National College of Ireland

Optimising Supply Chain Performance with Machine Learning for Predicting Late Deliveries

MSc Research Project
Data Analytics

Avni Rathi
Student ID: x22182918

School of Computing
National College of Ireland

Supervisor: Naushad Alam

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Avni Rathi
Student ID:	x22182918
Programme:	Data Analytics
Year:	2023 - 2024
Module:	MSc Research Project
Supervisor:	Naushad Alam
Submission Due Date:	12/08/2024
Project Title:	Optimising Supply Chain Performance with Machine Learning for Predicting Late Deliveries
Word Count:	6239
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Avni Rathi
Date:	16th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Optimising Supply Chain Performance with Machine Learning for Predicting Late Deliveries

Avni Rathi
x22182918

Abstract

This research explores the application of machine learning approaches to enhance e-commerce supply chain management, focusing on two critical aspects: managing late deliveries and improving sales prediction accuracy. Since late deliveries can have a major effect on sales performance, the study combines both factors to give a thorough understanding of their interrelated effects. This work examines the effectiveness of machine learning models in predicting the risk of late deliveries, employing Random Forest, Decision Tree, and Logistic Regression. The study utilized a comprehensive dataset containing sales and delivery information to evaluate model performance. The findings reveal that the Random Forest Classifier at 97.58% accuracy outperformed other models in predicting late deliveries after applying hyperparameter tuning to optimize the performance of the model, demonstrating the highest accuracy and robustness. Feature importance method was performed to find key predictors which impact model results. Further for sales prediction, this work investigates how machine learning algorithms can improve sales behavior prediction. The models evaluated include Linear Regression, Decision Tree Regression, and Lasso Regression. Among these, Decision Tree Regression achieved the best performance, with an exceptional R-squared value at 0.996 indicating superior accuracy in forecasting sales behavior. The research highlights the significance of effective data transformation, feature engineering, and categorical encoding in model performance.

1 Introduction

1.1 Background & Motivation

Supply chain management is one of the leading processes in modern business operations that covers the flow of goods and services from suppliers to customers. It offers timely delivery of goods, except for covering all key processes that promote products. It is no wonder that every entrepreneur defines transporting goods on time as the key element to measuring success. Although supply chain management is one of the most effective strategies to deliver high-quality products to customers at the required time but it faces a lot of challenges. While delays in transportation, inventory-related issues, and bottlenecks in production schedules are among the most widespread ones, late deliveries can be especially disadvantageous. The main reasons are that late deliveries can lead to higher costs, diminished customer satisfaction, and lost business opportunities.

1.2 Problem Statements

Many e-commerce businesses use traditional methods to optimize supply chain performance, which is not satisfactory without advanced technologies such as machine learning, which has been proven to be impressive for improving late deliveries and sales. Previous research has shown numerous issues such as data quality, lack of procedural integration, and the use of different standards that are currently hampering the adoption and full application of predictive analytics in the e-commerce supply chain Abideen et al. (2021). It is necessary to predict and prevent delays in the delivery schedule because a failure to perform operations on time can lead to a loss of business efficiency and satisfaction with the service. If companies use predictive analytics correctly, they can manage inventories, respond efficiently to demand fluctuations, and also enhance the overall supply chain performance.

1.3 Research Aim & Objective

This study aims to optimize eCommerce supply chain performance by leveraging machine learning for predictive analytics, focusing on reducing late deliveries and improving sales behavior prediction to enhance operational efficiency and customer satisfaction.

The objectives of this study are to assess the current state of predictive analytics adoption in eCommerce supply chains, identify and analyze gaps and limitations in existing predictive analytics practices, develop and implement ML models for predicting late deliveries, investigate the effectiveness of ML algorithms in predicting sales behavior, and propose recommendations for integrating predictive analytics into operational workflows to enhance supply chain management and customer satisfaction.

1.4 Research Questions

1. To what extent can machine learning algorithms be applied to supply chain data to manage and predict the risk of late deliveries?

2. How well can machine learning algorithms enhance the prediction of sales behavior in supply chain management?

1.5 Significance, Scope & Limitations

This study makes a strong theoretical contribution to the understanding of how machine learning can be utilized in eCommerce SCM. It gives an understanding of integrating predictive analytics to enhance theoretical applications. This reveals that advanced analytics are an important addition to the broader supply chain foundation. In the context of practical application, the results of this study could be used by businesses to improve supply chain performance and the degree of customer satisfaction Baryannis et al. (2019).

This study analyzed the Dataco Supply Chain dataset sourced from the Kaggle website, which includes detailed shipping information, delivery schedules, and inventory logs. Several machine learning models and related techniques were applied to predict the number of late deliveries and forecast sales behavior, selected for their compatibility with complex data and the ability to enhance the accuracy of the predictions focused on the e-commerce supply chain Brintrup et al. (2020). There are a few limitations related to

the quality and relevance of the dataset and models, which may affect the reliability and validity of the outcomes.

1.6 Structure of the Report

This report is structured to explore optimizing eCommerce supply chain performance with machine learning. The introduction describes the background of the research, its aims, and objectives. Related work (Section 2) constructing a theoretical framework. Methodology (Section 3) describes the stages of the study: data collection, preprocessing, model development, and evaluation. The discussion section interprets the results and investigates their implications for practice. Section 7 Concluding remarks outline the key findings and the practical implications for industry. The structure of the report ensures a thorough exploration of how machine learning can enhance eCommerce supply chain performance and provides both theoretical insights and practical recommendations.

2 Related Work

Supply chain management is one of the most significant application areas for emerging predictive analytics and ML to enhance performance due to the growing complexity of supply chains worldwide. The literature review explores the present situation of the field critically about machine learning modelling and its application in supply chain management while enhancing sales prediction and predicting late deliveries.

2.1 Theoretical Frameworks and Conceptual Foundations

The theoretical frameworks in supply chain management are essential for understanding and optimization of operational practices. This study discusses important theoretical perspectives such as supply chain network theory and the resource-based view Gayam et al. (2021). The theory underscores the complexity and interdependent nature of the network, helping to an overall understanding of product distribution, interaction, and collaboration among stakeholders. Similarly, the resource-based view suggests the distinctive advantage of the organization's unique resources in the competitive environment.

