series models, paired with the state-of-the-art methods, including neural networks, natural language processing (NLP), and predictive analytics, can help the system discover nonlinear relationships between external drivers and demand to avoid stockouts, overstocks, and agile responses to replenishment decisions. Another common thread in these contributions is that it highlights some of the practical issues, such as the quality of data, the necessity to join up heterogeneous sources, and the necessity to train the models on a continuous basis as the demand pattern and the market conditions change.

The AIs are also used to incorporate the notion of short-term demand sensing with medium- to long-term forecasting within the same architecture, especially across critical or high-risk supply chains. The recent activity on the crucial and medical products incorporates the real-time signal extraction

with adaptive forecasting models updating themselves with emerging information and assisting in decisions that are critically uncertain and disrupted. These structures underline the purpose of AI in the context of forecasting points except the accuracy of the forecast, enhancing the strength and the adaptability of end-to-end planning, such as safety-stock policies, capacity assignment, and emergency sourcing plans.

Deep learning–based demand forecasting underpins modern demand sensing through the capture of rich temporal and nonlinear patterns that are often missed by more classical statistical models, especially under regime shifts. LSTM and its related recurrent architectures often combined in ensembles with other ML models provide better bias–variance trade-offs and outperform traditional approaches on error metrics such as MAPE and MSE. Recent work to address data sparsity and intermittency at granular levels integrates ADI-CV–based demand segmentation with augmentation of time–series data using approaches such as MBB, T-CGAN, and TTS-CGAN to create realistic synthetic histories that improve generalization in volatile environments. Consistent empirical results show 20–50% error reductions and attendant gains in inventory, service levels, and capacity utilization, although robust data infrastructure, cross-functional governance, and processes to manage model drift and integrate AI outputs into human planning are also discussed as important.

3. Problem Statement

Conventional supply chain risk management approaches have poorly delivered the accurate and timely insights needed for decision-makers. Most of these methods are slow, reactive, and rigid, grounded on fragmented data and manual analytics that cannot bear the realities of volatility and complex patterns of disruption. This research will introduce an Artificial Intelligence–driven framework for Supply Chain Risk Reduction that integrates advanced machine learning, risk sensing, and simulation based on digital twins. The proposed approach seeks to overcome identified limitations and achieve more reliable, proactive, and scalable risk mitigation across end-to-end supply networks.

4. Proposed Methodology

Supply chain risk detection in this study involves collecting multiple datasets, including historical and semi-annual port congestion data (time spent in ports), port activity data (number of port calls), sales training/validation data, selling prices, and external maritime indicators such as vessel size/age statistics. Once we collect the data, we align the daily sales figures with monthly shipping logs. We then create features—like time lags and moving averages—to track how delays ripple through the supply chain. Finally, we train Random Forest and LightGBM models to predict demand, identifying risk by spotting where accuracy drops during congestion spikes.

4.1. Data Collection

The data collection in this study includes both internal and external supply-chain datasets. The internal data consists of historical sales data, validation data, and selling prices. The external data includes annual and semi-annual port congestion statistics (time spent in ports) and port activity datasets (number of port calls) downloaded from the UNCTADstat database. In addition to this, supplementary datasets were collected from Kaggle to support demand patterns and external risk indicators. Together, these datasets provide a complete view of sales behaviour, pricing changes, and real-world port conditions needed to study supply chain risks.

4.2. Data Pre-processing

To remove errors, missing values, and noise from the raw data, several cleaning techniques were applied. First, missing sales and price values were filled using forward-fill and median replacement to ensure smooth demand patterns. Outliers in sales and price data were detected using statistical thresholds and corrected to avoid misleading spikes. The port congestion and port call datasets were pre-processed by standardizing date formats, converting units, and merging different time periods into a single consistent timeline. Additional steps included generating lag features, moving averages, and calendar features to enhance the model's ability to learn patterns.

4.3. Feature Engineering

Feature Engineering converts our access raw sales, port congestion, and maritime data into features,

which are actually learnable by the models. First, there are calendar characteristics like- day, week, month, and holiday flags that seek advantage of seasonality and repeats in demand. Then lag characteristics and rolling averages are constructed using past demand and congestion levels to ensure the models associated with such historical disruptions with the current demand. Lastly, the external variables such as the price, promotion, lead time, and disruption indicators are coded into the internal demands of the models so that the port risk and logistics delay can be transformed into a fluctuation in downstream demand.

4.4. Model Training

To prevent data leakage, Model training starts with a chronological train-validation-test split. Linear and Ridge models are easy to interpret and establish simple baselines, whereas the random-forest and light-gbm identify non-linear relationships in the features of demand and disruption. Time-based cross-validation is applied to all the models and the most suitable model is chosen based on the RMSE and MAE on the held out test set.

