PREDICTING LATE DELIVERIES IN HUMANITARIAN SUPPLY CHAINS USING AN ENSEMBLE MACHINE LEARNING APPROACH

Fahd Alfarsi and Mohammad Alshehri
University of Jedda, Saudi Arabia

Abstract

Purpose: This study focuses on predicting late deliveries within humanitarian supply chains using a case study dataset to address this critical issue.

Design/ methodology/ approach: This paper proposes several ensemble machine learning methodologies aimed at mitigating shipment risks by predicting the likelihood of late deliveries based on a detailed understanding of the procurement process.

Findings: Our findings demonstrate that an ensemble algorithm-based prediction model can effectively forecast the severity of late deliveries by suppliers in a representative case study of humanitarian supply chains.

Originality/ value: By providing a better understanding of shipment risks, this study aims to reduce uncertainty and improve the efficiency of humanitarian supply chain operations. This research contributes to the existing literature by showcasing the applicability of advanced machine learning techniques in enhancing the resilience and effectiveness of humanitarian supply chains.

Keywords: Humanitarian Supply Chains, Late Delivery Prediction, Supply Chain Resilience, Machine Learning, Delivery Forecasting

Introduction

The complexity of humanitarian supply chains presents unique challenges that differentiate them from commercial supply chains. Humanitarian logistics often operate in environments where timeliness is critical, as delays can have significant and sometimes life-threatening consequences. Whether responding to natural disasters, conflict zones, or public health crises, humanitarian organizations must navigate uncertain and dynamic environments with limited resources, fragile infrastructure, and unpredictable demand patterns. In this context, the ability to accurately predict late deliveries is crucial to improving the efficiency and effectiveness of humanitarian relief efforts.

Humanitarian supply chains differ significantly from their commercial counterparts in several ways. First, they are often designed to operate under extreme uncertainty, where demand can surge unexpectedly, as seen in disaster relief or during pandemics. Second, the geographical and logistical challenges—such as damaged infrastructure, political instability, or natural barriers—compound the difficulty of ensuring timely deliveries. Finally, humanitarian operations are mission-driven, prioritizing the well-being and survival of affected populations rather than financial profit, which often leads to challenges in resource allocation, coordination, and decision-making. Late deliveries in such contexts can lead to severe consequences, ranging from the failure to provide critical medical supplies, food, and shelter to displaced populations, to the exacerbation of crises due to delays in essential services. Therefore, mitigating delays in humanitarian supply chains is not merely a matter of operational efficiency but also one of ethical responsibility and social impact. In recent years, advances in machine learning have opened up new opportunities to enhance the predictability of supply chain outcomes. Traditional methods, such as statistical analysis and deterministic models, have struggled to account for the variability and complexity inherent in humanitarian logistics. Machine learning, particularly ensemble methods, offers a promising alternative by leveraging data-driven insights to make more accurate predictions about delivery times. Ensemble machine learning models, which combine the strengths of multiple algorithms, have shown significant improvements in prediction accuracy across various domains.

The present paper aims to apply various EML approaches to predict late deliveries in humanitarian supply chains, thereby improving the timeliness and reliability of relief efforts. The study leverages various machine learning techniques, namely Extremely Randomised Trees (ET), Gradient Boosting Machines (GBM), Adaptive Boosting (AdaBoost), and Stochastic Gradient Boosting (XGBoost), to

build a predictive model that can identify potential delays before they occur. By utilizing historical data from humanitarian logistics operations, the model is trained to recognize patterns that are indicative of late deliveries, such as geographical challenges, transportation bottlenecks, and supply chain disruptions. The key contributions of this research are threefold. First, it develops a novel ensemble-based predictive model specifically tailored for humanitarian logistics, addressing the unique challenges posed by these supply chains. Second, it offers insights into the most influential factors contributing to late deliveries, providing valuable information for supply chain managers and policymakers. Third, the study demonstrates the practical applicability of EML models in improving the resilience and responsiveness of humanitarian operations.