Conceptual foundations in supply chain management provide fundamental principles and structures that allow to modeling of supply chain operations. This leaflet sheds light on such foundational SCM concepts as logistics, inventory management, and demand forecasting, as well as highlights their importance in the context of predictive analytics and decision support. Logistics involves organizing the movement and storage of materials and goods across the supply chain Ghazal and Alzoubi (2021). Another foundational concept is inventories and their management or the balance between supply and demand, with techniques like JIT one of the inventory control techniques that, on the one hand, shortly before the products are due to be delivered and, on the other hand, makes sure that there is no excess stock of products. Forecasting demand refers to the use of statistics to understand how consumers make purchases.

2.2 Supply Chain Network Theory and Resource-Based View

Supply Chain Network study is based on a systematic understanding of the relations between entities in supply chains, focusing on the smooth distribution of materials, in-

formation, and finances Wallmann and Gerschberger (2021). This paper shows the importance of recognizing nodes and relationships for delivery performance. The current framework has an immediate effect on the organization of work for a late-delivery forecast, illustrating complex timings and relations between suppliers, manufacturers, distributors, and retailers. Therefore, the junction of this framework with machine learning has a positive influence on the accuracy of delivery delay forecasts by using real-time data from different nodes. Even after these improvements, there is still a need for more research to increase forecast accuracy and delivery performance because the current solutions are not enough.

The Resource-Based View is a managerial concept relevant to Supply Chain Management. According to the RBV theory, organizations must intelligently exploit their internal resources and capabilities to achieve advantages over a long period. Considering SCM, the RBV concept is used in the context of resources such as technology, human resources, and infrastructure, used to manage supply chains more efficiently and prospectively. If we try to apply this concept to predict late deliveries, it will be clear that it is accomplished by the availability and proficiency of machine learning, analytics, and integration capabilities. Therefore, by using the RBV concepts, organizations may design their commonality and differentiation personality-based predictive models Shibin et al. (2020). As a result, not only does the delivery discipline improve, but organizations can also handle the challenges of the environment proactively and promptly.

2.3 Research on Late Delivery Performance

Several factors influence delivery performance in supply chains, including inventory management, transportation and supplier relationships Feng and Zhao (2023). Each of these factors must be properly managed to ensure the timely provision of goods. Effective Inventory or warehouse management is the accounting for the purchase, sale, and storage of goods, and optimal data for monitoring delivery. The more effectively these processes are managed, the more the company can earn. Transportation not only ensures that the goods are delivered but optimizes the cost and reduces delivery times while supplier management ensures timely delivery in the right quantity.

Supply chain management is supported by multiple models and theories, which significantly improve delivery performance and operational efficiency. Lean Six Sigma model relies on minimizing waste output and focuses on improvement through the decreasing rate of the processes. Just-in-time principles concern only the inventory amount available and offered strategies of its management. The whole production process is organized, so that there is no need to produce large amounts of goods beforehand and store them at considerable expenses. The Total Quality Management strategy provides tools to improve operations at all levels of the manufacturing company Tirkolaee et al. (2021).

2.4 Machine Learning Application and Case Studies of Sales Prediction

Machine learning applications in supply chain management are diverse and targeted at different crucial areas. Demand forecasting analyses historical data, and market trends to predict future needs, enhancing order accuracy and service rates and can also be used to optimize inventory levels Feizabadi (2022). Optimal inventory levels ensure timely expenditure of all products based on demand, flow and supplier feedback with the help

of ML. Route optimization reduces transportation costs and minimizes delivery time and predictive maintenance anticipates technique failures to schedule timely repairs.

2.4.1 Amazon's Demand Forecasting

Amazon uses machine learning algorithms to accurately predict customer's demand which reduces the risk of overstocking and shortage. The approach includes the selection of thousands of products with the highest purchase rates over a given period, and the historical data on the selection used is divided into two groups. Among these products are traditional seasonal attributes and external that affect customer interest Khan et al. (2020). Thus, Amazon uses a proactive approach to improve delivery efficiency and maintain the company in a competitive position.

2.4.2 UPS Route Optimization & Walmart's Inventory Management

The United Parcel Service employs machine learning for global route optimization, minimizing the cost and time of delivery by using real-time information. This is a proactive rather than reactive optimization approach of the UPS, which allows it to operate effectively and at low cost during peak periods of strong demand Janinhoff et al. (2024). ML is an integral artificial intelligence approach that Walmart uses to streamline inventory management across its wide supply chain. Walmart dynamically adjusts inventory to ensure an optimal supply of products in stores based on the analysis by analyzing sales trends, and seasonal fluctuations and having insights into supplier lead times Mehrotra et al. (2024).

2.5 Predictive Modelling and Decision Support Systems and Late Delivery in SCM

Predictive modelling and decision support systems improve supply chain process by forecasting outcomes based on historical data and current needs. Such systems use regression, time series analysis, and many more, ML algorithms like random forests or neural networks to facilitate the prediction of the demand, inventory, and delivery time Helo and Hao (2022). The prediction module is linked with decision support systems, and, therefore, the tool is used for quick decision-making, resource optimization, and timely actions to avoid the risk of being late with the delivery. The tools help the managers adjust to the market changes promptly.

Data preprocessing and feature engineering are major activities related to adjusting a supply chain dataset to efficient predictive modeling of late delivery. Cleaning includes all activities that help rectify inconsistencies, such as missing values and errors, in the dataset and ensure the dataset is clean. Normalization pertains to making all numerical attributes have equivalent ranges, as this way all features can be fairly compared Rai et al. (2021). Outlier detection seeks to identify all abnormal existing instances and provide unwarranted actions concerning their treatment or removal. Feature selection pertains to the final selection of all worth feature attributes that strongly impact the late delivery performance and have an implication for the model efficiency. These preparations increase the accuracy and viability of the model.

2.6 Challenges and Limitations

Predictive modelling for supply chain management faces numerous challenges that limit its effectiveness. First, data quality concerns may be seen through inconsistencies, incompleteness, and inaccuracy present within the datasets utilized for supply chain management. Second, there are several integration problems, which are caused by mixed sources and supply chain partners hosting their data in different IT systems. Finally, the problem of model interpretability is related to the high complexity of models produced with the use of ML models. As a result, the adoption and application of such models are affected, which can also be attributed to the limited scalability of the developed solutions. These problems need to address to enhance the quality of data, improve its integration, ensure the use of simple interpretable models, and develop scalable models specific to the supply chain.