4.4.1. Linear Regression

Linear Regression is the initial layer in our model stack, as it offers an easy yet understandable manner to view responsiveness of demand to core drivers such as price, promotions, calendar characteristics and any type of disruption. The features are each weighted (coefficient) to indicate the amount of increase or decrease in demand to be realized when such a feature changes by one unit without any other factors being altered- such as how a promotion flag, or an increase in degree of congestion level, will alter sales.

$$\hat{y}_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t}$$

where β_0 is the intercept and $\beta_1,\ldots,\beta_k$ are learned by minimizing the sum of squared errors between actual and predicted demand. In our project, this baseline lets us compare more advanced models (Random Forest, LightGBM) to a simple, explainable reference and verify that added complexity genuinely reduces RMSE and MAE, especially during disruption periods.

4.4.2. Ridge Regression

Ridge Regression is used as a stronger linear baseline because many engineered features (lags, rolling statistics, congestion indicators) are highly correlated. Unlike simple Linear Regression, Ridge is less volatile under such multicollinearity and generalizes better to unseen periods.

Ridge addresses this by adding an L2 penalty that shrinks coefficients toward zero while still keeping all predictors in the model. The objective is to minimize squared error plus a penalty on the sum of squared coefficients, where $\lambda \geq 0$ controls the strength of regularization. In our case, this leads to more stable, less overfit demand forecasts that still remain interpretable, providing a fair comparison point before moving to nonlinear tree-based models.

4.4.3. Random Forest

The first nonlinear model of our stack is the Random Forest Regression used to model the complex interaction between demand and congestion features which cannot be modeled by the linear and Ridge models. Multiple decision trees are trained on varying bootstrap samples and random feature subsets (lags, rolling statistics, prices, promotions, port indicators) per SKUlocationtime and the results are averaged to have a stable forecast. This ensemble design enables the model to be resistant to noise and outliers as well as having a better RMSE and MAE during both the regular and disruption regimes.

For Random Forest Regression in your project, the prediction for demand at time t is: $\hat{y}_t = (1/M) \sum f_m(x_t)$, where M is the total number of trees in the forest, x_t is the feature vector at time t (lags, rolling stats, prices, promotions, congestion, linestore, calendar effects), $f_m(x_t)$ is the prediction made by the m-th decision tree given x_t, and $\hat{y}_t$ is the final forecasted demand, obtained by averaging all tree predictions.

4.4.4. LightGBM Regression

The latest model in the stack is LightGBM Regression, and it is aimed at the high-nonlinearity relationships between the demand and the rich feature sets, including lags, rolling statistics, price, promotion, calendar variables, and congestion indicators. LightGBM uses gradient boosting to create trees in series as opposed to the averaging of so many independent trees as with Random Forest, meaning that the new tree should target the residual errors of

the existing ensemble, and the model is incentivized to learn during hard-to-predict intervals such as disruption.

This causes LightGBM to be especially useful in structured sales and demand data where numerous weak signals are summed up into one very strong signal. Formally, LightGBM models the predicted demand at time t as a sum of M regression trees: $\hat{y}_t = F_m(x_t) = \sum f_m(x_t)$, where x is the feature vector at time t, and each f is a decision tree added at boosting step m. Training minimizes a regularized objective that balances data fit and model complexity. In practice, LightGBM uses second-order gradient information (both gradient and Hessian) and a histogram-based, leaf-wise tree growth strategy, which speeds up training on large historical demand datasets while improving accuracy, leading to the best RMSE and MAE in the project's experiments.

4.5. Evaluation Metric

In order to evaluate the performance of the suggested demand forecasting model, we measured it with the help of two commonly used statistical measure: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). These metrics give one a clear picture of the predictive ability of the model in absolute sense, which is critical in decision making when asking a retail and supply chain forecast.

4.6. Best Performing Model

LightGBM was the most successful model of the considered models, as the model has the highest accuracy of predictions and generalization on the test dataset. The model was trained and evaluated on the daily sales data hence suitable in time series prediction in retail and supply chain use.

4.7. Metric Results

The results of the evaluation of the LightGBM model against benchmark ranges used in similar studies in forecasting are given in the following table:

5. Result and Discussion

The four forecasting models we evaluated were Linear Regression, Ridge Regression, Random Forest and LightGBM. To evaluate the performance of each model, we used RMSE and MAE as performance metrics. The results of our comparison of these models on the test dataset are displayed in Table below.