Literature Review

A few studies adopted ML for managing SC disruptions, e.g., predicting late delivery. Thomas and Panicker (2022) used eight features, namely unit weight and quantity, plant code, customer ID, Carrier ID, export and destination ports and product ID, to develop their ML-based framework for predicting order delivery delay. Since the data was significantly imbalanced, with a ratio of 97:30 for on-time and delayed cases, respectively, a common technique, SMOTE, was used to oversample the minor class, which handles imbalanced datasets by generating synthetic samples for the minority class, has been employed. Using various conventional ML algorithms, Random Forest (RF), k Nearest Neighbors (k-NN), and Support Vector Classifier (SVC), the delayed delivery predictive model was built on only data captured from one day, which may arguably introduce a potential algorithmic bias.

Baryannis, Dani and Antoniou (2019) developed a delayed delivery risk prediction model, considering the trade-off between model interpretability for non-AI experts and the model's performance. The data was obtained from an aerospace manufacturing supplier (AMS) to partners in Europe and Asia and contains the historical deliveries of 450 suppliers. The features utilised for building the predictive model were 15 features, including details about the quantities prices of the purchased items, the detailed dates of the purchase orders, the due dates, the original request as well as delivery dates. Various performance metrics were adopted to examine the performance of the proposed model, including Recall, Precision, Matthews Correlation Coefficient, F1 and Accuracy.

Brintrup *et al.* (2020) used historical data from an Original Equipment Manufacturer (OEM), precisely a complex engineering assets company, to predict first-tier SC disruptions. The experiment involved three defined phases: (1) identifying potential representative variables that can be valuable in predicting disruptions, (2) developing performance metrics to assess the model performance successfully, and (3) experimentally analyse the performance of various predictors (algorithms), and finetuning their parameter. The study indicates that including further pre-processed variables with the dataset helped the model outperform other experiments with a promising accuracy of 80% on average.

Unlike the above studies, which built disruption prediction models, Kosasih and Brintrup (2022) examined the visibility of supply chains from a link prediction perspective and detected potential links unidentified to the buyer using Graph-based model. The study applied Unified Gradient to enhance the explainability of the proposed model by identifying key input features that impact GNN's decision-making process. Additionally, the benefits and limitations of GNN for link prediction were analysed. The dataset utilized in this research consists of real automotive data, including 8,663 firms and 68,812 procurement relationships from international companies.

Concerns Regarding the Present Studies

The above studies have novelly contributed to the area of applying ML for SC disruption detection and timely prediction. However, a few methodological concerns emerged based on in-depth reading of the surveyed works. One concern is the inappropriate selection of the metrics utilised to assess the model performance. While the metrics that are commonly used with imbalanced datasets have already been identified and used, including AUC and the three computed versions of F, reporting the Rec. at the level of class was surprisingly missing from all the surveyed works. This would provide insight into how the model successfully detected each class (delayed or not delayed) independently, which is a critical indicator when the dataset is imbalanced.

Discussing the reported metrics has consequently highlighted the misselection of the validation architecture (VA) in two of the surveyed studies, specifically Thomas and Panicker (2022) and Baryannis, Dani and Antoniou (2019). Adopting SMOTE, conducted by these two works, to handle

imbalanced datasets via oversampling minor classes with synthetic instances has been commonly known as a critical methodological step. However, using cross validation (CV) as a VA would potentially allow the model to train and test on the same data (original and synthetic). Therefore, the train-and-test splitting strategy is the proper VA for oversampled datasets.

One concern is the misrepresentation of some of the datasets used for building the disruption predictive models. While this limitation is typically out of the authors' control, it is important to note that the proposed models may be either ungeneralisable or biased. This concern is highlighted in Thomas and Panicker (2022), where the data used was captured from the supplies of one calendar day only (May 26th, 2013). Since building actionable and valid predictors requires various time points over the supply chain history, which is expected to inform various potential disruption events, Thomas and Panicker (2022) model may typically not be suitable to implement in the real-life world.

Preliminary filtration is a crucial step in data preprocessing, as it helps remove noisy and unrepresentative data, such as system logs and missing values. However, a significant limitation observed in some of the reviewed studies is the excessive filtration applied during experiments, which can reduce the generalizability of the results and limit the sample to a small subset of supply chains. One instance is Brintrup *et al.* (2020), who removed the Product Description in the first run of feature selection without weighing its potential correlation to the model binary outputs. While texts are relatively more challenging to preprocess than numerical data and occasionally require complex architectures, such as the state-of-the-art Natural Languages Processing (NLP) models, they may contain rich assets that would significantly help improve the disruption predive models.