2.7 Summary and Synthesis of Literature

The findings presented in the literature review highlight the value of using machine learning for the optimization of supply chain performance, especially in the prediction of late deliveries (sections 2.2 & 2.5). They demonstrate the critical influence of theoretical frameworks such as the Supply Chain Network Theory and the RBV in SCM, outlining the importance of understanding dependencies and utilizing internal competencies to achieve positive results. The challenges associated with data quality issues, inconsistent integration, and a lack of interpretability still need to be addressed. As a result, it seems that the integration of the most advanced machine learning techniques with supply chain management practices is likely to lead to a notable improvement in predictive accuracy and operational effectiveness. The achievements that can be derived through such a strategy would assist companies in relying on their networks with confidence, addressing weaknesses, and better managing volatility and disruptions.

3 Methodology

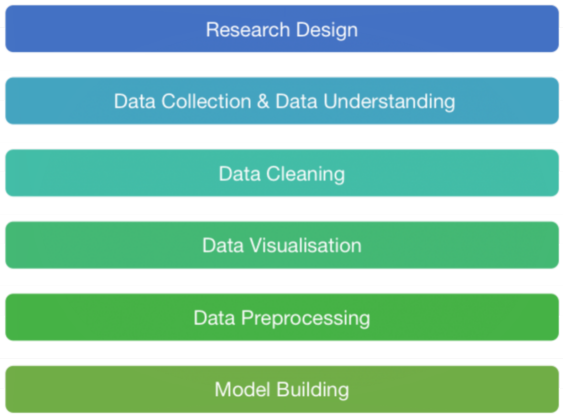

Figure 1: Methodology Structure

3.1 Research Design

The combined research design was employed in this paper aiming to deploy both the qualitative and quantitative approaches for optimizing supply chain performance to predict late deliveries and sales behavior. Therefore, the literature review was the first phase conducted to determine the major factors that impact delivery performance and sources that describe and explain the best ML in predicting late delivery and prediction of sales behaviour. The finding of the first step developed the conceptual framework for empirical analysis Lolla et al. (2022). The data collection process involved gathering historical supply chain data and considering the diverse sources to generate comprehensive data that can be studied. One of the major steps was technologies of feature engineering which allowed extracting significant features that truly may influence the delivery Tordecilla et al. (2021). Finally, models were created based on and preprocessing data, including supervised algorithms for learning such as logistic regression, random forest, decision trees and regression models for sales prediction. Assessment of models' performance was carried out with common metrics involves accuracy, precision, recall, F1-score and MSE, RMSE and MAE.

3.2 Data Collection & Data Understanding

The data for this study was taken from the Kaggle data source which is "DataCo Smart Supply Chain For Big Data Analytics" dataset [1]. The dataset consisted of extensive records regarding various supply chain operations, from order processing and shipment tracking to customer demographics. The dataset was chosen due to its comprehensive nature, containing relevant data with regard to supply chain performance metrics, which makes it suitable for predicting late deliveries and sales. The data has 180519 rows and 53 columns with relevant features which will help understand the supply chain management process.

3.3 Data Cleaning

The first step in data cleaning includes checking for null values in the data. It was identified that two columns, Product Description and Order Zipcode, had a lot of missing values, with one of them being completely blank. Since these columns were not useful, they were removed from the data. Additionally, two other columns, Customer Lname and Zipcode, had only a few missing values (8 and 3, respectively). To handle this, the missing values in these columns were also removed, ensuring the data was clean and prepared for further analysis. After this, dropped several columns which were not directly related to the task. These columns included 'Customer Email', 'Product Card Id', 'Order Item Cardprod Id'and many others. The next was to check for duplicate values also to handle these it is removed from the dataset. This step ensured that the data was streamlined for analysis, containing only relevant features for the machine learning models.

3.4 Data visualisation

The data analysis focuses on detecting patterns, trends, and relationships concerning the supply chain data for better understanding. Different types of visualisation are performed

[1] https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis/data

to get an insight from the data.

The Pie chart in Figure 2 for Late Delivery Risk displays the count of two categories of delivery risk, the first one is '0' meaning that the firm has no risk and the second one is '1' meaning that the firm is at risk. According to the pie chart, risk '0' in light blue, is relatively shorter than risk '1' in light orange.

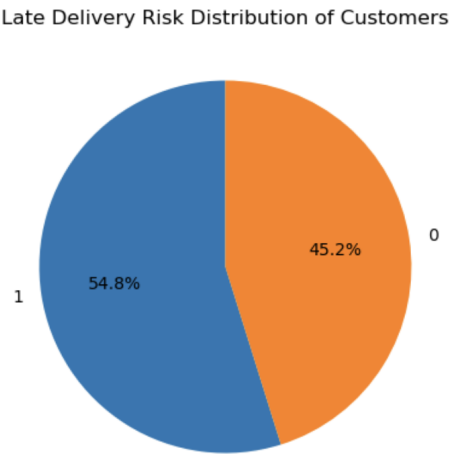

Figure 2: Pie Chart for Late Delivery Risk where '1' represents a risk of late delivery & '0' represents no risk of late delivery

Stacked Bar Chart in Figure 3 for Order Region vs Late Delivery Risk shows the 'Late Delivery Risk by Order Region" plot using a bar plot. It was used to visualize the relationship between 'Order Region' and 'Late delivery risk'. The chart was created using a cross-tabulation of these two variables, depicting the count of late delivery risks across different order regions. By stacking the bars, the chart illustrates the proportion of each risk level within each region, providing insights into regional patterns and risk distribution. This visualization helps identify regions with higher late delivery risks, aiding in targeted interventions to improve supply chain performance.

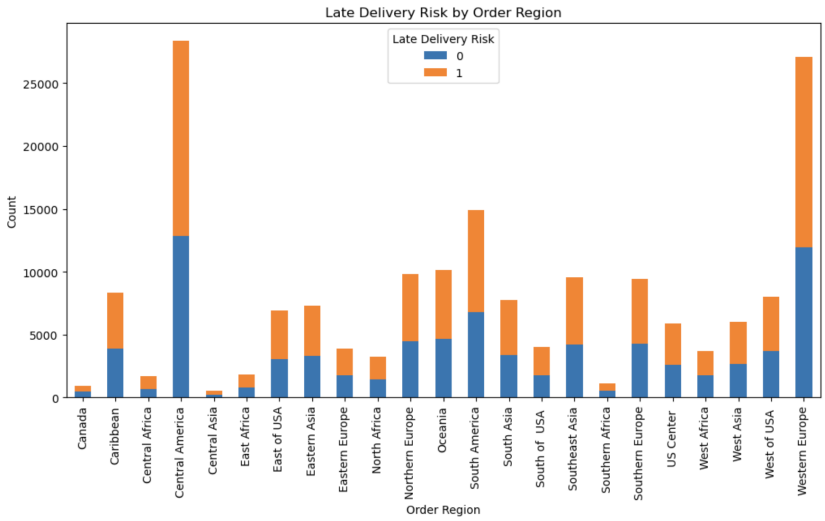

Figure 3: Stacked Bar Chart for Order Region vs Late Delivery Risk

The Sankey chart Figure 4 display the flow of deliveries all over the regions, shipping methods, and risk of late delivery. The regions are shown on the left side, and these regions are connected with shipping methods, while the right side shows late delivery. The number of deliveries made using each delivery method is displayed by the lines between these categories. The thickness of the lines shows how often deliveries fit into each category relative to one another, giving a clear picture of how various transportation methods and geographical locations affect the possibility of late delivery.

