Role of Data Analytics to overcome Challenges in Supply Chain Management

MSc Research Project
MSC Data Analytics

Muddassir Ahmed
Student ID: x23138688

School of Computing
National College of Ireland

Supervisor: Arjun Chikkankod

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……Muddassir Ahmed…………………………………………………………………………………

Student ID: ……x23138688……………………………………………………………………………….……

Programme: ……MSC Data Analytics…………………………… **Year:** ……2024……………..

Module: ……MSC Research Project……………………………………………………………….………

Supervisor: ……Arjun Chikkankod……………………………………………………………………….………

Submission Due Date: ………12 August 2024……………………………………………………………………….………

Project Title: Role of data analytics to overcome challenges in supply chain management

Word Count: …………7311…………………… **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: ………Muddassir Ahmed……………………………………………………………

Date: …………12-08-2024………………………………………………………………………

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):	

Role of Data Analytics to overcome Challenges in Supply Chain Management

Muddassir Ahmed
x23138688

Abstract

Supply chain management is facing big challenges in business. In this paper, we are trying to find how data analytics can be used in supply chain management, specifically in demand forecasting of the products and late deliveries of the shipment. The main focus of this thesis is on two supply chain areas: demand forecasting of goods based on previous sales and late deliveries of the shipment analysis. By using different data-driven models to check demand forecasts based on previous sales of the products, we also analyze the main reasons behind late deliveries. Based on these, companies can take proper measures to reduce risks, and they can improve operational efficiency. The ordinary least squares regression is used to evaluate previous sales of the dataset, and classification models are used to evaluate the reasons for the late delivery of the shipment. We then use a confusion matrix to evaluate the performance of the classification models. Data analytics is used for the supply chain industry, where companies can improve their operations and resources, and can optimize their resources and operations. This study aims to find ways through predictive data modelling by using different machine learning algorithms to improve performance in supply chain management.

1 Introduction

Companies are facing late delivery and demand forecasting issues. Supply chain management (SCM) is the pillar of the world trade and smooth movement of the goods in all over the world. Despite this, globally there are many challenges in supply chain such as prediction of the future demand based on previous and long delays in shipments. Which is affecting operational efficiency of the trade in over all the world. Supply chain management (SCM) is playing a very crucial role in increasing a company's performance by adapting technologies. One of the prominent advancements in SCM is adoption of machine learning (ML) that is providing solutions of supply chain industry. Some recent research in the field of machine learning has huge impact in different aspects of supply chain management especially in demand prediction of the goods and sales forecasting.

1.1 Background Information:

In today’s competition of the global market supply chain is playing very important role of different products from manufacturing to delivery to end user. To maintain this competitive market challenges, supply chain should adopt and optimize processes. However, there are several factors which leads to interrelated challenges which reduce the performance of the overall process. Data analytics in supply chain industry is increasing day by day due to its broad capabilities that is including customer behavior analysis in different sectors trend analysis and the demand prediction of the goods. This research is exploring in different way just to predict big data analytics applications in supply chain demand forecasting of the goods that is providing a classification models of these applications which is identifying research gaps in supply chain and giving insights for future of the research.

1.2 Fluctuating of the Demand:

It is very difficult to find consumer behavior that what he wants in future. Supply chain always deal with uncertainty in the demand of the products this is the main problem. If any error occurs in predicting of the goods demand that can leads to inventory imbalance and stock shortage which is major challenge for any organization.

1.3 Late Deliveries:

It is very important for supply chain to make sure goods arrive its destination on the specified time period just to keep customer happy and satisfied. Timely delivery of the goods is also important for any business success. If goods deliver late to the customer place that can leads to customer unhappy and becomes high operating cost. If any company want to improve supply chain process, then it is important to understand main reasons of the late deliveries.

Data analytics and machine learning can be used to identify problems for late deliveries. By identifying trends and outliers in the dataset organizations can get valuable insights which can resolve the issues. This study aims to bridge the gap between traditional methods and advanced analytical techniques those are used to predict future demand bas on the previous sale and reasons for the late deliveries. Data driven analytics provide broad picture within supply chain that helps companies to respond better to market change and they can resolve the issue which cause late deliveries.

1.4 Goal of the Project:

This research is conducted to create a plan in supply chain industry and its use of data analytics. This study focus on two major areas. One is predicting future sale based on previous sale and why deliveries get delayed.

1.5 Research Questions:

1) Predicting Demand (Sale): How accurately can we forecast sales using OLS linear regression based on various order-related and customer-related features?

2) Late Deliveries: Which classification algorithm performs best in predicting late delivery risk based on order and customer data?

2 Related Work

Supply chain management (SCM) in the field of machine learning has huge impact in different aspects of supply chain management especially in demand prediction of the goods and sales forecasting. In this research author (Husna and Shah; 2021) examine the applications of the machine learning in supply chain industry which is emphasizing D&SF. They highlight the importance of machine learning for handling complex problems and improving forecasting accuracy. This study tells about Fuzzy-ANN approaches excel in managing inaccurate data for example weather forecasting. Decision tree model and random forest give valuable interpretation in this matter. Furthermore, data pre-processing techniques reduce model complexity that balance the accuracy and computing time. A critical comparison conducted by (Cadavid and Grabot; 2018) between machine learning and traditional forecasting methods within the study that is indicates machine learning techniques. Which outperform traditional methods in the terms of accuracy and especially when dealing with variables. This paper concludes by recommendation for further research on data pre-processing techniques that can reduce the computational cost. It is also emphasizing the need to make the advance technologies which will be accessible to the small and medium size enterprises (SMEs) companies, which has low financial resources or expertise to implement them.

Risk management often relies on post event analysis and historical data in supply chain. Which is effective for addressing real time disruptions. (Feizabadi et al.; 2022) proposed methodology advocates for a forward looking approach which is predictive analytics to predict potential disruptions. This research explores various predictive analytics methods for example time series analysis for some anomaly detection and NLP field. Optimization for the risk assessment models that are using machine learning algorithms that are ensuring their accuracy in ML field and adaptability in dynamic environments of the research. (Carbonneau and Vahidov; 2008) highlights some most important synergistic relationship between different predictive analytics machine learning models and the supply chain industry by using theoretical discourse in the research and practical evidence of models. Case studies from different sectors which highlight underscore the proposed strategy's effectiveness and the benefits.

Data analytics in supply chain industry is increasing day by day due to its broad capabilities that is including customer behavior analysis in different sectors trend analysis and the demand prediction of the goods. (Singh and Panse; 2022) is exploring in different way just to predict big data analytics applications in supply chain demand forecasting of the goods that is providing a classification models of these applications which is identifying research gaps in supply chain and giving insights for future of the research.

(Zohdi and Salamiraad; 2022) categorizes big data analytics methods used in supply chain management into seven main techniques. That are time series forecasting, K-nearest-neighbours, clustering, support vector machines, neural networks, regression analysis and support vector regression. The neural networks and regression analysis are most frequently used techniques due to their effectiveness in handling complex and large datasets. This study also highlights the comparative advantages and limitations of each technique which is providing a detailed analysis of their applications in demand forecasting.

Big data analytics has impact on retail supply chain business for enhancing various performance metrics. (Ampazis et al.; 2015) evaluates BDA tools against performance criteria such as supplier integration, capacity utilization customer integration, demand management, cost, flexibility and time. This analysis highlighting issues in different retail companies between customer loyalty to its brand and cost of implementing BDA practices. This is the important for selecting the appropriate BDA tools to balance these competing priorities.

Moreover, this research is particularly relevant to the Indian retail supply chain management though they may vary across different industries. This study suggests that the prominent values and priorities of BDA practices might change. It is depending on the specific industry such as cost being the main criterion in one industry while demand management is the paramount in another. (Kilimci and Ekmis; 2019) researcher provide a framework for evaluating BDA practices in the retail. This Framework is very important and it is emerging in different retail firms in selecting the most effective BDA tools based on their specific supply chain performance measures which is offering a tailored approach to enhancing overall supply chain efficiency.

This paper classifies SCMF into key areas where predictive analytics is most frequently applied. That is demand management and procurement. When user want to check accurate demand forecasting for the goods and sensing along with sourcing risk and supplier of the products. Selection is identified as the foremost applications in different machine learning and big data analytics (BDA) enabled predictive models.

