International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Descriptive and Predictive Analytics for Supply Chain Data with Machine Learning Technique

Yahya Oussoulous[1], Sofianita Mutalib*[2], Kamalia Azma Kamaruddin[3], Norsariah Abdul Rahman[4], Irwan Ibrahim[5], Ja'afar Pyeman[6]

Submitted:10/03/2024 **Revised**: 25/04/2024 **Accepted**: 02/05/2024

Abstract: This study investigates how supply chain analytics, or SCA, can be used to examine past supply chain performance in the Asia Pacific region. The study assesses different supply chain elements, such as client demographics, product categories, payment options, and transportation, using historical sales data. The dataset is observed through descriptive analysis with scatter plots, box-plots, skewness analysis and bar chart. Regression-based machine learning models, such as linear regression, random forest and boosting were utilised to forecast demand (order item total), allowing businesses to analyse their logistical and production processes. The demand was well forecasted by Random Forest, and the danger of late delivery was well-predicted by ensemble learning models, according to the results. The study comes to the conclusion that, in order for businesses to increase productivity and customer happiness, both descriptive and predictive approaches are essential. While predictive analytics enables proactive decision-making and risk avoidance, descriptive analytics offers a thorough understanding of supply chain operations. The study contributes to the supply chain analytics field in the Asia Pacific region by setting a theoretical framework that allows organisations to adopt proactive strategies to optimize their performance based on different supply chain components. However, the paper has several limitations related to data collection, as it does not involve many companies and manufacturers. We need further research to enhance prediction performance and optimize the learning process.

Keywords: *Supply chain analytics, demand forecasting, machine learning, supervised methods.*

1. Introduction

As an industrial hub, the emerging market of Asia Pacific has witnessed a remarkable transformation in its market landscape, characterized by rapid economic growth and urbanization, resulting in one of the most promoting markets, attracting customers from all over the world. Thus, the region has seen a substantial increase in the export of goods and services, reaching a value of 9.66 trillion US dollars in 2022 [1]. This growth has also led to the development of a complex supply chain and logistics system. However, the dynamic and uncertain international business environment has introduced tremendous disruptions to the global supply chain. Supply chain disruptions arise from a "combination of an unintended and unexpected triggering event that occurs somewhere in the upstream supply chain (the supply network), the inbound logistics network, or the purchasing (sourcing) environment, and a consequential situation, which presents a serious threat to the normal course of business operations of the focal firm" [2]. Nevertheless, the COVID-19 pandemic highlighted that unexpected and rare events which cause supply chain disruptions are not "black swans" [3]. In order to address these challenges, supply chain analytics (SCA) has emerged as an effective tool for predicting and managing disruptions in the supply chain. SCA allows organizations to measure and improve their supply chain performance by identifying the root causes of disruptions and making informed business decisions [4].

Analytics, the process of extracting meaningful insights from data, has been empowered by the advance of Machine Learning (ML) techniques which have demonstrated remarkable capabilities in complex decision-making tasks. According to Hackett Group [5] survey, 66% of supply chain managers believe that analytics capabilities are crucial to their business operations. However, recent review articles have identified a predominance of descriptive analytics rather than predictive analytics [6]. Regarding the increasing complexity and uncertainty of supply chains, the ability to predict disruptions before they occur is crucial. As well as the important role of the planning phase where SCA can assist top management in decision-making about supply chain operations, which frequently comprises demand planning, procurement, production, inventory, and logistics

[1]*School of Industrial Management, Mohammed VI Polytechnic University, MOROCCO*
ORCID ID : 0009-0004-2803-5499
[2]*School of Computing Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, MALAYSIA*
ORCID ID : 0000-0001-8384-3131
** Corresponding Author Email: sofianita@uitm.edu.my*
[3]*School of Computing Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, MALAYSIA, ORCID ID : 0000-0001-5860-5182*
[4]*Procurement Division, Pertubuhan Keselamatan Sosial, Menara PERKESO, 281 Jalan Ampang, 50538 Kuala Lumpur, MALAYSIA*
ORCID ID : 0009-0005-6485-6304
[5]*Faculty of Business and Management, Universiti Teknologi MARA Cawangan Selangor, 42300 Puncak Alam, Selangor, MALAYSIA*
ORCID ID: 0000-0002-0887-2394
[6]*Faculty of Business and Management, Universiti Teknologi MARA Cawangan Selangor, 42300 Puncak Alam, Selangor, MALAYSIA*
ORCID ID: 0000-0001-8866-3244

[7]. SCA plays a vital role in lean, agile, resilient, and sustainable supply chains by reducing production lead time and transportation time, minimizing uncertainties and risks, and enhancing integration [8]. Despite the various benefits of SCA and the vivid interest among practitioners in adopting SCA, the research in this area is limited [9][10].

This paper aims to investigate the application of data analytics in the supply chain using ML models to forecast demand. This is by analyzing the flow of goods from the Asia Pacific market using descriptive analytics first, and then predicting the demand for the Asia Pacific goods from customers from the USA and Puerto Rico which raises questions about the delivery time regarding the distance between these markets and other factors that might generate many disruptions. Thus, it is necessary to predict late delivery risk as well. By doing so, the study contributes to the field of SCA by showcasing the application of predictive data analytics within supply chains. It also contributes to the field of supply chain risk management by presenting a set of methods that hold the potential to predict supply disruptions. The remainder of the paper is structured as follows: Section 2 provides a literature review of related works on supply chain and data analytics. Section 3 presents materials and method. The results of the study are discussed in Section 4. Finally, Section 5 concludes the paper.