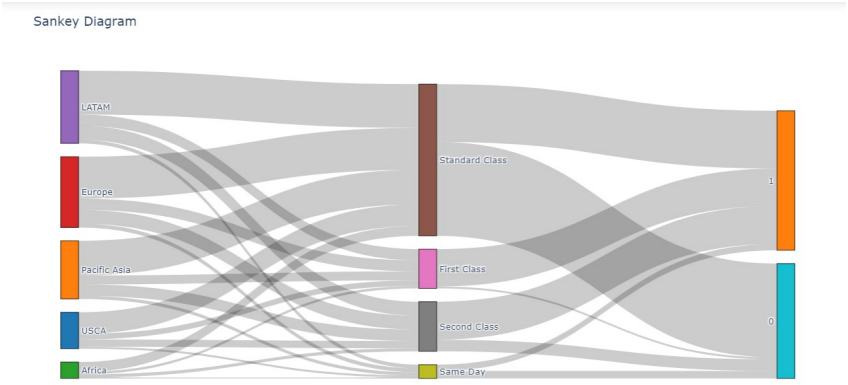

Figure 4: Sankey Chart of Order Region, Shipping Mode & Late Delivery Risk

A bar chart Figure 5 was utilized to display the distribution of orders from different categories. The data was grouped by 'Category Name' and the count of orders in each category was counted. The resulting bar chart, with categories on the x-axis and the number of orders on the y-axis, was colour-coded based on order counts. This visualization helps in understanding which categories have higher sales volumes, providing valuable insights into sales behavior.

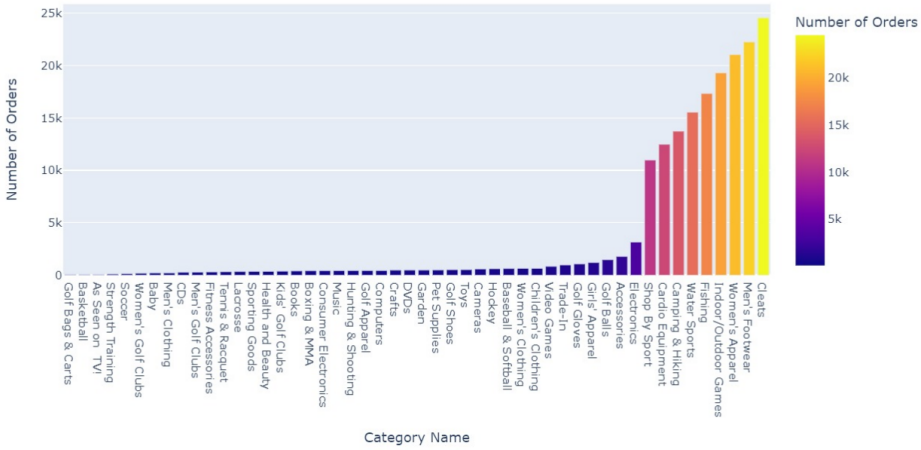

Figure 5: Sales Distribution by Product Category & Number of Orders

The correlation heatmap Figure 6 shows the relationship between the variables visually in the data. The red colours show a strong positive correlation and the blue colours show a negative correlation. The variables 'sales' and 'Order Item Total' are highly correlated with each other. On the other hand, 'Latitude' and 'Longitude' show negative correlations with other variables. The heatmap is useful for finding relationships between features within the data which helps in feature selection and multicollinearity.

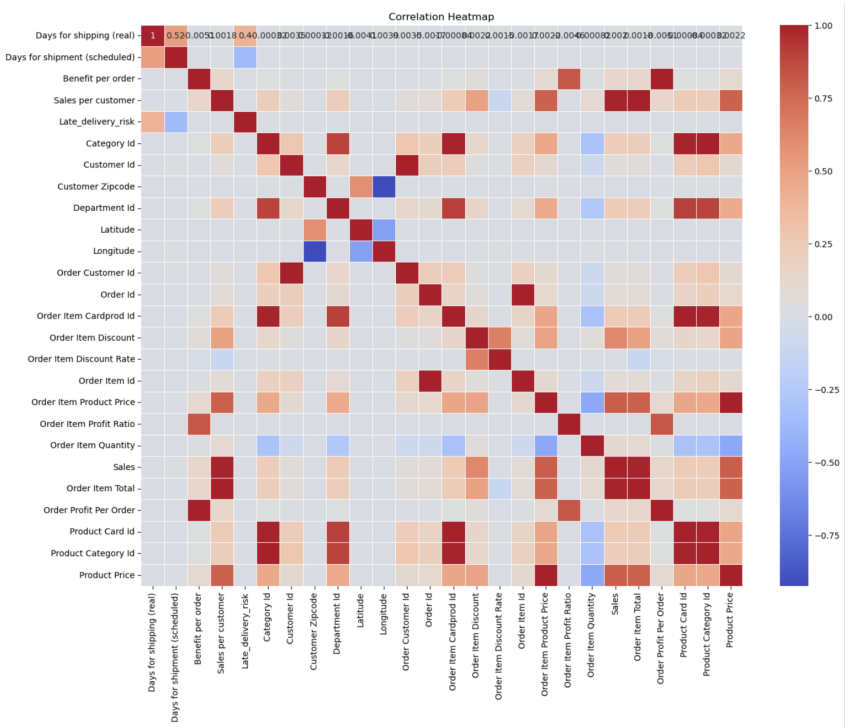

Figure 6: Heatmap showing correlation of target variable with independent variables in the dataset.

3.5 Data Preprocessing

The first step in preprocessing is to drop irrelevant columns after visualising the data and it is identified that some of the columns had the same values in different columns which is considered as duplicate columns which were removed from the dataset.