(Sengar and Ahmed;2019) identifies network based and support vector regression (SVR) based models are most use for demand prediction which is emphasizing their effectiveness in improving SCM functions. One of the main key contribution of this research is the development of a taxonomy for SCM result components that is includes all necessary elements for effectively implementing SCMF by using predictive analytics.

Techniques in machine learning is essential for improving predictive capabilities specific in the context of supply chain 4.0. (Zhu et al.; 2021) study explores the application of ML algorithms to predict the late deliveries and providing organizations to enhance customer retention and forecast of the future behavior. The study utilizes the SelectKBest method and ANOVA's f_classif() function to identify the most relevant predictors. (Lahmer and Akantous; 2022) research suggests future work will involve expanding the study to other supply chain processes which will enable to achieve high performance for combining machine learning and deep learning models.

When we are talking in the context of logistics and supply chain management then efficiency of transportation is crucial for overall operational success. This study investigates the impact of delays in warehouses of Indonesia for goods to move from the warehouse door to the port.

(Mahraz and Berrado; 2022) study uses ordinary least squares (OLS) regression model to analyze the relationship between warehouse loading delays (independent variable X) and the delay in the door-to-port process (dependent variable Y). The regression equation obtained is Y=2.2564+1.3546XY = 2.2564 + 1.3546XY=2.2564+1.3546X with an R2R^2R2 value of 0.9584 that is indicating 96% of the variation in the door to port delay that can be explained by the warehouse loading delay. This high coefficient of determination is underscore the significant impact of the warehouse delays on the overall logistics process. (Wiyanti and Nugroho; 2021) research gives future recommendation that incorporate additional variables that might be affect the door to port overall processing speed and explore other regression models for the comparative analysis. This could be giving more inclusive understanding of the factors which are influencing logistics performance and they offer insights into further optimizing supply chain processes. In recent years, the integration of big data analytics in supply chain industry has become a focal point for researchers. This is aiming to enhance the sustainability and efficiency of supply chain. Researchers explore the application of BDA in optimizing SCM and emphasizing the importance of a robust strategy for managing the supply chain.

(Carbonneau and Laframboise; 2007) researcher highlight that traditional SCM approaches often overlook critical parameters like customer preferences clearing complexity and natural calamities that are vibrant for a comprehensive supply chain strategy. To address the gap this study, propose an optimal strategy for maintaining supply chains by leveraging a dataset comprising car attributes such as age, sale price, market value and model, that is applying statistical investigations such as regression analysis. This approach then compares with the outputs of the data analytics algorithms by using feature selection techniques just to predict car supply more accurately. (Vairagade et al.; 2019) findings of the research is the effectiveness of the covering approach for feature selection in improving the predictive model's accuracy. This method, which is involves selecting features that is to enhance the performance of the model and demonstrated superior results compared to traditional statistical methods such as chi-squares and multi-regression analysis.

(Tirkolaee and Aeini; 2021) research provides a comprehensive analysis of machine learning and business intelligence applications in the domain of demand forecasting of the goods and underscoring their significance in modern enterprises. The authors highlight that accurate demand forecasting is crucial for optimizing operational efficiency and enhancing the overall business performance. This is particularly relevant in today's dynamic market environment where precise predictions can be substantially having much impact on strategic decision making and operational outcomes.

(Mediavilla and Palm; 2022) paper emphasizes the integration of business intelligence with machine learning techniques to forecast future demand for goods. Data is collected from different sources and the study demonstrates how ML algorithms can predict weekly monthly and quarterly demand with high accuracy. Researcher use AWS SageMaker just to implement machine learning models that are focusing on the comparison between predicted and actual sales of data to validate the forecasting accuracy. This approach achieved up to 92.38% accuracy in real-time organizational data that is illustrating the effectiveness of their method. (Saha et al.; 2022) research explores the role of time series analysis and rule based forecasting methods for enhancing demand prediction. This study finds that DeepAR models which is a type of deep learning architecture which tailored for time series forecasting and its offer superior accuracy compared to traditional old methods. This shows their ability to handle complex patterns in large datasets those leads to the more reliable forecasts.

This study (Arif and Sany; 2019) highlights the transformative potential of machine learning in converting large amounts of data into actionable insights so therefore enhancing supply chain efficiency. This review is based on analysis of 79 selected papers from an initial pool of 1870 which is focusing on the most relevant machine learning algorithms which is used in demand forecasting of the goods. These algorithms are including neural networks, artificial neural networks, support vector regression and support vector machine.

This research (Terrada and Ouajji; 2022) categorizes the machine learning applications into three main sectors: industry, agriculture and services. The majority of applications (65%) are from industrial sector that are highlighting the sector's readiness and need for advanced predictive analytics just to optimize operations. Only 5% of the reviewed studies focus on the agricultural sector which is highlighting a significant research gap. The limited application of machine learning in the agriculture sector underscores the need for the more targeted studies.

Study emphasize (Solanki et al.; 2020) adoption of machine learning algorithms in demand forecasting can lead to more accurate predictions. It reduced computational costs in supply chain management. Researcher also highlight the importance of expanding machine learning and its application to the agricultural and service sectors. It is also covering transportation and healthcare which are critical for economic growth and stability in COVID-19 pandemic.

The study highlights the need of future research to address the identified gaps particularly in the agricultural sector. It is more comprehensive investigation how machine learning can effectively implement in this sector just to enhance supply chain efficiency.

Application of machine learning in supply chain industry has become significant attention. The predictive use of machine learning model within supply chain industry has particularly use for forecasting late deliveries by suppliers. (Bousqaoui and Tikito; 2019) study develops and test a regression based machine learning model just to predict the late deliveries of the goods.

ML applications in SCM have predominantly focused on classification algorithms to predict whether delays will occur, rather than the severity and timing of these delays. These classification methods often struggle with the curse of dimensionality, which limits their applicability in settings characterized by low volume and high variety. This study by (Mitra and Kumar; 2022) deviates from this trend by employing regression algorithms to predict not just the occurrence, but the severity of delivery delays, thus providing a more nuanced and actionable forecast for manufacturers. (Mohamed and Abdelaziz; 2020) study offers several practical implications. By using regression algorithms, the proposed model not only predicts the occurrence of delays but also their severity, addressing a significant gap in existing research. This capability is particularly beneficial for low-volume-high-variety manufacturers who face unique challenges in supply chain. (Khan and Mohamed; 2020) study shows that it is possible to mitigate the curse of the dimensionality, which is making the model applicable to diverse manufacturing settings.

3 Research Methodology

This study use regression and classification models to figure out how to improve supply chain management for prediction of the demand and understanding of the late deliveries.

3.1 Demand Prediction: Regression Models

For demand forecasting we used ordinary least square regression model. This model is used to estimate coefficient of linear regression. We are evaluating models using metrics such as mean absolute error (MAE) and root mean square error (RMSE) to check how well they predict demand based in past sales.

We have checked mean absolute error (MAE) and root mean square error (RMSE) that how accurate our regression models are forecasting.

Handling object type data: Object type attributes are dealing without effecting performance of the model. Instead of deleting these columns we have tried to find ways to include them effectively.

3.2 Analysis of Late Delivery: Classification

We use classification for late delivery analysis, we use different classification models such as random forest, support vector machines (SVM), logistic regression, linear discriminant analysis and gaussian naive bayes. These models help us to determine whether shipment deliver to customer place on time or late.

We evaluate these models by using accuracy, recall, and f1 Score just to see how well they predict delivery status.

3.3 Evaluation of Prediction Outcome:

- True Positive (TP): This shows how the prediction that a delivery will be late so appropriate action can be taken to minimize its delay.
- True Negative (TN): This shows how correctly predicting that a delivery will be on time.
- False Positive (FP): This shows how wrongly predicting that a delivery will be on time but actually delivery will be late.
- False Negative (FN): This shows how wrongly predicting that a delivery will be late but actually delivery will be on time.

The main purpose of this study is to make supply chains industry more efficient by accurately prediction of future demand and identifying the main causes of late deliveries by using data analytics. This study will evaluate which model such as regression and classification is best for supply chain management.