LightGBM model is the lowest in RMSE and MAE, which proves a much higher degree of forecasting accuracy. Random Forest model showed a moderate increment on the linear models and still could not match the predictive performance of LightGBM. Linear Regression as well as Ridge Regression produced the highest error values, which indicates a low degree of aptitude to describe the complicated dynamics found in data.

Also, it was experimentally established that the training on raw sales data only resulted in greater forecast. The addition of the engineered characteristics, including lag variables, rolling window statistics, holiday indicators, and port congestion signals, resulted in significant increases in the accuracy of the predictions of all the types of models.

As it is seen when compared with other models like Linear Regression, Random Forest and XGBoost our model gives the lowest RMSE (Root Mean Square Error) and MAE (Mean Absolute Error). Making our model a better choice when compared to other existing models.

5.1. Discussion

These results indicate that LightGBM is better than classical models of regression and the Random Forest ensemble when it comes to short-term supply chain demand forecasting. The higher performance is explained by the framework of gradient boosting of the model that makes it possible to represent nonlinear trends, intricate interactions of features, and minor changes in demand patterns. This is necessary where the changes are irregular or disruption based in nature in the environment where linear relationships cannot be applied.

The enhancements driven by feature engineering highlight that it is a very important activity in forecasting time-series activities. Models were able to adjust to seasonality and autoregressive structure by using temporal characteristics like lagged demand and rolling averages and sensitivity to external drivers of operation by using contextual features like calendar effects and congestion indicators. The obtained results support the idea that the accuracy of the predictions is not only based on the model

architecture but also on the relevance and expressiveness of the input feature space.

This finding aligns with prior literature where a number of studies showed that the performance of boosting-based models is very strong in dynamic forecasting applications. The findings of this study further confirm the hypothesis that traditional regression methods might be too simplistic for actual-world supply chain scenarios, where demand patterns are associated with nonlinearity, uncertainty, and operational disruptions.

From an application viewpoint, LightGBM's improvement in the accuracy of forecasting allows for better efficiency in supply chain decision-making. Improved predictive accuracy can help firms optimize inventory placement, reduce safety stocks, and proactively meet the ever-changing customer demand. The reliability of such models, however, remains dependent on continuous retraining and further high-quality, updated data.

The future directions may involve the integration of hyperparameter optimization frameworks, extending the coverage of exogenous variables, and exploring hybrid architectures for boosting with neural time-series forecasting models.

6. Conclusion

AI-driven approaches to supply chain risk management are reshaping how organizations anticipate, absorb, and recover from disruption. This literature review shows that techniques such as machine learning, natural language processing, optimization, and digital twins can transform fragmented operational data into forward-looking risk insights and prescriptive recommendations. When embedded in cognitive control towers and other decision-support platforms, these capabilities enable faster detection of emerging threats, more agile response strategies, and better alignment between resilience, cost, and service objectives.

At the same time, the review highlights that realizing this potential depends on overcoming persistent challenges in data quality, model transparency, and lifecycle management. Issues such as data silos, model drift, and limited explainability can undermine trust in AI outputs and restrict adoption in high-stakes planning and execution contexts. Addressing these gaps calls for robust MLOps practices, explainable AI techniques, and governance frameworks that clarify roles, responsibilities, and risk ownership across the supply chain ecosystem.

Overall, AI is emerging not merely as an incremental tool but as a strategic capability for building antifragile supply chains that improve through exposure to volatility rather than being weakened by it. By systematically integrating advanced analytics into risk sensing, scenario planning, and autonomous decision-making, firms can move beyond reactive firefighting towards a probabilistic, self-improving operating model that is better equipped for an uncertain future.

References

[1] Baryannis, G., Validi, S., Dani, S., & Antoniou, G. (2019). Supply chain risk management and artificial intelligence: State of the art and future research directions. International Journal of Production Research, 57(7), 2179–2202.

[2] Ivanov, D. (2017). Ripple effect and supply chain disruption: A review of quantitative studies. International Journal of Production Research, 55(7), 2079–2096.

[3] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. PLOS ONE, 13(3), e0194889.

[4] Schoenherr, T., & Speier-Pero, C. (2022). The digital transformation of supply chains: A review and research agenda. Journal of Business Logistics, 43(1), 4–26.

[5] Taleb, N. N. (2007). The Black Swan: The Impact of the Highly Improbable. Random House.

[6] Choi, T.-M. (2020). Future of artificial intelligence and its influence on supply chain risk management. Computers & Industrial Engineering, 149, 106833.

[7] Wang, Y., & Chen, X. (2023). AI in supply chain risk assessment: A systematic literature review. arXiv preprint arXiv:2401.10895.

[8] Li, J., & Kumar, S. (2024). The role of artificial intelligence in supply chain risk management: Towards an integrated conceptual framework. SSRN Electronic Journal.