3.5.1 Feature Engineering

In this section, extracting new columns, categorical encoding, and label encoding are performed to prepare the data for modelling. Several new columns were created to enhance the predictive abilities of the dataset from the date columns. The order year, order month, order day, shipping year, shipping month, and shipping day were extracted from the order date and shipping date. 'order date (DateOrders)' and 'shipping date (DateOrders)' two columns were removed to produce a much cleaner dataset designed to predict delivery time intervals.

To begin with, the process of categorical encoding that used pandas' get dummies function has significantly increased the size of the data frame. It also helped convert each categorical variable into a set of binary indicators required by a machine learning model to interpret the information. There are a lot of categorical columns that need to be encoded to perform modelling. Thus, no ordinal assumptions were made about categories; instead, all of them were included in the analysis to allow a richer evaluation of the input variables Potdar et al. (2017).

The label encoding for ordinal columns (ordinal cols) in the data frame is also implemented. It works in such a way that the label encoding process generates a LabelEncoder instance (le) for each ordinal column, which in turn transforms its categorical values into numeric labels. The latter part is essential for utilizing the ordinal columns in numerical format, where each category in the original data receives a unique integer number Ganesh and Kalpana (2022).

3.6 Modelling

Different machine learning models were used in the study to predict late deliveries and sales. Models including Random Forest, Decision Trees and Logistics Regression were chosen for predicting late delivery since they are strong when working with data containing complex relationships between inputs. For predicting sales, regression models were performed such as Linear Regression, Decision Tree Regression and Lasso Regression. For modelling, the data is split into train and test sets which are used to prepare data for modelling in machine learning. Then, the target variable is defined. The test size of 0.3 means that approximately 30% data will be reserved for the test set, while the model is trained with the remaining 70% of data.

4 Design Specification

4.1 Architectural Framework

This section presents the design specifications for the application of ML models. It includes a detailed of different type of models which were used in this paper. The architecture of the models is an essential framework integrated into the design of these tools for supporting the e-commerce supply chain. Fundamentally, the framework encompasses

several major components that define the operation of the models, that is, data preprocessing, feature engineering, and ML algorithms. The process of data preprocessing is an initial task that is associated with cleaning and transforming raw data to be able to analyse it Stranieri and Stella (2022).

4.2 Choosing Machine Learning Models for Late Delivery Prediction

To examines the effectiveness of ML models in managing and predicting late delivery risks, three distinct models were selected: Random Forest, Decision Tree, and Logistic Regression. The Random Forest model was chosen due to its robustness and capacity to handle complex data effectively.Model leverages an ensemble of decision trees to enhance predictive accuracy and manage a wide range of input features. Its ability to aggregate the predictions from multiple trees helps mitigate the risk of over fitting and provides a comprehensive analysis of late delivery risks.

The Decision Tree model was included for its high interpretability. This model provides a clear, visual representation of decision-making processes and is useful for understanding the factors influencing late deliveries. Despite its simplicity, Decision Trees can deliver reliable predictions and offer insights into the data structure and decision rules.

Logistic Regression was also employed to provide a comparative perspective. Although it is a simpler model compared to Random Forest and Decision Tree, Logistic Regression allows for an evaluation of how well a linear approach can perform in predicting late delivery risks. Including this model ensures a thorough assessment of predictive capabilities, offering a baseline for comparison with more complex algorithms Modgil et al. (2022).

4.3 Choosing Machine Learning Models for Sales Prediction

To investigate how machine learning algorithms can enhance sales behavior prediction in supply chain management, three models were employed: Linear Regression, Decision Tree Regression, and Lasso Regression. Linear Regression was utilized as a baseline model to establish a foundational performance benchmark. Its straightforward approach assesses the linear relations between features and sales behavior, providing a clear measure of basic predictive capability. Decision Tree Regression was selected for its ability to model complex, non-linear relations in the dataset. This model can capture intricate patterns and interactions between features, offering high accuracy and detailed insights into the factors influencing sales behavior. Its flexibility in handling various data types makes it particularly suitable for understanding and predicting complex sales patterns. Lasso Regression was included to address feature selection and regularization Nagar et al. (2021).

5 Implementation

In the implementation section, the study includes the model development, tuning and description of various models to address both the research question which is predicting late delivery risks and enhancing sales prediction in supply chain management.

5.1 Data Preparation and Splitting

The initial step in the implementation was to prepare the data for the modelling. The data was cleaned and preprocessed to make sure it gave accurate results. First, the feature selection step was done to keep relevant features from the dataset for the modelling stage as it was a crucial step. Second, the dataset was split into train and test sets using a 70-30 split, ensuring the models were trained on a substantial section of the data while reserving a portion for evaluation.

5.2 Model Development for Research Question 1

The research question 1 focuses on predicting the risk of late delivery using a ML model. Three classification models were applied for prediction.

5.2.1 Random Forest Model

The Random Forest model, employing 100 decision trees, was particularly effective due to its ensemble approach, that aggregates predictions from the multiple trees to improve accuracy and handle complex data. This model demonstrated superior performance across various metrics, including precision, recall, and F1 score, making it highly efficient for predicting late deliveries.

5.2.2 Decision Tree Model

The Decision Tree model, known for its simplicity and interpretability, also provided reliable results, though it is more prone to overfitting with complex datasets. The model is simple to understand and for gaining insights into the variables which cause late delivery. The Decision Tree classifier was trained on synthetic data with 1000 samples with a max depth of 3.

5.2.3 Logistic Regression Model

Logistic Regression was implemented because of its efficiency in binary classification which makes it an ideal choice for predicting late delivery. The maximum iteration is set to 1000 to give the model enough iterations to converge and ensure robust performance. The logistic Regression classifier was instantiated and trained on the synthetic data that was split into train and test sets. After extracting the data, the classifier was fit with the training data and tested on the test data.

5.3 Hyperparameter Tuning and Feature Importance

After applying all the models for late delivery prediction, the hyperparameter tuning was applied for better performance of each model. The 2 types of techniques were used in this method such as Grid search and Random Search to analyze various combinations of hyperparameters. In this process, each model's parameters were modified to get the best results possible. The tuning process helps to increase the accuracy and boost the performance of each of the models Probst et al. (2019). After the tuning of the model, feature importance analysis was performed on each of the models for research question 1. This method helps to find the top 10 features that are most influenced the late delivery prediction. It also helps to understand these key features which give useful insights into

which columns have the most impact in identifying delivery results and giving practical guidance for risk reduction Wang and Ni (2019).

5.4 Model Development for Research Question 2

The research question 2 aimed to enhance the prediction of sales behaviour in SCM. Three regression models were employed for prediction. Before the modelling, the standardised method is used on features of the dataset. This process helps to standardize each column to make sure that each feature contributes equally to the model which improves the performance of the model.