Proper model output setting, explicitly defining the target variable, is a crucial step of data preprocessing; nevertheless, this was not done appropriately in some of the surveyed studies. For example, Baryannis, Dani and Antoniou (2019) selected the delivery status (early: when it is 8 or more working days ahead of schedule, on time: when it is between 7 working days earlier than the schedule and 3 working days behind the schedule, late: if it is 4 or more working days behind the schedule) as their predictive model output. Next, and for simplification as claimed, both early and on time were merged as on-time delivery. Early delivery may impose some challenges based on the nature of the procured items, and, therefore, it may not necessarily be considered an advantage. This includes (1) the increased holding cost, (2) inventory disruption, e.g., excess inventory leading to tying up capital and warehouse space, (3) insufficient space resulting in stockouts and production delays and (4) potential for overproduction.

Another methodological concern observed is the adoption of incorrect feature selection methods. Deciding on the correct feature selection mainly relies on the type of data in the dataset itself. More specifically, the input and output variables' data type (numerical or categorical) must be considered before choosing which feature selection approach can be adopted. We noticed that some studies, such as Baryannis, Dani and Antoniou (2019), randomly adopted one or more feature selection approaches regardless of the type of input and output features analysed. This would result in selecting fewer representative features and, hence, lower performance for the predictive model.

Methodology

Data Collection and Preprocessing

The dataset used in the present research was obtained from the U.S. Agency for International Development (USAID). The dataset includes comprehensive data about all health commodity orders (dated between 31/01/2016 and 05/03/2024). The dataset contains 105 features including:

- General information (e.g., Order Number, Country Destination, Type of Order, Method of Fulfilment, Mode of Transportation, Category of Item Tracer and Product, Product ID, Name of Product, Unit of Measure, Ordered Quantity, Shipped Quantity)
- Financial information (e.g., Funding Health Programme, Funding Source, Fiscal Year, Illustrative Price)
- Delivery Information (e.g., Estimated Lead Time in Days, Vendor Incoterm, Destination Incoterm, On Time Delivery, delay in Days, Delivery Progress).
- Dates (e.g., Decided Delivery Date, Revised Agreed Delivery Date, Order Entry Date, Requested Delivery Date, Estimated Delivery Date, Actual Delivery Date).
- Times in days for (Order Cycle Time, shipment Approval by USAID).

Ensemble Models
These types of models use several algorithms in the perdition processes. It aggregates the outcomes from different classifiers to produce a majority vote output. While each individual algorithm learns differently and its performance is application- and data-dependent, ensemble models address this by delivering a generalizable result, as they rely on multiple "voters" (algorithms). This approach also helps build a reliable model that reduces variance and mitigates the risk of overfitting (Asharf et al., 2020).

Ensemble model can be implemented through bagging or boosting. The former trains n base learners utilising the technique of random samples with replacement, allowing for parallel processing. In contrast, boosting works sequentially, with each model focusing on the misclassifications of the previous one (Teja, 2019). In bagging, different decision trees (DTs) are trained on various samples of the training data, and the final prediction is an average of all the trees' outputs. A prime example of this technique is Extremely Randomized Trees (ET). In boosting, ensemble members are added sequentially to correct the errors of previous models, adjusting the input data to emphasize the misclassified instances. The final output is a weighted average of the predictions (Brownlee, 2021). Boosting models contain adaptive boosting (AdaBoost), gradient boosting machines (GBMs), and stochastic gradient boosting.

Extremely Randomised Trees (ET)
This method shares a similar structure with Random Forest (RF), as both select a random set of features for node splitting. However, ET differs in two key areas: sampling and node splitting. This method samples from the whole dataset when constructing a tree, whereas RF uses bootstrapped samples (input subsampling with replacement). This reduces bias in ET by ensuring more diverse data subsets. Additionally, ET randomizes node splitting, reducing the influence of specific input variables and lowering variance. This enhanced structure has demonstrated improved performance on benchmark datasets compared to RF (Geurts, Ernst and Wehenkel, 2006).

Adaptive Boosting (AdaBoost)
AdaBoost is an early boosting-based ensemble algorithm developed to address challenges related to some specific games. The algorithm preforms a combination of weaker learners to build a robust model (Freund and Schapire, 1997). As its name suggests, adaptation occurs through the errors of previous trials, assigning less weight to misclassified instances and focusing on those in subsequent models. This adaptiveness helps correct earlier misclassifications (Schapire, 2013).