2. Related works

Big Data Analytics (BDA) pertains to utilizing sophisticated analytical techniques such as predictive methods, statistics, data mining, and Artificial Intelligence (AI) on extremely large and unstructured datasets [11]. BDA encompasses two perspectives: the first involves big data (BD), and the second involves business analytics (BA). BD refers to high-volume, high-velocity, and high-variety sets of dynamic data that surpass the capabilities of traditional data management methods [13]. On the other hand, BA involves the study of skills, technologies, and methodologies employed to evaluate organization-wide strategies and operations continuously to obtain insights and guide the business planning of an organization. Such assessments span from strategic management and product development to customer service, utilizing evidence-based data, statistical analysis, operational scrutiny, predictive modeling, forecasting, and optimization techniques [14].

Predictive analytics is a group of methods that uses statistical and other empirical techniques to predict future events, based on past occurrences. Previous studies highlighted the importance of the leverage of ML models to gain insights from data and predict future events. The reviews conducted by several researchers find that although the main data analytics methods currently employed are statistics, simulation, optimization, the big promise of big data is in machine learning [15][16][17]. Furthermore, Zhong et al. [18] emphasized that machine learning should be embedded in decision models so they can have continuous learning capability.

The capabilities of BDA can offer support to various supply chain (SC) functions, encompassing procurement, warehousing, manufacturing, demand management, transportation/logistics, and overall SC operations [19]. The current literature highlights additional research papers focused on manufacturing and transportation/logistics [20]. Critical activities in SC manufacturing involve production planning and control (PPC), research and development of products, maintenance and diagnostics, and quality management. The integration of BDA into PPC is attracting interest from many researchers, and BDA tools and techniques in this area are relatively mature [21].

SCA has been described in two main ways: as a collection of abilities such as management, talent, and technology, and as a range of qualitative and quantitative tools, techniques, and approaches [22][23][24]. Mubarik et al. [4] (2019) combined these two perspectives and defined SCA as a firm's capacity to analyze data through the utilization of quantitative methods. However, the definition provided by the author focuses only on capabilities related to data analysis and quantitative methodologies. Moreover, research acknowledges various types of SCA, including descriptive and diagnostic analytics, predictive analytics, and prescriptive analytics [25].

Recently, several review papers highlighted the use of predictive analytics in supply chain research. Tiwari, Wee, and Daryanto [16] provided instances of predicting customer behavior such as purchasing patterns and identifying trends in sales activities. In a similar vein, Zhong et al. [18] outlined predictive analytics to encompass aspects within supply chains like marketing and finance. This includes activities such as social media tracking, and tracking of exchange rates for trade. Cohen [26] and Sharma and Garg [27] proposed that connecting the internal production system with external partners, including both suppliers and consumers, is essential to leverage the potential of big data analytics in managing inventory and implementing automated inventory control strategies. Similarly, Wang et al. [15] used analytics to optimize inventory ordering decisions while Katchasuwanmanee et al. [28] deployed a combination of internal data supplemented by external, unstructured data to enhance the efficiency of production processes. On the other hand, Zhong et al. [21] employed big data from radio frequency identification (RFID) to facilitate logistical planning and scheduling on the shop floor, and then create an industrial IoT (Internet of Things) proof of concept system through the utilization of RFID tags.

Despite its power in predictive analytics, there is limited existing research that explores the use of machine learning within supply chains. The reason behind its omission from

the extensive literature on big data supply chain studies might stem from the scarcity of research focusing on machine learning-based techniques in this domain. Based on research conducted by IBM, the predominant discovery indicated that most organizations were in the very early stages of predictive analytics. The majority of those applying this approach were concentrating on sales and customer behavior forecasting, rather than supply chain operations [29] . Brintrup et al. [30] applied machine learning classification models such as Random Forest (RF), Support Vector Machine 3 (SVM), Logistic Regression, and K-Nearest Neighbour (KNN) to predict first-tier supply chain disruptions. Indeed, several authors have emphasized the potential of utilizing big data to predict supply chain disruptions.

According to Wang et al. [15], the use of supply chain analytics could extend to procurement to manage supply risks and suppliers' performance. This approach enables global supply chains to adopt a proactive stance rather than a reactive one in response to supply chain risks. He et al. [31] proposed a theoretical framework including natural language processing to extract potential risks from sources like news outlets, followed by risk classification and a simulation engine that predicts the potential impact of the risks on a company's KPIs. Furthermore, Fan et al. [32] introduced a structure that integrates big data analytics into supply chain risk management by incorporating both internal and external big data sources. Additionally, Dani [33] proposed a proactive strategy for managing risks within the supply chain, which involves the use of data mining techniques to uncover potential risk sources. Lastly, Brintrup et al. [22] suggested a method centered around graph mining to estimate how disruptions such as earthquakes may cascade in global supply networks.

In this paper, empirical analysis is used to demonstrate how data analytics, more specifically machine learning, can be used to forecast one of the key indicators that can improve supply chain performance, which is demand, and to predict supply chain disruptions such as late delivery risk. The analysis specifically focuses on the flow of goods from the Asia Pacific region to customers in the USA and Puerto Rico.

3. Methodology

The main research objectives are to observe the application of data analytics in the supply chain, to assess the impact of descriptive and predictive analytics on supply chain decision-making, to explore the effectiveness of predictive models in demand forecasting and finally to examine the influence of feature selection and model accuracy on predictive analytics outcomes. The study's key concepts and variables revolve around the adoption of predictive analytics, especially ML models, by organizations in their supply chain processes, the quality of the data used to do predictive analytics, the type of ML models used to do prediction, such as Linear Regression, Logistic Regression, SVM, Random Forest and etc, the accuracy of models in demand forecasting, as shown in Figure 1.