3.4 Limitations of the Study

The accuracy for the prediction of future sale demand and late deliveries of the shipment can be influenced by the quality and completeness of the data. Incomplete data or incorrect information can affect the performance of regression models which way give less accurate classification result.

Second, supply chain management is a complex system which is interacting with multiple factors that may impact delivery times. Implemented a model that covers every variable is very challenging due to its factors such as transportation method and products handled at various stages.

Despite these limitations our study providing valuable insights into supply chain management.

4 Design Specification

Aim of this research is to increase the supply chain process for moving goods from one location to other location by using advanced statistical methods. Goal is to explore the main problems of the late delivery and what opportunities by using ML to increase supply chain efficiency.

We are building predictive models to forecast demand of goods and analyze late deliveries issues focusing on real world supply chain management. It solves common supply chain problems using a dataset from mendeley data which publically available on

‘ https://data.mendeley.com/datasets/8gx2fvg2k6/5’

This rich dataset is perfect for training on our predictive models to uncover valuable insights.

4.1 For demand forecasting:

we used Ordinary least square regression model. This model is used to estimate coefficient of linear regression. We are evaluating models using metrics such as mean absolute error (MAE) and root mean square error (RMSE) to check how well they predict demand based in past sales.

4.2 For analysing late deliveries:

we are using classification models such as Gaussian Naive Bayes, Linear Discriminant Analysis, Support Vector Machines and Random Forest, Logistic Regression Classification.

These models helps to predict the orders being late that will allow the company to take early actions to prevent delays and keep the supply chain running smoothly.

Throughout this project we are pre-processing data, Feature extraction, training the models, testing the model and then evaluating of the performance.

Outcomes of this research include practical insights for the organizations to help them boost efficiency in their supply chain operations. While reducing risks that could mainly cause of customer dissatisfaction. Additionally, this study aims to add valuable information in supply chain optimization by using advanced data analytics that will contribute to the academic research in data analytics field.

4.3 Data Pre-processing:

This dataset has fifty-three (53) columns and one lac eighty thousand and five hundred eighteen (180518) rows.

4.4 Data Cleaning: Some columns found with missing values, customer last name, product description, customer zip code, order zip code. Late delivery column is a binary form (0,1) Its mean 0 has no delay and 1 for delays in delivery. Some orders are late its showing value 0.578 that mean orders are getting late that is leading customer disappointment and company will reduce its profit.

4.5 Missing Values: We replaced last name of the customer with Not Determined and zero is replace in columns zip code and order zip code. Column name description has missing values and it will not use in our processing so its dropped from the dataset.

4.6 Last Name of the Customer: Last name missing is filled with not determined and then created a new column full name and it combined first name and last name.

4.7 Outlier Handling: Outliers is not found in the dataset. Our model is non-parametric and it can handle extreme values as well but we did not remove any outlier. Our model is used to capture the patterns in the dataset.

4.8 Feature Engineering: Data Correlation

Heatmap provide some clear insights of the dataset. As we have found some duplicate columns name. After Removing some un used attributes because they have missing values and also these are not contributing in analysis. By heatmap we are choosing some attributes for further processing.

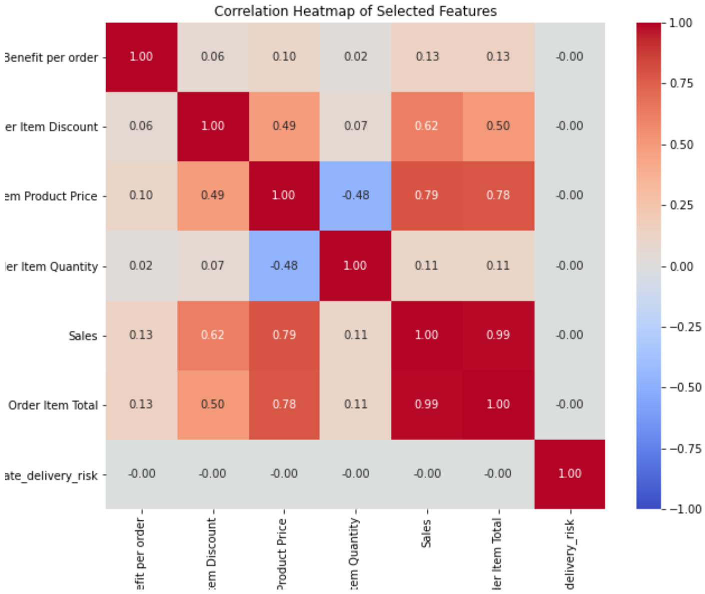


Figure 1: Correlation Heatmap of Selected Features

The above heatmap gave us some useful insights:

- Some columns have strong links with other. For example, order item total is with order discount, order quantity and sales.
- Column name sale has closely relationship with some other attributes of dataset, which are discount and product price.
- Analysis identified some relationships which are opposite in different columns such as quantity of the items and price of the product. Higher price items have few orders while products price between $10 and $100 has more frequent orders. This suggests that expensive items are ordered very less as compare to moderate price items.

4.9 Exploratory Data Analysis

Different statistical methods are used for visualizations just to understand the information such as product category performance, market trends and delivery. The main goal is to identify any interesting relationships correlations or any anomalies that can help us to understand the dataset better way and it can improve business decisions.

4.10 Customer Segment Analysis

Business strategy always make after knowing the customer and it is very crucial part. This analysis showing different factors such as age, gender and race to help us break down our customer based into different segments. Our aim is to spot common traits or behavior so that we can understand our customers better and then we can make our marketing, product development and service delivery just to meet their expectations.

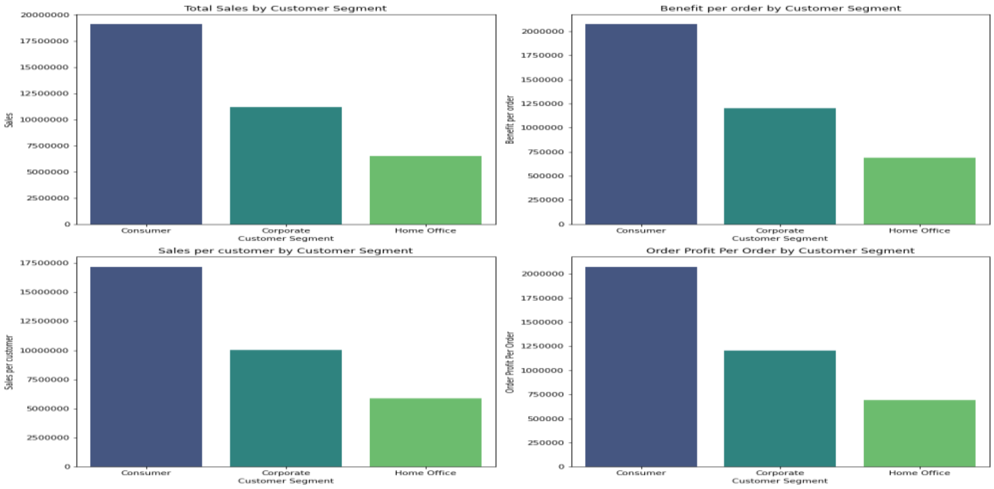


Figure 2: Customer Segment for sales, profit and benefits per order

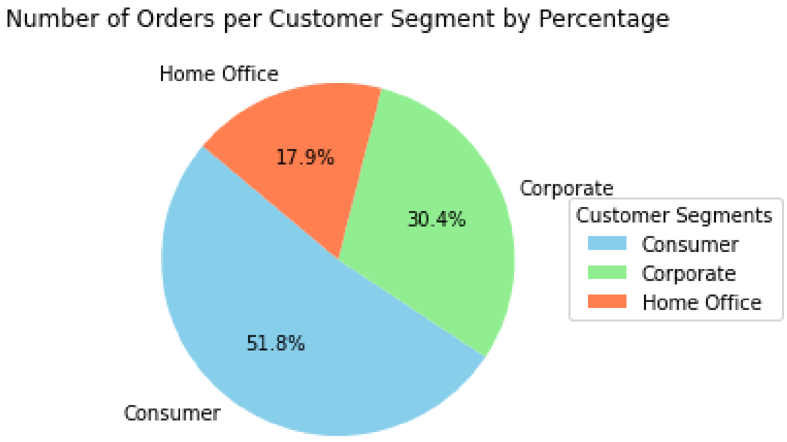


Figure 3: Number of Orders baes on home, corporate and consumer

4.11 Market Analysis

Understanding the market is also very important factor for any business that wants to do long term business. By keeping an eye on what is going on around them companies can make better strategies and they can take good decisions about how to compete with competitors. This research helps in predicting future changes more accurately and giving business to take quick action to fulfil customer demand.