Gradient Boosting Machines (GBM)
Unlike AdaBoost, which focuses on penalizing misclassifications, GBM minimizes a loss function (such as log loss for classification or mean absolute error for regression) to improve prediction accuracy. It continuously reduces the loss using gradient descent until the model reaches an optimal point (Friedman, 2001).

Stochastic Gradient Boosting (XGBoost)
XGBoost is an optimised version of GBM with enhanced efficiency, flexibility, scalability and better performance in tasks that handles sparse data. As a more regularized and extended version of GBM, XGBoost improves generalization and overall model performance (Chen and Guestrin, 2016).

Evaluation Metrics
Several performance metrics were utilised to assess the models developed in this study. These include standard metrics such as Rec and Prec, which can be derived from the confusion matrix. Recall measures a model's ability to predict positive instances, while precision calculates the ratio of predicted positives that are indeed true positives. The formulas for these metrics are: *Rec = True Positives (TP) / (True Positives (TP)+False Negatives (FN))* and *Precision = True Positives (TP) / (True Positives (TP) + False Positives (FP)).*

Since these metrics evaluate class-level performance, additional metrics such as *F1* score and Weighted Recall (*WRec*) were adopted to assess the model's performance more comprehensively. The *F1* score balances precision and recall, while WRec computes each class's recall being weighted by the proportion of instances in that class relative to the entire dataset. This ensures that the performance on larger classes has more influence on the overall recall score, preventing smaller classes from disproportionately affecting the result.

Results and Discussion

Table 1 presents the performance of the four ensemble ML algorithms adopted for the present study for predicting late deliveries within a humanitarian supply chain. Each classifier's performance is evaluated using several key metrics: Precision (*Prec.*), Recall (*Rec.*), F1-Score (*F1*), and Weighted Recall (*WRec*). These metrics are calculated for both class 0 (on-time delivery) and class 1 (late delivery), offering valuable insights into how well each model performs in different delivery scenarios.

Classifier	**Class**	***Prec.***	***Rec.***	***F1***	***WRec***
ET	0	0.97	0.98	0.97	0.96
	1	0.91	0.89	0.90	
AdaBoost	0	0.97	0.97	0.97	0.96
	1	0.91	0.89	0.90	
GBM	0	0.97	0.98	0.97	0.95
	1	0.91	0.88	0.89	
XGBoost	0	0.97	0.95	0.96	0.94
	1	0.85	0.89	0.87	

Table 1: Late delivery prediction results distributed by classifier,
class 0 = on-time delivery, class 1 = late delivery.

Extra Trees (ET) demonstrates high effectiveness in predicting on-time deliveries, with an impressive Prec. of 0.97 and a Rec. of 0.98, hence F1 = 0.97. This indicates a reliable ability to classify on-time deliveries accurately. For late deliveries, ET shows a slight decline in performance, with Prec. and Rec. of 0.91 and 0.89, respectively, and an F1-score of 0.90. Nevertheless, the model remains robust in predicting late deliveries, as indicated by the weighted recall of 0.96, suggesting a well-balanced approach to both classes.

AdaBoost performs similarly to ET, with equally strong results for on-time delivery predictions. Both Prec. and Rec. for class 0 are 0.97, resulting in an F1-score of 0.97. For late deliveries, AdaBoost's performance closely mirrors that of ET, with a Prec. of 0.91, Rec. of 0.89, and an F1-score of 0.90. This consistency across classes suggests that AdaBoost is a reliable classifier for predicting both on-time and late deliveries, evidenced by its weighted recall of 0.96, which aligns with ET's performance.

Gradient Boosting Machine (GBM) also achieves comparable results, with class 0 Prec. of 0.97 and Rec. of 0.98, resulting in an F1-score of 0.97. However, its performance for class 1 (late deliveries) is slightly lower than ET and AdaBoost, with a precision of 0.91, recall of 0.88, and an F1-score of 0.89. Despite this small reduction, GBM still performs well overall, as reflected by its weighted recall of 0.95, demonstrating that it is a reliable model, though slightly less balanced than ET and AdaBoost. In contrast, XGBoost performs slightly less effectively than the others for predicting on-time deliveries, with a Prec. of 0.97, Rec. of 0.95, and an F1-score of 0.96. More notably, XGBoost shows a more significant drop in performance for class 1 predictions, where its precision is 0.85, recall is 0.89, and the F1-score is 0.87. These results suggest that XGBoost is less effective at predicting late deliveries compared to the other models. The weighted recall of 0.94 further reflects this decline, indicating that XGBoost is less balanced in handling both classes, particularly when predicting late deliveries.