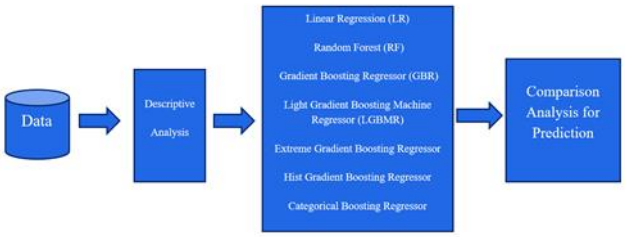


Fig. 1. The Proposed Models in Supply Chain Analytics

3.1. Data Collection & Preparation

The data used in this study was collected online in Mendeley Data [34] and also published in Kaggle, which is a widely recognized online platform that hosts data science competitions, provides datasets for analysis, and fosters a community of data enthusiasts, scientists, and machine learning practitioners. The dataset provides historical data on the flow of goods purchased from all over the world to customers in the USA and Puerto Rico. The dataset presents features such as order ID, product name, product category, product price, customer details, demanded quantity, sales and etc. The focus of this study was on the Asia Pacific market. Consequently, the data of this market was extracted from the dataset, and the variables (features set) that significantly contribute to the prediction were examined. The observation about the dataset, will be discussed as descriptive analysis.

3.2. Modelling

In this research, supervised machine learning models for regression will be utilized to forecast the demand. The main regression models we are going to use in this study are Linear Regression (LR), Random Forest (RF), Gradient Boosting Regressor (GBR), Light Gradient Boosting Machine Regressor (LGBMR), Extreme Gradient Boosting Regressor (XGBR), Hist Gradient Boosting Regressor (HGBR), Categorical Boosting Regressor (CBR).

3.2.1. Linear Regression (LR)

The goal of linear regression is to find the values of β1, β2, ..., βp that minimize the sum of squared differences between the observed y values and the predicted values ŷ based on the linear equation, as in (1):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \quad (1)$$

Where:

y is the target variable;

x_1, x_2, ..., x_p are the independent variables;

β_1, β_2, ..., β_p are the coefficients of the independent variables;

ϵ represents the error term, accounting for the variability that is not explained by the model.

3.2.2. Random Forest (RF)

The RF Regressor is a robust ML algorithm that is part of the ensemble learning family. It is employed for both classification and regression tasks. Ensemble learning involves combining multiple independent models to generate a more robust and accurate predictive model. In the case of the RF Regressor, it combines various decision trees to create a regression model that is more stable and accurate. RF starts by creating S subsets Ni (bootstrapped samples), where each subset contains n samples drawn randomly with replacement from the original dataset (if the bootstrap argument is valid), as shown in (2).

$$N_i = (x_1, y_1), (x_2, y_2), \dots\dots, (x_n, y_n) \quad (2)$$

For each decision tree, a random subset of features is selected for splitting at each node. This introduces diversity among the trees, preventing overfitting and capturing different aspects of the data. Decision trees are constructed using recursive binary splitting, as shown in Figure 3. At each node, the algorithm selects the best feature and split point based on criteria like mean squared error. The tree is grown until a stopping criterion is met, such as maximum depth or minimum samples per leaf.

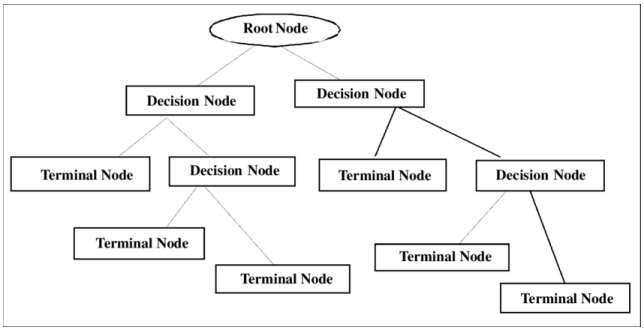


Fig.2. Decision Tree

Once all decision trees are built, the ensemble model is formed. Predictions from individual trees are combined to make the final prediction. The predictions from all trees are averaged to produce the ensemble prediction, as shown in (3).

$$y_{ensemble}(\mathrm{x}) = \frac{1}{\mathrm{S}}\sum_{i=1}^{S} \hat{y}_i(x) \quad (3)$$

3.2.3. Gradient Boosting Regressor (GBR)

The Gradient Boosting Regressor (GBR) is an ML algorithm that creates an ensemble of weak learners, typically decision trees, in a sequential manner, with each learner attempting to correct the errors of the previous ones.

The first step is to initialize the prediction $F_0(\mathrm{x})$ using the mean of the target variable, as in (4):

$$F_0(\mathrm{x}) = \mathrm{mean}(\mathrm{y}) \quad (4)$$

Next, the model calculates the residuals of the loss function, in each iteration k, regarding the current model's prediction for each data point, as in (5):

$$r_{ik} = -\frac{\partial \mathrm{L}\left(y_i, F_{k-1}(x_i)\right)}{\partial F_{k-1}(x_i)} \quad (5)$$

After this, the model fits a decision tree dt_k to the residuals r_{ik}, as in (6):

$$dt_k(\mathrm{x}) = \mathrm{DecisionTree}(\mathrm{X}, r_m) \quad (6)$$

The learning rate η controls the impact of each decision tree's prediction. Finally, the model makes a final prediction by summing the prediction of all iterations, as in (7):

$$F_n(\mathrm{x}) = F_0(x) + \sum_{k=1}^{n} \eta \cdot dt_k(\mathrm{x}) \quad (7)$$

3.2.4. Light Gradient Boosting Machine Regressor (LGBMR)

The Light Gradient Boosting Machine Regressor (LGBMR) is a machine learning algorithm that falls under the gradient boosting framework. It is built to effectively handle vast datasets and provides elevated functionality with considerably less memory usage in contrast to conventional gradient boosting algorithms. LGBMR uses a specific data structure called "Histogram-based Learning" which groups feature values into discrete bins. This allows for faster computation and memory efficiency. It follows the same algorithm as the GBR model. The LGBMR's key strengths include its ability to handle large datasets efficiently, its speed, and its capacity to capture complex patterns.