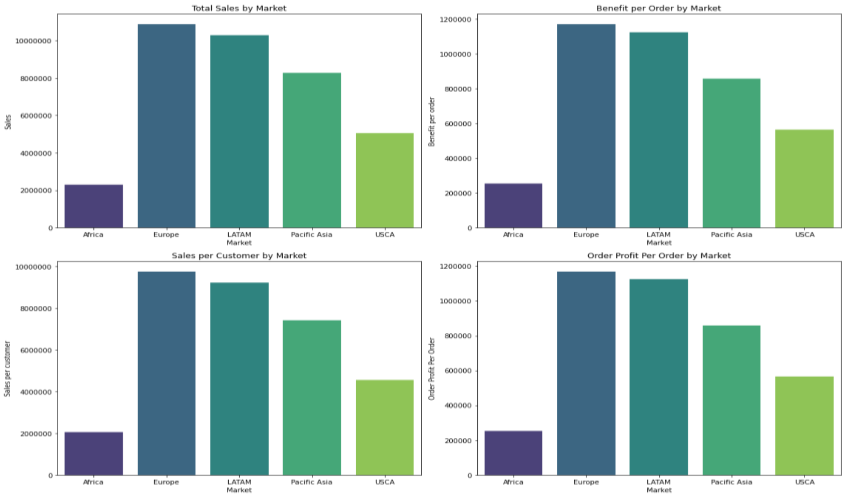


Figure 4: Analysis of market based on sales and benefits by market

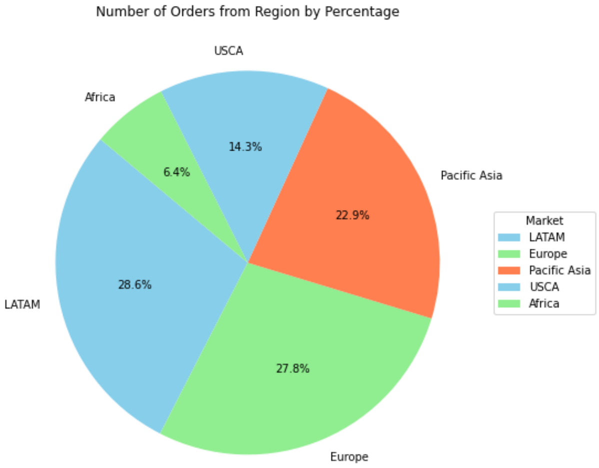


Figure 5: Number of Order by Percentage by market regions

4.12 Product Category

Analysing items individually within a category provides better insights into their sales volume and profitability. This comparison is helping to assess how each item performs relative to others in the same category. Which is giving a clear understanding of their financial impact over the time.

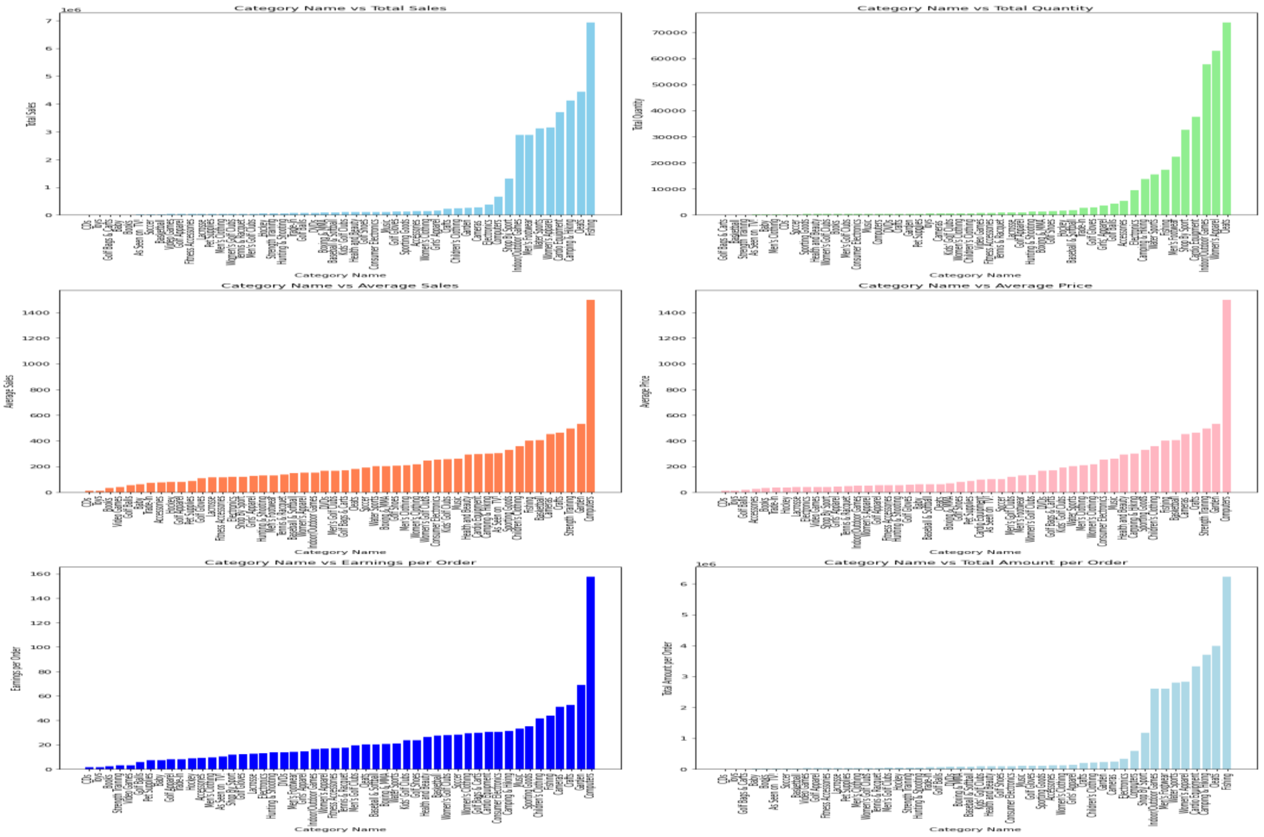


Figure 6: Category of the product based on category name and average price

4.13 Late Delivery and Revenue

A crucial aspect for supply chain managers to monitor the impact of the goods delivery delays on revenue aspects. It is important to understand how late deliveries affect overall revenue of the company and customer satisfaction. This analysis investigates the relationship between revenue which is generated by logistics companies and the frequency of late shipments over a defined period of time. The objective is to determine there is a direct correlation between revenue from the sales of the goods. It is also important to understand provision of services and the time taken to deliver products from the seller to the buyer.