Across all classifiers, the performance for predicting on-time deliveries (class 0) remains consistently high, with precision, recall, and F1-scores generally above 0.95. This consistency indicates that all models are highly effective at correctly identifying on-time deliveries. However, when it comes to predicting late deliveries (class 1), all classifiers show slightly lower performance, with XGBoost displaying the most significant drop. Both ET and AdaBoost strike the best balance between class 0 and class 1, while GBM follows closely behind, and XGBoost lags in late delivery prediction accuracy.
In a practical sense, the prediction of late deliveries is particularly crucial for humanitarian supply chains, where timely deliveries can significantly impact relief efforts. The results suggest that Extra Trees and AdaBoost are the most reliable classifiers for this task. Gradient Boosting (GBM) also performs well but is slightly less consistent, particularly in predicting late deliveries. XGBoost, though powerful in many machine learning contexts, appears to underperform in this specific scenario, especially in predicting late deliveries.

Conclusion

The study successfully demonstrates the potential of ensemble machine learning techniques to predict late deliveries in humanitarian supply chains. Through a representative case study, we show that models such as Extra Trees and AdaBoost outperform other approaches, providing a well-balanced capability to predict both timely and late deliveries. While Gradient Boosting Machine yields acceptable results, its slightly lower performance in predicting late deliveries suggests it may not be ideal in critical contexts where late deliveries are of paramount concern. XGBoost, despite its widespread use in various domains, proves less effective in this particular scenario, highlighting the need for more tailored approaches in the humanitarian supply chain domain. This research adds several contributions to the existing body of knowledge by offering valuable understandings into the integration of advanced machine learning techniques, improving both the resilience and operational efficiency of humanitarian supply chains.

References

- Asharf, J., Moustafa, N., Khurshid, H., Debie, E., Haider, W. and Wahab, A. (2020) 'A review of intrusion detection systems using machine and deep learning in internet of things: Challenges, solutions and future directions', *Electronics,* 9(7), pp. 1177.
- Baryannis, G., Dani, S. and Antoniou, G. (2019) 'Predicting supply chain risks using machine learning: The trade-off between performance and interpretability', *Future Generation Computer Systems,* 101, pp. 993-1004.
- Brintrup, A., Pak, J., Ratiney, D., Pearce, T., Wichmann, P., Woodall, P. and McFarlane, D. (2020) 'Supply chain data analytics for predicting supplier disruptions: a case study in complex asset manufacturing', *International Journal of Production Research,* 58(11), pp. 3330-3341.
- Brownlee, J. (2021) 'A Gentle Introduction to Ensemble Learning Algorithms', *Machine Learning Mastery*.
- Chen, T. and Guestrin, C. 'Xgboost: A scalable tree boosting system'. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794.
- Freund, Y. and Schapire, R. E. (1997) 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of computer and system sciences,* 55(1), pp. 119-139.
- Friedman, J. H. (2001) 'Greedy function approximation: a gradient boosting machine', *Annals of statistics*, pp. 1189-1232.
- Geurts, P., Ernst, D. and Wehenkel, L. (2006) 'Extremely randomized trees', *Machine learning,* 63(1), pp. 3-42.
- Kosasih, E. E. and Brintrup, A. (2022) 'A machine learning approach for predicting hidden links in supply chain with graph neural networks', *International Journal of Production Research,* 60(17), pp. 5380-5393.
- Schapire, R. E. (2013) 'Explaining adaboost', *Empirical inference*: Springer, pp. 37-52.
- Teja, P. S. (2019) 'Bagging and Boosting'. Available at: https://medium.com/@saitejaposam9/bagging-and-boosting-2b6cd4a6bda1 (Accessed 06/10/2022).
- Thomas, A. and Panicker, V. V. 'Application of Machine Learning Algorithms for Order Delivery Delay Prediction in Supply Chain Disruption Management'. *Conference of Innovative Product Design and Intelligent Manufacturing System*: Springer, 491-500.