3.2.5. Extreme Gradient Boosting Regressor (XGBR)

The XGBoost (Extreme Gradient Boosting) Regressor is a highly influential and widely used ML algorithm within the gradient boosting framework. It is engineered to deliver exceptional predictive accuracy, handle diverse data types efficiently, and prevent overfitting. XGBR follows the GBR framework, which involves building an ensemble of weak learners (usually decision trees) to create a strong predictive model. The strength of XGBR lies in its ability to handle complex relationships in the data, prevent overfitting, and provide interpretable feature importance scores.

3.2.6. Hist Gradient Boosting Regressor (HGBR)

Histogram-Based Gradient Boosting Regressor (HGBR) is a variant of the gradient boosting algorithm that leverages

histogram-based techniques to improve both the efficiency and effectiveness of training decision trees in boosting. It is particularly suitable for large datasets and provides fast training times while maintaining high predictive accuracy. Like the LGBMR, the HGBR is a robust choice for large datasets where memory efficiency and faster training are crucial. It combines the strengths of gradient boosting with the benefits of histogram-based techniques, providing accurate predictions while minimizing computational costs.

3.2.7. Categorical Boosting Regressor (CBR)

The CatBoost Regressor (CBR) is a gradient boosting-algorithm tailored to handle categorical features effectively and accurately while providing high predictive accuracy. It is known for its automatic handling of categorical data, strong performance, and robustness against overfitting. CatBoost introduces a unique approach to handle categorical features. Unlike other gradient boosting algorithms, CatBoost natively handles categorical data without the need for pre-processing like one-hot encoding. It employs an ordered boosting strategy that deals with categorical variables more effectively. CatBoost's automatic handling of categorical features and its robustness against overfitting make it a valuable tool for a wide range of ML tasks. It is also known for its strong performance and ability to produce accurate predictions with minimal hyperparameter tuning.

3.3. Evaluation metrics

The following metrics will be used to evaluate the performance of the models predicting our target variable.

3.3.1 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is a widely used metric for evaluating the accuracy of regression models by measuring the average magnitude of the errors between predicted and actual values. It provides a measure of how well the model's predictions match the actual observed values. The equation of the RMSE metric is referred to (8):

$$\text{RMSE} = \sqrt{\frac{1}{\text{N}}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \quad (8)$$

Where:

N is the number of observations;

$\hat{y}_i$ is the predicted value for observation i;

y is the actual value of the observation i.

Lower RMSE values indicate better model performance, as they indicate that the model's predictions are closer to the actual values.

3.3.1. R^2

The R^2 metric is employed to evaluate how well a regression model fits the data. It quantifies the portion of the variance in the dependent variable that can be attributed to the independent variables in the model. R^2 serves as an indicator of how accurately the model's predictions align with the actual data.

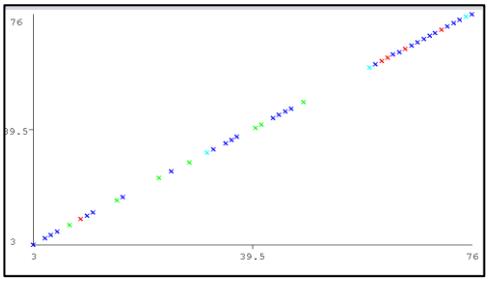
$$R^2 = 1 - \frac{\text{RSS}}{TSS} \quad (10)$$

The R^2 value ranges from 0 to 1. A higher R^2 value indicates that a larger proportion of the variance in the dependent variable is explained by the independent variables in the model. R^2 values closer to 1 indicate a better fit of the model to the data. However, R^2 should be interpreted in conjunction with other evaluation metrics such as RMSE, as it can be influenced by the number of variables and the complexity of the model.

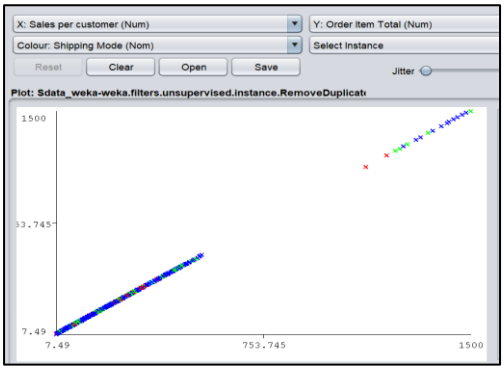
4. Results and Discussion

4.1. Exploratory Data Analysis

The data used in this research contains 53 variables and 41260 rows. 44.7% of the data is categorical, and 27.7% is respectively integer and float. The dataset contains many duplicated columns, such as order item product price and product price that should be removed. As well as some empty columns such as the order Zipcode and Product description. The correlation analysis shows that some variables are highly correlated with a correlation rate that can reach up to 99%. As shown in Figure 3, the variables can be considered redundant attributes and using one of the variables can be removed for modelling.



a. Category ID (x) vs Product Category ID (y)



b. Sales per customer (x) vs Order item total (y)

Fig. 3. Sample of scatter plot graphs (a) and (b) for redundant variables analysis

The dataset contains a few missing values, 5 missing values in the customer's last name column and 1 record in the customer zip code column out of 41260. So, we can drop them without any processing, this will not affect the results of the predictions. However, the dataset contains plenty of outliers, that should be handled, in many variables such as longitude, latitude, order item discount, order item profit ratio, sales, order item total, order profit per order, product price, department ID, and order ID. Figure 4 shows the sample of for variable longitude and order item discount for the box-plot graphs.