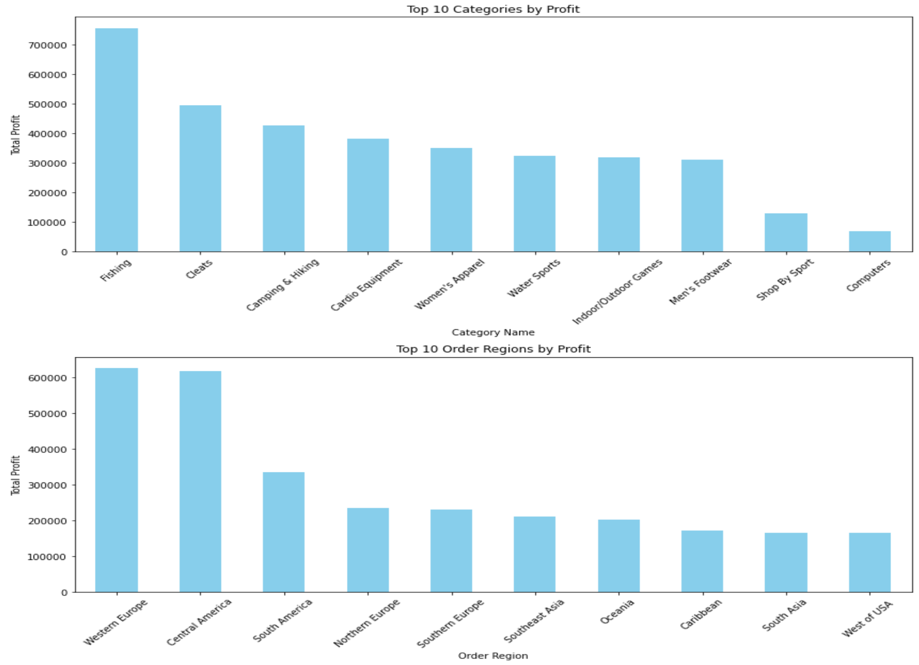


Figure 7: Top 10 profit analysis based on different regions

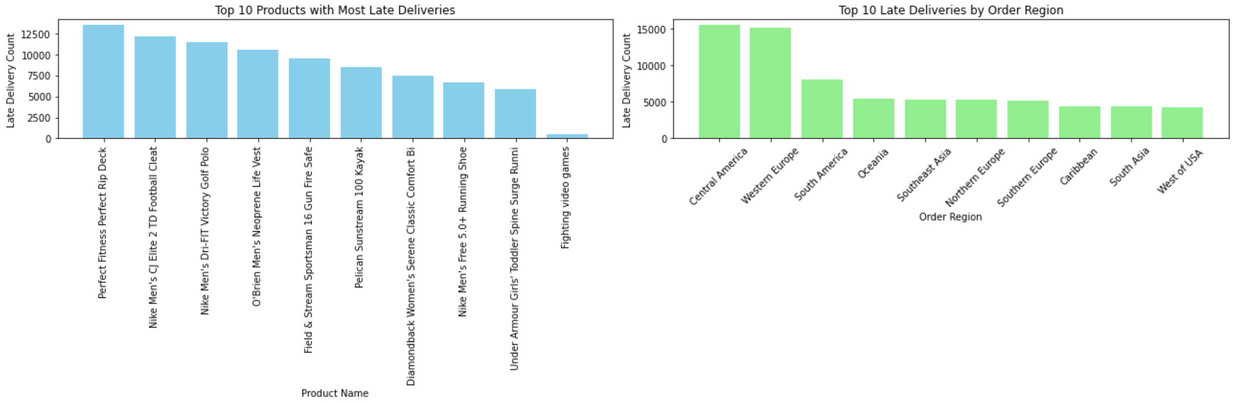


Figure 8: Top 10 Late Deliveries based on different regions

4.14 Delivery Status

This analysis aims to assess the current state of deliveries of from the processing of the orders to its destination. Finding the reasons of delays by comparing key metrics such as time of the order processing transit time and the delivery completion against industry. Real time tracking can enable prompt intervention in the case of delay that will mitigate customer dissatisfaction and it can ensure that goods are delivered as expected time.

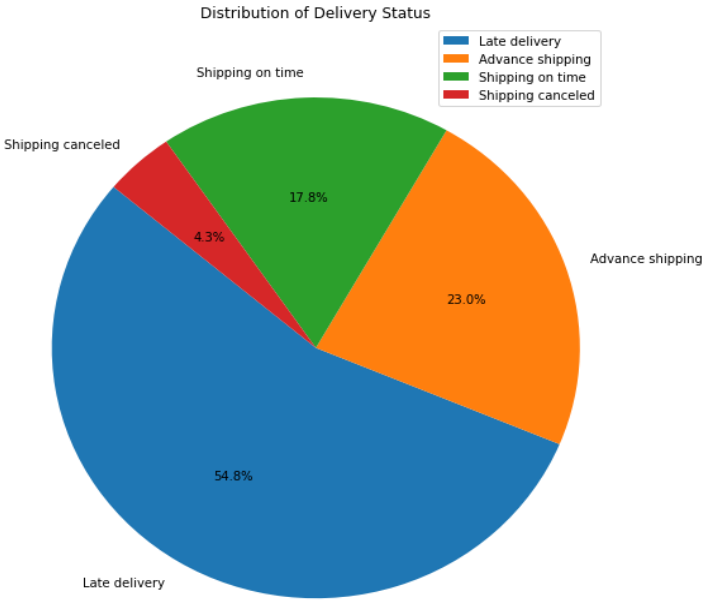


Figure 9: Delivery Status of goods

4.15 Delivery Status by Shipping Mode

Analysing delivery status in different shipping methods can discloses the performance levels of each approach. This visualization provides clarity on how different shipping modes can impact on efficiency of the delivery and its effectiveness which enable more informed evaluation of strengths and weakness.

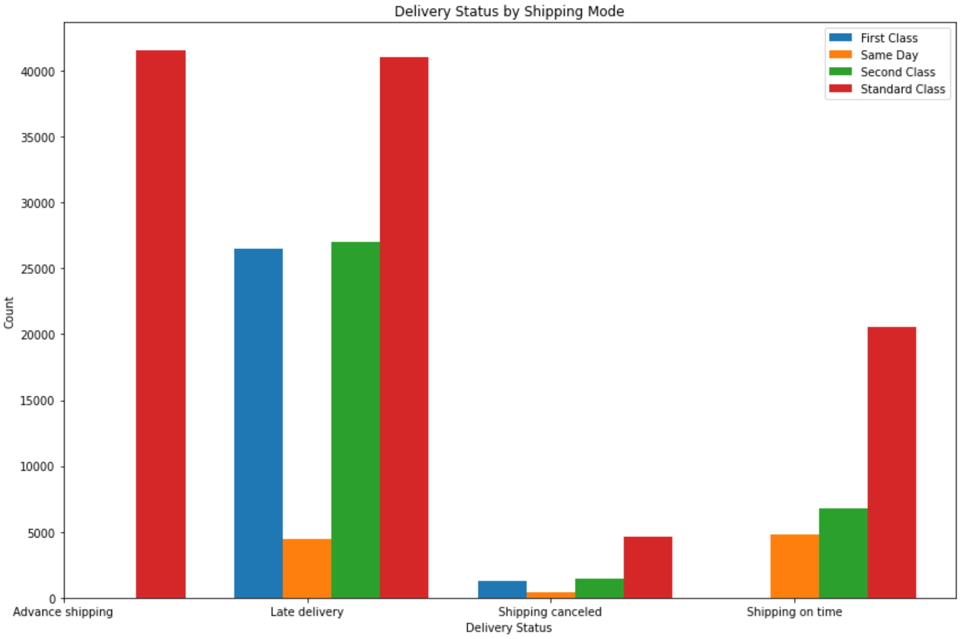


Figure 10: Delivery Status by shipping mode

4.16 Method for the payment

Understanding current trends and consumer preferences in the context of payment methods. It is essential for enhancing the checkout experience and ensuring smooth transactions. This insight allows organizations to optimize payment process for better align with customer expectations and it will improve overall transaction efficiency.