5.4.1 Linear Regression Model

The Linear Regression model was applied to predict sales behaviour and the model is known for its interpretability and simplicity which make it a strong choice for understanding the relation between the target variable and independent variable. The performance is evaluated from various metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The feature importance method is also applied to get the top 10 features which influence the most prediction Sharma et al. (2022).

5.4.2 Decision Tree Model

The second model which is the Decision Tree Regression model is applied making it highly effective despite its susceptibility to overfitting. This model is very good at predicting complex sales behaviors because it improves at obtaining non-linear relationships in the data. The performance is evaluated from the same metrics which are given in the above model and feature importance is also applied in this model.

5.4.3 Lasso Regression Model

The third is the lasso regression model which is applied for sale prediction. This model is used for its ability to perform regularization and variable selection. It was useful for find the most relevant predictors. The metrics are the same and also feature importance method is applied to recognize the most impacted features for the model.

6 Evaluation

The evaluation section includes a comprehensive analysis of the results and the findings from the models which were employed.

6.1 Experiment 1: Initial Model Application & Results for Research Question 1

The first case study focuses on assessing the effectiveness of different ML models applied for the prediction of late deliveries. The Random Forest classifier got an accuracy of 97.39%, the Decision Tree Classifier achieved an accuracy of 93.67% even though it's the

simplest model but it gives reasonable accuracy and the Logistic Regression model gives an accuracy of 97.58%. The below table provides the other metrics results.

Model	Precision	Recall	F1 Score	True Positives (TP)	True Negatives (TN)	False Positives (FP)	False Negatives (FN)
Random Forest Classifier	97.47%	97.39%	97.38%	29,423	23,314	1,300	116
Decision Tree Classifier	94.33%	93.67%	**93.61**%	29,538	21,187	3,427	1
Logistic Regression	**97.68**%	**97.58**%	97.57%	29,539	23,301	1,313	0

Table 1: Performance metrics for different classifiers

The classifier comparison has shown that all models perform highly efficiently. By means of an ensemble approach for managing complex data, the Logistic Regression model delivered the highest test set accuracy of 97.58% and excelled in both precision and F1 scores. The model has also shown the highest improvement in its efficiency in reducing overfitting. The Decision Tree model has a lower accuracy of 93.67%, having high interpretability but was more likely to give false positive results. Although the accuracy of the Random Forest Classifier does not lag far behind that of Logistic Regression, with an accuracy of 97.39%, the model showed strong performance but it had more false positives. Hyperparameter tuning could be used to enhance the accuracy and overall performance of these models.

6.2 Experiment 2: Hyperparameter Tuning & Feature Importance for Research Question 1

In this case study, hyperparameter tuning was applied to all the models to enhance the performance after the implementation of the initial models. This method provides fine-tuning parameters which helps to get better accuracy and other metrics. The Random Forest classifier gets an accuracy of 97.58%, the Decision Tree Classifier achieved an accuracy of 93.53% and the Logistic Regression model gives an accuracy of 97.58%. The below table provides the other metrics results.

Model	Precision (%)	Recall (%)	F1 Score (%)	True Positives	True Negatives	False Positives	False Negatives	Cross-Validation Score (%)
Random Forest	97.68	97.58	97.57	29,539	23,301	1,313	0	97.54
Decision Tree	97.63	97.53	97.52	29,509	23,307	1,307	30	97.50
Logistic Regression	97.68	97.58	97.57	29,539	23,301	1,313	0	97.54

Table 2: Performance metrics after Hyperparameter Tuning

The model performance has improved by applying hyperparameter tuning, with Logistic Regression and Random Forest getting an accuracy score of 97.58%. The Decision Tree model achieved close with 97.53% accuracy. The confusion matrices depict that logistic regression and random forest had no false negatives whereas the decision tree had a less as compared to others. High precision, recall, and F1 scores were shown by all models, indicating reliable performance. The best cross-validation scores were 97.50% for the decision tree and 97.54% for random forest and logistic regression which shows their strong performance.