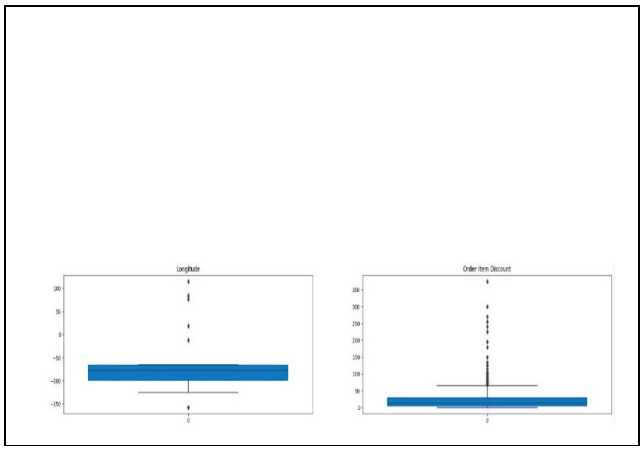
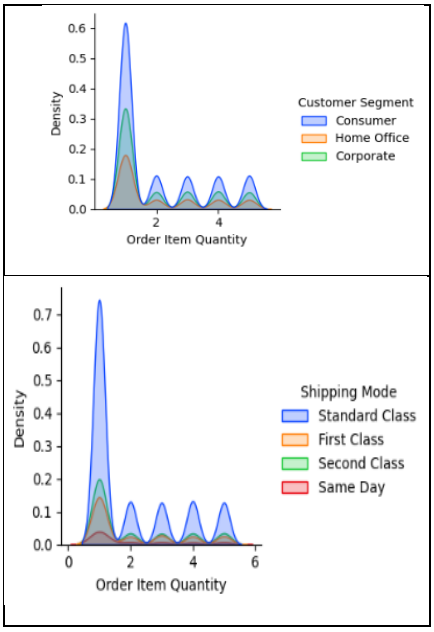


Fig. 4. Sample of box-plot graphs for two variables

The visualization of the distribution of demanded quantity in the function of variables such as customer segment, order status, and shipping mode shows that it is skewed to the left, as shown in Figure 5. Using the longitude and latitude information provided in the dataset, and the geopandas library in Python, we visualized the location of customers. The map, in Figure 6 shows that most customers are located in the USA and some of them are in Puerto Rico. The map shows some customers located in different other countries and that is due to outliers.



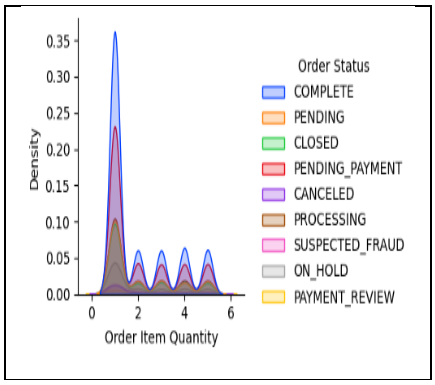


Fig. 5. Skewness analysis

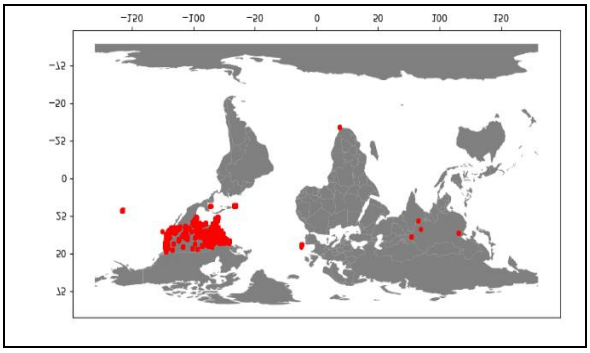


Fig. 6. Location of the Customers

Insights from historical data on the shipment of goods from the Asia Pacific region to Western countries are necessary to assess the performance of the supply chain in the region. In one point of view, the data is plotted to product category name to the total sales, total quantity, total amount per order and earnings per order placed, as shown in Figure 7. It helps the company to understand about the products and companies could have a better vision of the product categories, which categories are most demanded, and what should be improved to increase the sales of other categories.