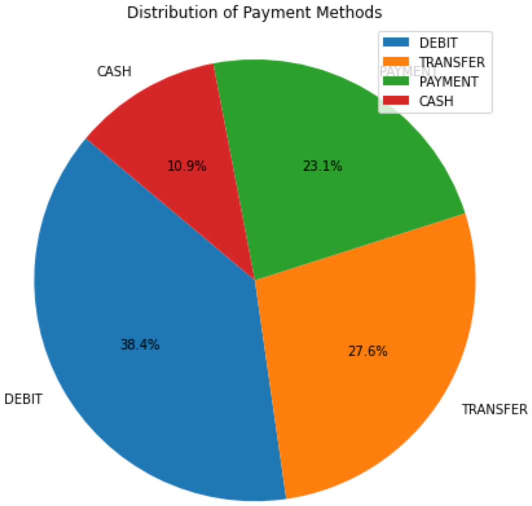


Figure 11: Method for the payment ratio

5 Implementation

5.1 Ordinary Least Squares (OLS):

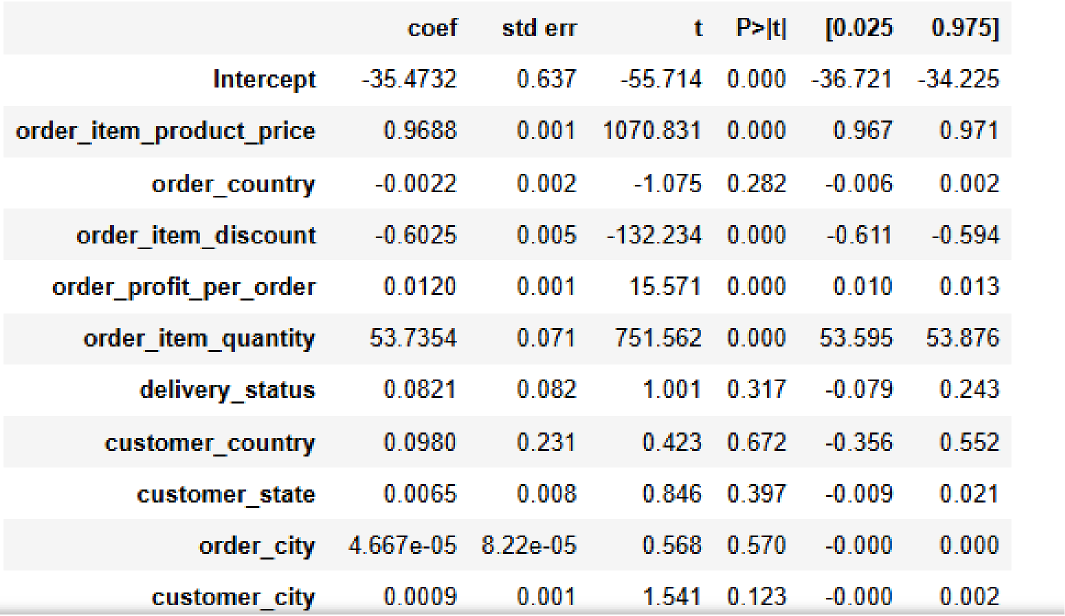
We are using Linear regression for prediction of the Sales by analysing the relationships of different variables. Variables including price, quantity, sales, customer type and market segment. Forecasting accurate sales is a crucial for firms aiming to efficiently meets the consumer demand.

OLS is a statistical technique used to identify best fit that related to predictor variables to a dependent response variable. OLS enables the organizations to make informed predictions and decisions so they can minimize the risk of supply shortages.

5.2 Data Preparation for OLS Regression

Some columns in the dataset contains object data types that cannot use in regression models. One option is to exclude these string columns but it is important to know that variables such as market, category, region, product and category it may have huge impact on 'total amount per order'. Therefore, these categorical variables were converted from object types to integer types before use just to facilitate their inclusion in regression analysis.

Table 1: Regression Coefficient Table



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-35.4732	0.637	-55.714	0.000	-36.721	-34.225
order_item_product_price	0.9688	0.001	1070.831	0.000	0.967	0.971
order_country	-0.0022	0.002	-1.075	0.282	-0.006	0.002
order_item_discount	-0.6025	0.005	-132.234	0.000	-0.611	-0.594
order_profit_per_order	0.0120	0.001	15.571	0.000	0.010	0.013
order_item_quantity	53.7354	0.071	751.562	0.000	53.595	53.876
delivery_status	0.0821	0.082	1.001	0.317	-0.079	0.243
customer_country	0.0980	0.231	0.423	0.672	-0.356	0.552
customer_state	0.0065	0.008	0.846	0.397	-0.009	0.021
order_city	4.667e-05	8.22e-05	0.568	0.570	-0.000	0.000
customer_city	0.0009	0.001	1.541	0.123	-0.000	0.002

Observations:

1. Assessment of P-values for Predictor Variables:

P-values is calculated just to determine statistical significance. Which values are less than 0.05 those were taken in the model because as this indicates that the null hypothesis is rejected.

2. Significant Predictors for:

P-values below 0.05 association with the response variable order item. Predictors include profit per order, product price, product name, category name, item discount, order item and market.

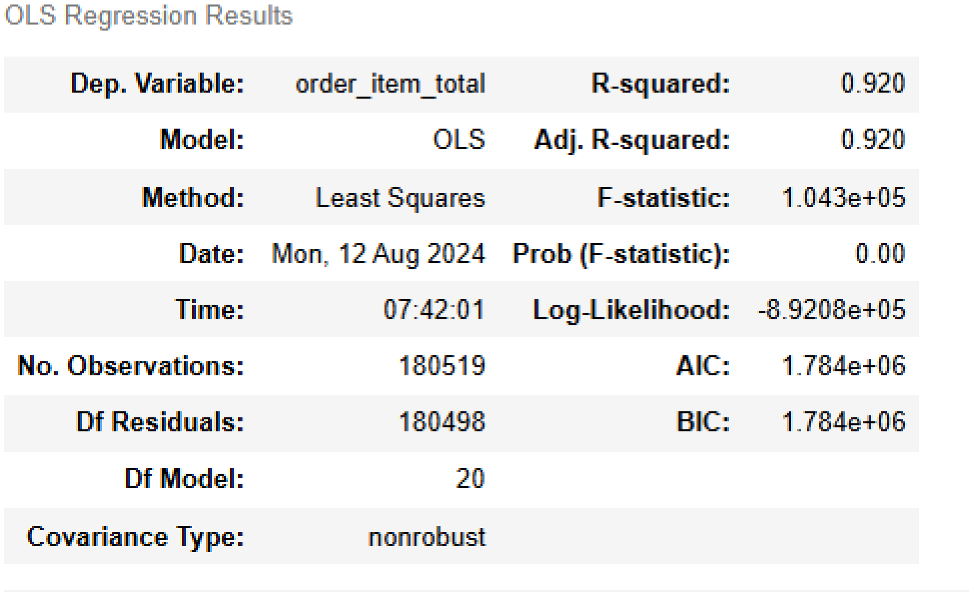
3. Insignificant Predictors:

Predictor variables with p-values exceeding 0.05 are considered to have no effect. Such as order region, country, customer state, order city, customer city, order state, order status, type, shipping mode, delivery status, customer segment.

5.3 Re-calibrating the OLS Regression Model:

The OLS regression model was recalibrated by excluding predictor variables with p-values higher than 0.05. This revision ensures a more robust and interpretable model by focusing on predictors that have a statistically significant impact on the response variable.

Table 2: OLS p-value < 0.05



OLS Regression Results

Dep. Variable:	order_item_total	**R-squared:**	0.920
Model:	OLS	**Adj. R-squared:**	0.920
Method:	Least Squares	**F-statistic:**	1.043e+05
Date:	Mon, 12 Aug 2024	**Prob (F-statistic):**	0.00
Time:	07:42:01	**Log-Likelihood:**	-8.9208e+05
No. Observations:	180519	**AIC:**	1.784e+06
Df Residuals:	180498	**BIC:**	1.784e+06
Df Model:	20		
Covariance Type:	nonrobust		

OLS p-value less than 0.05:

Coefficient (R-squared): Coefficient of determination that is R-square is 0.920 which is indicating that approximately 89.3% of the variance in the dependent variable. 'order_item_total' can be explained by the independent variables in the regression model.

Adjusted R-squared: The adjusted R-squared is 0.920 and which is more predictors variables.

F-statistic: It is showing 1.043e+05. Variables significant at any stage which is showing that those variables, that is explain to a proportion of the variation in the term of order item.

Statistically Significant Predictors: All predictor variables with p-values less than 0.05 are statistically significant in prediction.

Intercept Term: The intercept term with a coefficient of 0.1082 is statistically significant so it suggesting that it is predicting the order total cost per item by itself. This indicates that without considering other factors the intercept is playing a crucial role in the prediction.

Below is the table of the statistically significant predictors:

5.4 Positive Correlation:

Total amount is showing positive correlation with several predictors such as "Customer Segment_Consumer," "Customer Segment_Corporate," "Customer Segment_Home Office," "Sales," "Benefit per order," "Sales per customer," "Order Profit Per Order," and "Order Item Quantity." This means that when one of these variables increase its value then total amount per order also tends to increase on average.