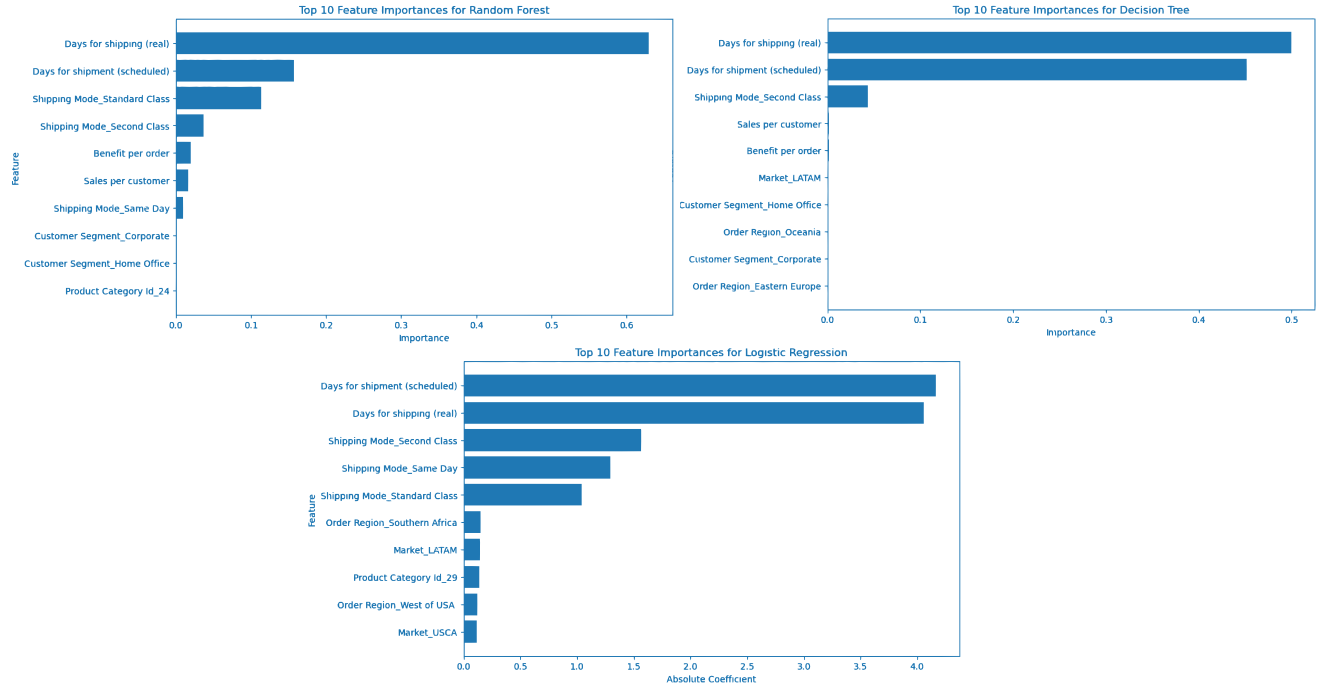

Figure 7: Feature Importance of Research Question 1

The feature importance method in Figure 7 was applied to each of the models to identify the top 10 features that were most influential in the dataset. Variables like "Days for shipping (real)" & "Days for shipment (scheduled)" were found to be the main factors in the prediction for the decision tree and random forest classifier. For logistic regression, the variable 'Days for shipping (real)' shows a positive impact, whereas "Days for shipment (scheduled)" has a negative coefficient. Both the model decision tree and random forest show the importance of other factors, like customer segments, regions, and shipping modes. This method helps to understand how different models can have different feature values, even when the outcome of the prediction is the same.

6.3 Experiment 3: Model Application and Results for Research question 2

This case study aimed to research question 2 in which three types of regression models are applied to predict sales such as Linear Regression, Decision Tree Regression and Lasso Regression. The models were evaluated using machine learning metrics such as MSE, RMSE, MAE and R^2. The Linear Regression got R^2 of 0.98, the Decision Tree Regression achieved R^2 of 0.99 and the Lasso Regression model gave R^2 of 0.97. The feature importance method is also employed to find the most impacted factors in predicting sales.

Model	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)
Linear Regression	334.45	18.29	13.56
Decision Tree Regression	51.21	7.16	0.97
Lasso Regression	348.44	18.67	13.75

Table 3: Performance metrics of different regression models

The Linear Regression Model depicts a high MSE and RMSE which shows a significant amount of prediction error. Also, the MAE is comparatively large which showcases that

the model's predictions usually differ from the actual values. This may suggest that the complexity of the data makes it challenging for Linear Regression to handle. The metrics of the decision tree regression model show lower MSE and RMSE as compared to Lasso and Linear Regression which indicates better accuracy and performance. The MAE is significantly lower which means that actual values are more closely in line with one another. This displays how well the model was able to represent non-linear relationships in the data. The Lasso Regression model gives similar results to linear regression on the basis of MSE and RMSE metrics. The MAE 13.75 which is larger shows it does not improve prediction accuracy as compared to linear regression.

Top 10 features for Linear Regression:

	Feature	Coefficient	Absolute Coefficient
7	Market_LATAM	-5.515629e+12	5.515629e+12
9	Market_USCA	-5.508364e+12	5.508364e+12
6	Market_Europe	-5.190127e+12	5.190127e+12
20	Order Region_Oceania	-2.374935e+12	2.374935e+12
24	Order Region_Southeast Asia	-2.306226e+12	2.306226e+12
8	Market_Pacific Asia	-2.279191e+12	2.279191e+12
28	Order Region_West Africa	-2.228873e+12	2.228873e+12
22	Order Region_South Asia	-2.087429e+12	2.087429e+12
18	Order Region_North Africa	-2.087008e+12	2.087008e+12
16	Order Region_Eastern Asia	-2.028261e+12	2.028261e+12

Top 10 features for Decision Tree Regressor:

	Feature	Importance
3	Sales per customer	0.990640
59	Product Category Id_43	0.001040
43	Product Category Id_17	0.001012
33	Product Category Id_4	0.000986
45	Product Category Id_24	0.000552
62	Product Category Id_46	0.000545
37	Product Category Id_9	0.000505
2	Benefit per order	0.000486
68	Product Category Id_63	0.000462
67	Product Category Id_62	0.000431

Top features for Lasso Regression with adjusted alpha:

	Feature	Coefficient	Absolute Coefficient
3	Sales per customer	124.431103	124.431103
61	Product Category Id_45	4.913475	4.913475
69	Product Category Id_64	4.462508	4.462508
59	Product Category Id_43	2.587412	2.587412
37	Product Category Id_9	2.400783	2.400783
73	Product Category Id_68	1.195465	1.195465
67	Product Category Id_62	1.080230	1.080230
71	Product Category Id_66	0.955672	0.955672
63	Product Category Id_48	0.799570	0.799570
68	Product Category Id_63	0.742840	0.742840

Figure 8: Feature Importance of Research Question 2

The feature importance analysis in Figure 8 shows that 'Sales per Customer' is the most influential feature with a 123.43 in both the Lasso regression and decision tree regression models. It was the most relevant feature with an importance score of 0.989 in Decision Tree Regression, but it had a significant positive coefficient in Lasso Regression. The decision tree feature importance suggests columns like Product Category Id 10" and "Product Category Id 43 play a part in determining the outcome even though they contribute less. The significance of various market regions was shown using linear regression, with features like "Market LATAM" and "Market USCA" showcasing big negative coefficients, indicating an important but inverse effect on the predictions.

6.4 Discussion

This section discusses the findings from the evaluations in relation to the literature reviewed and assesses the extent to which the research questions have been addressed. The literature review highlighted the potential of ML models in enhancing SCM, emphasizing the importance of predictive accuracy and model complexity. It identified various model's strengths in predicting late deliveries and predicting sales behaviour.

The evaluation has shown that the Random Forest classifier is the best fit for predicting late delivery risks among the models experimented within this work, as it was distinguished by the test set accuracy of 97.58% and the strong values of precision, recall, and F1 scores. It proves the efficiency of ensemble methods when working with more complicated input data and accounting for the reduction of overfitting. The Decision

Tree classifier, while slightly less accurate, should be considered useful due to the high interpretability of its approaches and contribution to understanding decision boundaries. As for Logistic Regression, the results are below the target levels, but they explain the challenge of attaining high predictive quality with the help of simpler models in high-dimensional datasets.

R^2 score of the Decision Tree Regression model is 0.9967. This proves near-perfect precision. It is evident that the Decision Tree model captures the complex relationships in the data, so it can be declared as the best-performing method. Respective performance metrics provided, Linear Regression and Lasso Regression present quite high accuracy. It demonstrates that models could be selected based on the relevant features of the datasets and specific requirements for prediction.

Overall, the findings validate the application of ML techniques to supply chain management, demonstrating their potential to significantly enhance predictive accuracy and decision-making processes. The research aligns well with existing literature, reinforcing the effectiveness of advanced machine learning models in complex and high-dimensional environments.