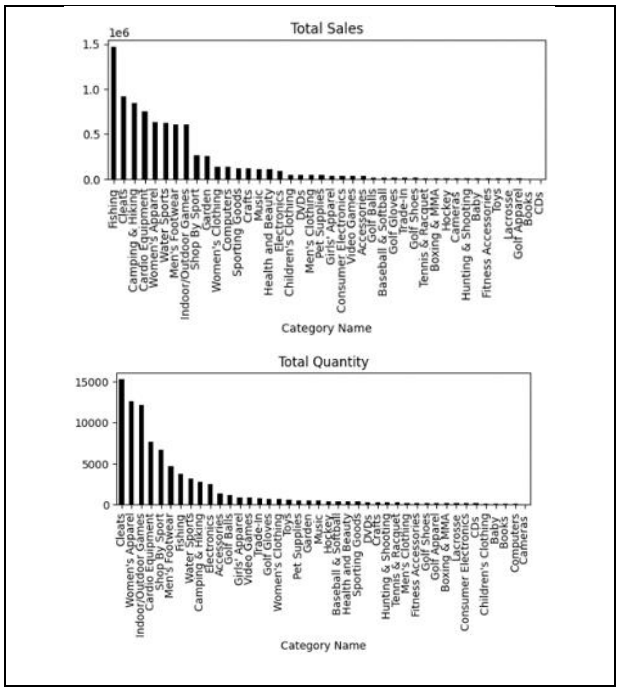


Fig. 7. Bar charts based on product category name

4.2. Predictive Analytics for Demand Forecasting

Predictive analytics are one of the most powerful tools to analyze and improve the organization's supply chain. In this study, ML regression was used to forecast demand, which includes models from Linear Regression, Random Forest, Gradient Boosting Regressor, LGBM, XGBoost, Hist Gradient Boosting, and CatBoost. Then, R2 and RMSE metrics were being compared. The selected list of attributes is shown in Table 1.

Table 1. List of input variables and output variable (Y)

Variable	*Description*
Type	Type of transaction made
Days for shipping (real)	Actual shipping days of the purchased product
Days for shipment (scheduled)	Days of scheduled delivery of the purchased product
Order Profit Per Order	Earnings per order placed
Sales per customer	Total sales per customer made per customer
Delivery Status	Delivery status of orders: Advance shipping, Late delivery, Shipping canceled, Shipping on time
Late_delivery_risk	Categorical variable that indicates if sending is late (1), it is not late (0).
Category Name	Description of the product category
Customer Country	Country where the customer made the purchase
Customer Segment	Types of Customers: Consumer, Corporate, Home Office
Customer State	State to which the store where the purchase is registered belongs
Department Name	Department name of store
Order City	Destination city of the order
Order Country	Destination country of the order
Order Item Discount Rate	Order item discount percentage
Order Item Product Price	Price of products without discount
Order Item Profit Ratio	Order Item Profit Ratio
Order Item Quantity	Number of products per order
Sales	Value in sales
Order Item Total (Y)	Total amount per order
Order State	State of the region where the order is delivered
Order Status	Order Status: COMPLETE, PENDING, CLOSED, PENDING_PAYMENT, CANCELED, PROCESSING, SUSPECTED_FRAUD, ON_HOLD, PAYMENT_REVIEW
Product Price	Product Price
Product Status	Status of the product stock: If it is 1 not available, 0 the product is available
Shipping Mode	The following shipping modes are presented: Standard Class, First Class, Second Class, Same Day

The prediction results performance is shown in Table 2.

Table 2. Values of performance measurement

Model	*R^2 Training Score*	*R^2 Testing Score*	*RSME Training Score*	*RSME Testing Score*
Linear Regression	0.25	0.27	1.23	1.23
Random Forest	0.96	0.78	0.27	0.68
LGBM	0.78	0.76	0.66	0.70
XGBoost	0.84	0.78	0.56	0.67
Hist Gradient Boosting	0.78	0.76	0.67	0.71
Gradient Boosting Regressor	0.69	0.69	0.79	0.80
CatBoost	0.80	0.77	0.63	0.68

It is noticed that Random Forest has the highest R^2 scores on both the train and test sets, and it also has the lowest RMSE scores on both sets. Nevertheless, the prediction using Linear Regression is not relevant as the R^2 scores are low and RMSE is high, and this may be due to the noisy and skewed data. If these models were compared, it can be concluded that Random Forest is the most performant model in forecasting demand. However, these results were obtained before tuning, and it is uncertain whether the best parameters were selected for each model. Thus, model hyperparameters need to be tuned using the tuning models cited in the methods section, and the results will be

compared, as shown in Table 3.

Table 3. Values of measurement after tuned.

Model	*R^2 Training Score*	*R^2 Testing Score*	*RSME Training Score*	*RSME Testing Score*
Linear Regression	0.25	0.27	1.23	1.23
Random Forest	0.96	0.78	0.27	0.68
LGBM	0.88	0.79	0.50	0.66
XGBoost	0.93	0.80	0.36	0.64
Hist Gradient Boosting	0.83	0.78	0.58	0.67
Gradient Boosting Regressor	0.87	0.79	0.51	0.66
CatBoost	0.86	0.80	0.52	0.65

After tuning, it can be noticed that the scores of LGBM, XGBoostR, HistGradientBoosting, and CatBoost improved. Random Forest still has the highest R^2 score and lowest RMSE on the train set, but the XGBoost and CatBoost have the highest R^2 score and the lowest RMSE on the test set. The model hyperparameters have a remarkable effect on the performance and prediction accuracy. Hyperparameters should be selected carefully and tuned, so they can enable the model to give accurate predictions. The RMSE scores remain high, and other tools should be considered to improve the performance of the models. However, overall, Random Forest is deemed a powerful model for forecasting demand in our case. The features importance diagram of the most accurate model gives insight into the factors that companies should consider improving their sales and satisfy the demand. Figure 8 shows five important features that is related to the prediction of demand. In our case, the profit is one of the deterministic factors of the demand trend, and this may be due to the price of the items as the profit has a direct effect on the price. The less expensive the items are, the higher the demand.