5.5 Negative Correlation:

Negative correlation has some categories for example category, market and the department. If we increase in any factor that can be leads to a decrease in value in another one.

Negative correlations are as follows:

1. Impact on Market:

When more offer given to customer then sale increase which leads to higher figure in the regions with larger percentage discounts. In different European regions and Latin America for most prominent. The significant reduction by sellers in some popular areas explain the negative relationship between these regions.

2. Products by Categories:

Some classes such as fishing gear and cleat shoes experience high volumes sold alongside percentage discounts. When discounts were offered their sale goes high as per expected. When company's markdown its price then more units purchased by customers which is gives the overview of some variables and amount per order.

3. Departmental Analysis:

Some departments sell a large number of products in different categories at lower prices often offer significant discounts. For example, the fan shop department generates large revenue as compare to others that coincide with significant discounts. This shows that different departments can react differently to pricing strategies that is a negative affects the total order amount.

Understanding the behavior of the customer and it is making the strategic decisions according to the company. It means different predictors variables can influence on different variables.

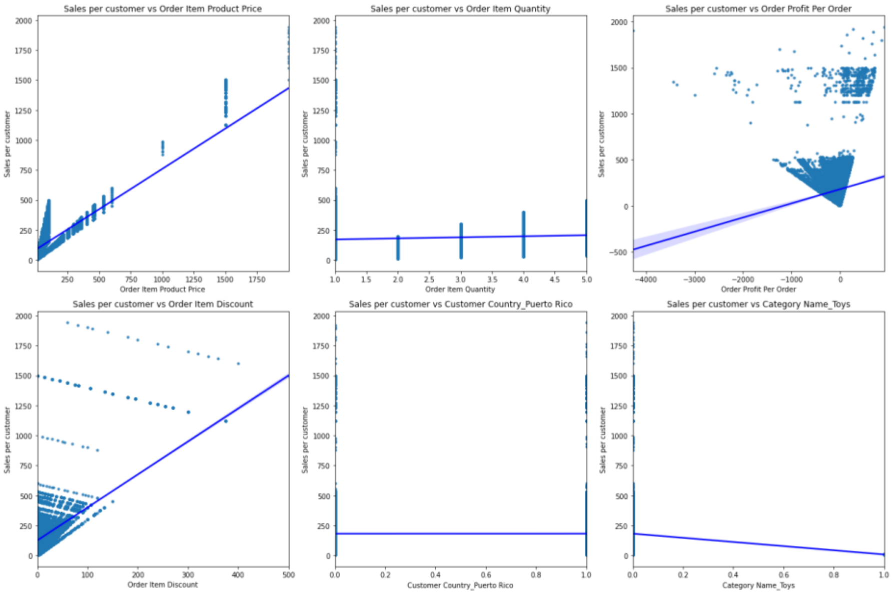


Figure 12: Linear Regression of sales per customer based on different variables

5.6 Model Interpretation:

The regression equation developed to provides insights into how different predictors contribute to the sales. To understand the key drivers behind their revenue. Which can enable them to make informed decisions on price strategies discount policies product assortment and departmental focus.

5.7 Business Perspective:

By analysing the regression model and the corresponding scatter plots we can observe that some variables have a more substantial impact on total sales per customer.

1. Positive Predictors: Some variables `Order Item Product Price`, `Order Item Quantity` and `Order Profit Per Order` have positive correlations with `Sales per customer`. This suggests that increases in these variables generally lead to the higher sales volumes. For example, higher order quantities and greater profit per order can contribute positively to the total sales per customer.

2. Negative Predictors: The model also reveals negative correlations with variables such as market, department name, Item discount, name of the product and the category name. This indicates that higher discounts or some product categories may lead to the lower total sales per customer.

5.8 Order Item Discount: Discounts can leads temporarily boost sales but it may be reduced the overall amount spent per transaction. Therefore, discounts can drive volume so they might not be sustainable for the long-term revenue growth.

5.9 Department and Market: Departments with higher markdowns or some market regions can show a decrease in total sales per customer. Which reflects the impact of aggressive discount strategies?

6 Evaluation

6.1 Regression Models

It’s a crucial part to estimate the quantity of ordered items for demand forecasting of the goods and making informed supply chain decisions. Regression models are instrumental in identifying the trends demands and the relationships among different variables within the datasets. Applying machine learning models into supply chain management can reduce prediction errors by up to 50% according to McKinney company.

In this analysis different regression models were applied. OLS Linear Regression is used to predict order item quantity. The performance of these models is evaluated by using mean absolute error (MAE) and root mean square error (RMSE) which is the key metrics in assessing forecasting accuracy.

Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) is used to measure the quality of regression models in forecasting tasks.

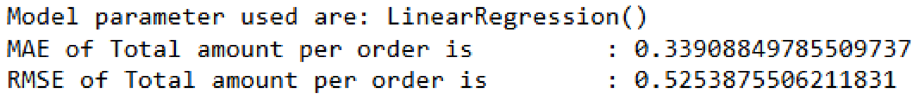
Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Root Mean Square Error (RMSE)

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

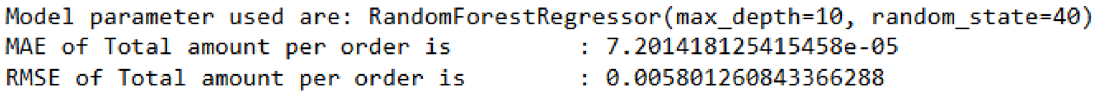
OLS Linear Regression:



```
Model parameter used are: LinearRegression()
MAE of Total amount per order is          : 0.33908849785509737
RMSE of Total amount per order is         : 0.5253875506211831
```

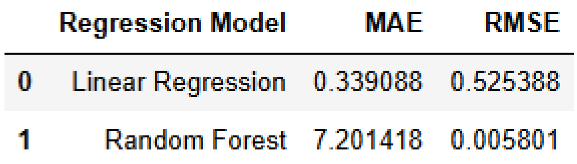
OLS linear regression model shows MAE value 0.3390 and RMSE 0.5253. These values suggest that on average the predictions are close to the actual values.

Random Forest Regressor:



```
Model parameter used are: RandomForestRegressor(max_depth=10, random_state=40)
MAE of Total amount per order is          : 7.201418125415458e-05
RMSE of Total amount per order is         : 0.005801260843366288
```

Results Comparison:



	Regression Model	MAE	RMSE
0	Linear Regression	0.339088	0.525388
1	Random Forest	7.201418	0.005801

Business Perspective:

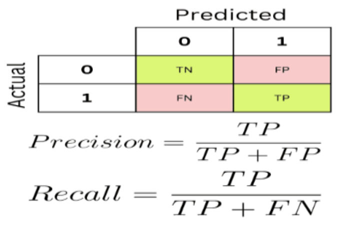
Accurate demand forecasting based on previous sale is crucial in supply chain management. These forecasts playing an important role in guiding strategic decisions related to the manufacturing, purchasing of goods and optimizing capital expenditure.

When organization can accurately forecast future demands then they can enhance overall productivity which is leading to increased profit of the organization. These profits can reinvest to meet different customer needs and desires. By improving customer satisfaction and generating revenue for the organization.

6.2 Classification Model for Late Delivery

We estimated different classification models to predict the late delivery by using metrics such as accuracy, recall and f1 score. These models are very crucial for forecasting of future sale and managing delays, by identifying late delivery reasons organization can take appropriate action before occur.