7 Conclusion and Future Work

Although the stated research aims of optimizing eCommerce supply chain performance were accomplished. The study successfully met all the goals of constructing and deploying the ML models for late delivery risk and sales behaviour. Therefore, the most effective model for mitigating delivery risks and enhancing supply chain reliability was introduced with the purpose of answering Research Question 1. Specifically, it was found that the Random Forest classifier yielded the highest accuracy at 97.58% and performed excellently on numerous tests, making it an outstanding tool for avoiding the specified risks.

In response to Research Question 2, which aimed to enhance the prediction of sales behaviour, the Decision Tree Regression model proved to be the best choice. It delivered outstanding accuracy with a high R^2 score and minimal error, showcasing its effectiveness in forecasting sales trends. These findings highlight the substantial benefits of applying advanced ML techniques to eCommerce supply chains.

This study provides valuable insights into predicting the risk of late delivery and sales using different machine learning models, but future research could explore several areas. This means evaluating model performance on higher dimensions and more complex data can help avoid overfitting. Also, using deep learning algorithms like the neural network could also be a good idea due to the enhanced accuracy of the model developed because they can simulate complex relationships and patterns. By improving the interpretability of complex models like Random Forests will benefit supply chain managers, and incorporating external data like weather and economic indices could refine predictions.

References

Abideen, A. Z., Sundram, V. P. K., Pyeman, J., Othman, A. K. and Sorooshian, S. (2021). Digital twin integrated reinforced learning in supply chain and logistics, *Logistics* **5**(4): 84.

Baryannis, G., Dani, S. and Antoniou, G. (2019). Predicting supply chain risks us-

ing machine learning: The trade-off between performance and interpretability, *Future Generation Computer Systems* **101**: 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): 3330–3341.

Feizabadi, J. (2022). Machine learning demand forecasting and supply chain performance, *International Journal of Logistics Research and Applications* **25**(2): 119–142.

Feng, Y. and Zhao, H. (2023). Multi-supply chains optimization mechanism based on machine learning and double auctions, *Fractals* **31**(6).

Ganesh, A. D. and Kalpana, P. (2022). Future of artificial intelligence and its influence on supply chain risk management–a systematic review, *Computers & Industrial Engineering* **169**: 108206.

Gayam, S. R., Yellu, R. R. and Thuniki, P. (2021). Optimizing supply chain management through artificial intelligence: Techniques for predictive maintenance, demand forecasting, and inventory optimization, *Journal of AI-Assisted Scientific Discovery* **1**(1): 129–144.

Ghazal, T. M. and Alzoubi, H. M. (2021). Modelling supply chain information collaboration empowered with machine learning technique, *Intelligent Automation & Soft Computing* **29**(3): 243–257.

Helo, P. and Hao, Y. (2022). Artificial intelligence in operations management and supply chain management: An exploratory case study, *Production Planning & Control* **33**(16): 1573–1590.

Janinhoff, L., Klein, R. and Scholz, D. (2024). Multitrip vehicle routing with delivery options: A data-driven application to the parcel industry, *OR Spectrum* **46**(2): 241–294.

Khan, M. A., Saqib, S., Alyas, T., Rehman, A. U., Saeed, Y., Zeb, A., Zareei, M. and Mohamed, E. M. (2020). Effective demand forecasting model using business intelligence empowered with machine learning, *IEEE Access* **8**: 116013–116023.

Lolla, R., Harper, M., Lunn, J., Mustafina, J., Assi, J., Loy, C. K. and Al-Jumeily OBE, D. (2022). Machine learning techniques for predicting risks of late delivery, *The International Conference on Data Science and Emerging Technologies*, Springer Nature Singapore, Singapore, pp. 343–356.

Mehrotra, P., Fu, M., Huang, J., Mahabhashyam, S. R., Liu, M., Yang, M., Wang, X., Hendricks, J., Moola, R., Morland, D. and Krozier, K. (2024). Optimizing walmart's supply chain from strategy to execution, *INFORMS Journal on Applied Analytics* **54**(1): 5–19.

Modgil, S., Singh, R. K. and Hannibal, C. (2022). Artificial intelligence for supply chain resilience: learning from covid-19, *The International Journal of Logistics Management* **33**(4): 1246–1268.

Nagar, D., Raghav, S., Bhardwaj, A., Kumar, R., Singh, P. L. and Sindhwani, R. (2021). Machine learning: Best way to sustain the supply chain in the era of industry 4.0, *Materials Today: Proceedings* **47**: 3676–3682.

Potdar, K., S., T. and D., C. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers, *International Journal of Computer Applications* **175**(4): 7–9.

Probst, P., Wright, M. N. and Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **9**(3).

Rai, R., Tiwari, M. K., Ivanov, D. and Dolgui, A. (2021). Machine learning in manufacturing and industry 4.0 applications, *International Journal of Production Research* **59**(16): 4773–4778.

Sharma, R., Shishodia, A., Gunasekaran, A., Min, H. and Munim, Z. H. (2022). The role of artificial intelligence in supply chain management: mapping the territory, *International Journal of Production Research* **60**(24): 7527–7550.

Shibin, K. T., Dubey, R., Gunasekaran, A., Hazen, B., Roubaud, D., Gupta, S. and Foropon, C. (2020). Examining sustainable supply chain management of smes using resource based view and institutional theory, *Annals of Operations Research* **290**: 301–326.

Stranieri, F. and Stella, F. (2022). A deep reinforcement learning approach to supply chain inventory management, *arXiv preprint* **arXiv:2204.09603**.

Tirkolaee, E. B., Sadeghi, S., Mooseloo, F. M., Vandchali, H. R. and Aeini, S. (2021). Application of machine learning in supply chain management: a comprehensive overview of the main areas, *Mathematical Problems in Engineering* **2021**(1): 1476043.

Tordecilla, R. D., Juan, A. A., Montoya-Torres, J. R., Quintero-Araujo, C. L. and Panadero, J. (2021). Simulation-optimization methods for designing and assessing resilient supply chain networks under uncertainty scenarios: A review, *Simulation Modelling Practice and Theory* **106**: 102166.

Wallmann, C. and Gerschberger, M. (2021). The association between network centrality measures and supply chain performance: The case of distribution networks, *Procedia Computer Science* **180**: 172–179.

Wang, Y. and Ni, X. S. (2019). A xgboost risk model via feature selection and bayesian hyper-parameter optimization, *International Journal of Database Management Systems* **11**(01): 01–17.