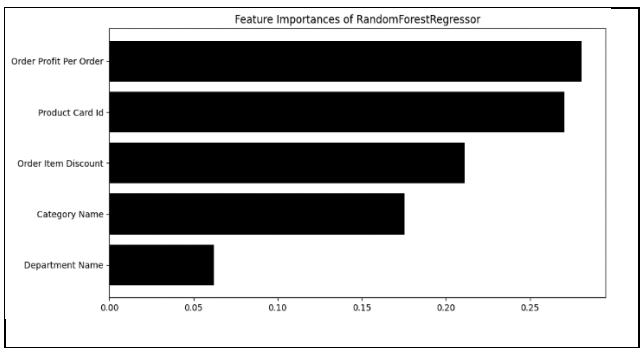


Fig. 8. Feature importance of Random Forest Regressor

5. Conclusion

The Asia Pacific market is a dynamic market with significant economic potential, driven by factors such as robust economic development, technological advancements, changing demographics, and increasing urbanization. This has profound implications for supply chain management, as the region becomes a hub for economic activity and trade. Supply chains in the Asia Pacific market are becoming more complex and interconnected, requiring businesses to adopt advanced strategies and tools to enhance their performance. The adoption of supply chain analytics becomes imperative, leveraging data-driven insights and predictive models to optimize procurement, inventory management, demand forecasting, and logistics. This study analyzed the performance of the Asia Pacific market supply chain using descriptive analytics and metrics such as demand, sales, and earnings. Predictive analytics techniques using machine learning (ML) were employed to forecast demand for Asia Pacific market products from Western customers. The study contributes to the supply chain analytics field in the Asia Pacific region by setting a theoretical framework that allows organizations to adopt proactive strategies to optimize their performance based on different supply chain components. However, the paper has several limitations related to data collection, as it does not involve companies and manufacturers, and the use of predictive analytics techniques is sensitive to unstructured and skewed data. Further research is needed to improve prediction performance and optimize the learning process.

Acknowledgement

The authors would like to express the gratitude to Ministry of Higher Education Malaysia for the Fundamental Research Grant Scheme FRGS/1/2022/SS01/UITM/01/4 for the research funding, also to Research Management Center, UiTM and also to College of Computing, Informatics and Mathematics, UiTM, for the internship placement for the international student. The authors would like to acknowledge late Dr Shahrin Nasir, one of the members in the team who had contributed to the research and also Nik Nur Alissa Sufi Nik Aziz as the research assistant.

References

[1] "World Bank Open Data," World Bank Open Data. https://data.worldbank.org/indicator/NE.EXP.GNFS.CD?locations=Z4

[2] C. Bode and J. R. Macdonald, "Stages of Supply Chain Disruption Response: Direct, Constraining, and Mediating Factors for Impact Mitigation," Decision Sciences, vol. 48, no. 5, pp. 836–874, Sep. 2016, doi: 10.1111/deci.12245.

[3] B. Avishai, "The Pandemic Isn't a Black Swan but a Portent of a More Fragile Global System," The New Yorker, Apr. 21, 2020. [Online]. Available: https://www.newyorker.com/news/daily-comment/the-pandemic-isnt-a-black-swan-but-a-portent-of-a-more-fragile-global-system

[4] M. Mubarik and R. Z. B. R. M. Rasi, "Triad of Big Data Supply Chain Analytics, Supply Chain Integration and Supply Chain Performance: Evidences from Oil and Gas Sector," Humanities and Social Sciences Letters, vol. 7, no. 4, pp. 209–224, Jan. 2019, doi: 10.18488/journal.73.2019.74.209.224.

[5] "Supply Chain 4.0: A luxury or a necessity ?," Social Innovation, Jan. 19, 2024. [Online]. Available: https://social-innovation.hitachi/en-us/think-ahead/manufacturing/supply-chain-4-0/

[6] F. Steinberg, P. Burggräf, J. Wagner, B. Heinbach, T. Saßmannshausen, and A. Brintrup, "A novel machine learning model for predicting late supplier deliveries of low-volume-high-variety products with application in a German machinery industry," Supply Chain Analytics, vol. 1, p. 100003, Mar. 2023, doi: 10.1016/j.sca.2023.100003.

[7] M. A. Waller and S. E. Fawcett, "Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management," Journal of Business Logistics, vol. 34, no. 2, pp. 77–84, Jun. 2013, doi: 10.1111/jbl.12010.

[8] R. D. Raut, S. K. Mangla, V. S. Narwane, M. Dora, and M. Liu, "Big Data Analytics as a mediator in Lean, Agile, Resilient, and Green (LARG) practices effects on sustainable supply chains," Transportation Research Part E Logistics and Transportation Review, vol. 145, p. 102170, Jan. 2021, doi: 10.1016/j.tre.2020.102170.

[9] N. A. T. Oyewole, N. C. C. Okoye, N. O. C. Ofodile, and N. E. Ejairu, "Reviewing predictive analytics in supply chain management: Applications and benefits," World Journal of Advanced Research and Reviews, vol. 21, no. 3, pp. 568–574, Mar. 2024, doi: 10.30574/wjarr.2024.21.3.0673.

[10] T. Schoenherr and C. Speier-Pero, "Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential," Journal of Business Logistics, vol. 36, no. 1, pp. 120–132, Feb. 2015, doi: 10.1111/jbl.12082.

[11] M. W. Barbosa, A. De La Calle Vicente, M. B. Ladeira, and M. P. V. De Oliveira, "Managing supply chain resources with Big Data Analytics: a systematic review," International Journal of Logistics Research and Applications, vol. 21, no. 3, pp. 177–200, Aug. 2017, doi: 10.1080/13675567.2017.1369501.

[12] "Transforming Data With Intelligence," TDWI. https://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx

[13] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Information Sciences, vol. 275, pp. 314–347, Aug. 2014, doi: 10.1016/j.ins.2014.01.015.