Confusion Matrix:



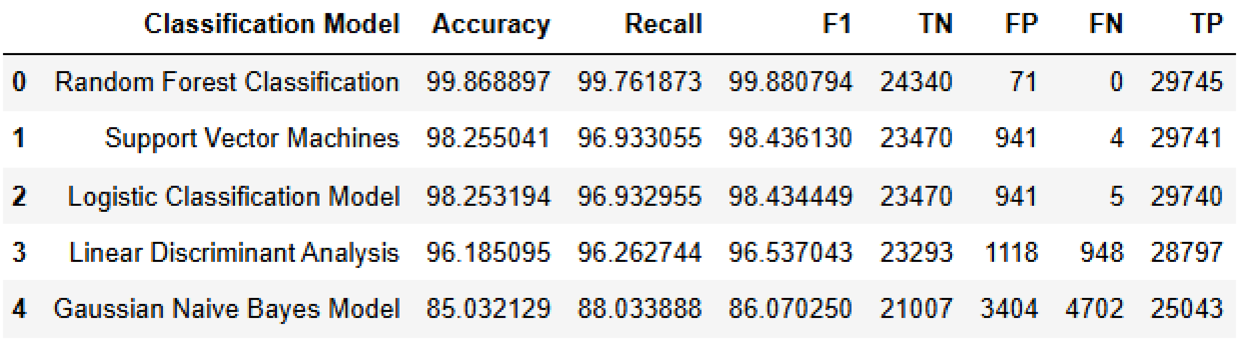
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Model Performance:

We have calculated accuracy, recall, f1 score, True Negatives (TN), True Positives (TP), False Negatives (FN), and False Positives (FP). These counts helping us understand how well each model is performing in predicting late deliveries and on-time deliveries.

Table 3: Classification models result



	Classification Model	Accuracy	Recall	F1	TN	FP	FN	TP
0	Random Forest Classification	99.868897	99.761873	99.880794	24340	71	0	29745
1	Support Vector Machines	98.255041	96.933055	98.436130	23470	941	4	29741
2	Logistic Classification Model	98.253194	96.932955	98.434449	23470	941	5	29740
3	Linear Discriminant Analysis	96.185095	96.262744	96.537043	23293	1118	948	28797
4	Gaussian Naive Bayes Model	85.032129	88.033888	86.070250	21007	3404	4702	25043

7 Conclusion and Future Work

This study is focused on supply chain management which leads to customer satisfaction by using dataset that is called mendeley data. We have implemented different regression and classification models and data analysis techniques just to find critical insights and company operations.

Ordinary least squares (OLS) regression highlights there are some factors which effect on buying the products. Coefficient, standard error calculated which are giving insights on discount, price of the product and how many products are purchased. According to these findings companies can make strategies of product pricing and they can make better supply chain process.

Classification models such as random forest, Support vector machine, logistic regression, linear discriminant analysis and Gaussian naïve bayes implemented just to check late deliveries of the shipment and then accuracy, recall and f1 values calculated and confusion matrix highlights about which product is late or not.

Based on the results of the models, companies can easily make pricing strategies according to the need of the market. It will also help to optimize the process of supply chain management based on these factors customer satisfaction will increase.

Further research can conduct in this field of supply chain management there are many factors can involve to streamline the overall process. For example, country market trends, economic condition of the country and special occasion of any specific country can increase the demand of the product and also politically stability can impact on the deliveries.

In conclusion this study highlights issues in supply chain management. Companies can optimize their delivery process and can gain customer satisfaction. They can predict future demands of the goods based on the sale of the previous sale of the products and they can manage inventory based on these results. In today’s world of highly uncertainty, companies can take help of data analytics and machine learning for making best decisions.

References

Husna, A., Amin, S.H. and Shah, B., 2021. Demand forecasting in supply chain management using different deep learning methods. In Demand forecasting and order planning in supply chains and humanitarian logistics (pp. 140-170). IGI Global.

Cadavid, J.P.U., Lamouri, S. and Grabot, B., 2018, July. Trends in machine learning applied to demand & sales forecasting: A review. In International conference on information systems, logistics and supply chain.

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

Carbonneau, R., Laframboise, K. and Vahidov, R., 2008. Application of machine learning techniques for supply chain demand forecasting. European journal of operational research, 184(3), pp.1140-1154.

Singha, D. and Panse, C., 2022, February. Application of different machine learning models for supply chain demand forecasting: comparative analysis. In 2022 2nd international conference on innovative practices in technology and management (ICIPTM) (Vol. 2, pp. 312-318). IEEE.

Zohdi, M., Rafiee, M., Kayvanfar, V. and Salamiraad, A., 2022. Demand forecasting based machine learning algorithms on customer information: an applied approach. International Journal of Information Technology, 14(4), pp.1937-1947.

Ampazis, N., 2015. Forecasting demand in supply chain using machine learning algorithms. International journal of artificial life research (IJALR), 5(1), pp.56-73.

Kilimci, Z.H., Akyuz, A.O., Uysal, M., Akyokus, S., Uysal, M.O., Atak Bulbul, B. and Ekmis, M.A., 2019. An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. Complexity, 2019(1), p.9067367.

Sengar, R.S., Ahmed, D.F. and SIRT, B., 2019. Review on Trends in Machine Learning Applied to Demand & Sales Forecasting. SMART MOVES JOURNAL IJOSCIENCE, 5(6), p.4.

Zhu, X., Ninh, A., Zhao, H. and Liu, Z., 2021. Demand forecasting with supply-chain information and machine learning: Evidence in the pharmaceutical industry. Production and Operations Management, 30(9), pp.3231-3252.

El Filali, A., Lahmer, E.H.B., El Filali, S., Kasbouya, M., Ajouary, M.A. and Akantous, S., 2022. Machine Learning Applications in supply chain Management: A Deep Learning Model Using an Optimized LSTM Network for Demand Forecasting. International Journal of Intelligent Engineering & Systems, 15(2).

Mahraz, M.I., Benabbou, L. and Berrado, A., 2022. Machine learning in supply chain management: A systematic literature review. International Journal of Supply and Operations Management, 9(4), pp.398-416.

Wiyanti, D.T., Kharisudin, I., Setiawan, A.B. and Nugroho, A.K., 2021, June. Machine-learning algorithm for demand forecasting problem. In Journal of Physics: Conference Series (Vol. 1918, No. 4, p. 042012). IOP Publishing.

Carbonneau, R., Vahidov, R. and Laframboise, K., 2007. Machine learning-based demand forecasting in supply chains. International journal of intelligent information technologies (IJIIT), 3(4), pp.40-57.

Vairagade, N., Logofatu, D., Leon, F. and Muharemi, F., 2019. Demand forecasting using random forest and artificial neural network for supply chain management. In Computational Collective Intelligence: 11th International Conference, ICCCI 2019, Hendaye, France, September 4–6, 2019, Proceedings, Part I 11 (pp. 328-339). Springer International Publishing.

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), p.1476043.

Mediavilla, M.A., Dietrich, F. and Palm, D., 2022. Review and analysis of artificial intelligence methods for demand forecasting in supply chain management. Procedia CIRP, 107, pp.1126-1131.

Saha, P., Gudheniya, N., Mitra, R., Das, D., Narayana, S. and Tiwari, M.K., 2022. Demand forecasting of a multinational retail company using deep learning frameworks. IFAC-PapersOnLine, 55(10), pp.395-399.

Arif, M.A.I., Sany, S.I., Nahin, F.I. and Rabby, A.S.A., 2019, November. Comparison study: product demand forecasting with machine learning for shop. In 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 171-176). IEEE.

Terrada, L., El Khaili, M. and Ouajji, H., 2022. Demand forecasting model using deep learning methods for supply chain management 4.0. International Journal of Advanced Computer Science and Applications, 13(5).

Makkar, S., Devi, G.N.R. and Solanki, V.K., 2020. Applications of machine learning techniques in supply chain optimization. In ICICCT 2019–System Reliability, Quality Control, Safety, Maintenance and Management: Applications to Electrical, Electronics and Computer Science and Engineering (pp. 861-869). Springer Singapore.

Bousqaoui, H., Achchab, S. and Tikito, K., 2019. Machine learning applications in supply chains: Long short-term memory for demand forecasting. In Cloud Computing and Big Data: Technologies, Applications and Security 3 (pp. 301-317). Springer International Publishing.

Mitra, A., Jain, A., Kishore, A. and Kumar, P., 2022, September. A comparative study of demand forecasting models for a multi-channel retail company: a novel hybrid machine learning approach. In Operations research forum (Vol. 3, No. 4, p. 58). Cham: Springer International Publishing.

Mohamed-Iliasse, M., Loubna, B. and Abdelaziz, B., 2020, October. Is machine learning revolutionizing supply chain?. In 2020 5th International Conference on Logistics Operations Management (GOL) (pp. 1-10). IEEE.

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, pp.116013-116023.