[14] N. Chen, N. Chiang, and N. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," MIS Quarterly, vol. 36, no. 4, p. 1165, Jan. 2012, doi: 10.2307/41703503.

[15] G. Wang, A. Gunasekaran, E. W. T. Ngai, and T. Papadopoulos, "Big data analytics in logistics and supply chain management: Certain investigations for research and applications," International Journal of Production Economics, vol. 176, pp. 98–110, Jun. 2016, doi: 10.1016/j.ijpe.2016.03.014.

[16] S. Tiwari, H. M. Wee, and Y. Daryanto, "Big data analytics in supply chain management between 2010 and 2016: Insights to industries," Computers & Industrial Engineering, vol. 115, pp. 319–330, Jan. 2018, doi: 10.1016/j.cie.2017.11.017.

[17] K. Spanaki, Z. Gürgüç, R. Adams, and C. Mulligan, "Data supply chain (DSC): research synthesis and future directions," International Journal of Production Research, vol. 56, no. 13, pp. 4447–4466, Nov. 2017, doi: 10.1080/00207543.2017.1399222.

[18] R. Y. Zhong, S. T. Newman, G. Q. Huang, and S. Lan, "Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives," Computers & Industrial Engineering, vol. 101, pp. 572–591, Nov. 2016, doi: 10.1016/j.cie.2016.07.013.

[19] K. Govindan, T. C. E. Cheng, N. Mishra, and N. Shukla, "Big data analytics and application for logistics and supply chain management," Transportation Research Part E Logistics and Transportation Review, vol. 114, pp. 343–349, Jun. 2018, doi: 10.1016/j.tre.2018.03.011.

[20] Z. Inamdar, R. Raut, V. S. Narwane, B. Gardas, B. Narkhede, and M. Sagnak, "A systematic literature review with bibliometric analysis of big data analytics adoption from period 2014 to 2018," Journal of Enterprise Information Management, vol. 34, no. 1, pp. 101–139, Mar. 2020, doi: 10.1108/jeim-09-2019-0267.

[21] R. Y. Zhong, S. Lan, C. Xu, Q. Dai, and G. Q. Huang, "Visualization of RFID-enabled shopfloor logistics

Big Data in Cloud Manufacturing," The International Journal of Advanced Manufacturing Technology, vol. 84, no. 1–4, pp. 5–16, Aug. 2015, doi: 10.1007/s00170-015-7702-1.

[22] A. Brintrup, Y. Wang, and A. Tiwari, "Supply Networks as Complex Systems: A Network-Science-Based Characterization," IEEE Systems Journal, vol. 11, no. 4, pp. 2170–2181, Dec. 2017, doi: 10.1109/jsyst.2015.2425137.

[23] G. C. Souza, "Supply chain analytics," Business Horizons, vol. 57, no. 5, pp. 595–605, Sep. 2014, doi: 10.1016/j.bushor.2014.06.004.

[24] S. F. Wamba and S. Akter, "Understanding supply chain analytics capabilities and agility for data-rich environments," International Journal of Operations & Production Management, vol. 39, no. 6/7/8, pp. 887–912, Dec. 2019, doi: 10.1108/ijopm-01-2019-0025.

[25] B. Pochiraju and S. Seshadri, Essentials of Business Analytics. 2019. doi: 10.1007/978-3-319-68837-4.

[26] M. A. Cohen, "Inventory Management in the Age of Big Data," Harvard Business Review, Aug. 08, 2019. https://hbr.org/2015/06/inventory-management-in-the-age-of-big-data

[27] M. Sharma and N. Garg, "Inventory Control and Big Data," in Advances in logistics, operations, and management science book series, 2016, pp. 222–235. doi: 10.4018/978-1-4666-9888-8.ch011.

[28] K. Katchasuwanmanee, R. Bateman, and K. Cheng, "Development of the Energy-smart Production Management system (e-ProMan): A Big Data driven approach, analysis and optimisation," Proceedings of the Institution of Mechanical Engineers Part B Journal of Engineering Manufacture, vol. 230, no. 5, pp. 972–978, Jun. 2015, doi: 10.1177/0954405415586711.

[29] D. Wardman, "Bringing Big Data to the Enterprise - Gaining new insight with Big Data capabilities," 2013. [Online]. Available: https://public.dhe.ibm.com/software/os/systemz/pdf/09_-_Dan_Wardman_-_Bring_Big_Data_to_the_Enterprise_.pdf

[30] A. Brintrup et al., "Supply chain data analytics for predicting supplier disruptions: a case study in complex asset manufacturing," International Journal of Production Research, vol. 58, no. 11, pp. 3330–3341, Nov. 2019, doi: 10.1080/00207543.2019.1685705.

[31] N. M. He, N. H. Ji, N. Q. Wang, N. C. Ren, and R. Lougee, "Big data fueled process management of supply risks: Sensing, prediction, evaluation and mitigation," Proceedings of the Winter Simulation Conference 2014, Dec. 2014, doi: 10.1109/wsc.2014.7019960.

[32] Y. Fan, L. Heilig, and S. Voß, "Supply Chain Risk Management in the Era of Big Data," in Lecture notes in computer science, 2015, pp. 283–294. doi: 10.1007/978-3-319-20886-2_27.

[33] S. Dani, "Predicting and Managing Supply Chain Risks," in International series in management science/operations research/International series in operations research & management science, 2009, pp. 53–66. doi: 10.1007/978-0-387-79934-6_4.

[34] F. Constante, F. Silva, and A. Pereira, "DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS," Mar. 12, 2019. https://data.mendeley.com/datasets/8gx2fvg2k6/5