IMPLEMENTATION OF AI FOR UNDERSTANDING SUPPLIER DISRUPTION

CAPSTONE PROJECT REPORT SUBMITTED

TO THE

INDIAN SCHOOL OF BUSINESS

FOR THE

ADVANCED MANAGEMENT PROGRAMME IN BUSINESS ANALYTICS

SUBMITTED BY

AADRIKA V [12110055]

MANJARI SHRIVASTAVA [12110082]

SHASHANK JHA [12110023]

VINAYAK C BURJI [12110056]

UNDER THE GUIDANCE OF

PROF. PEEYUSH TAORI

INSTITUTE OF DATA SCIENCE

INDIAN SCHOOL OF BUSINESS

DEC 2022

CAPSTONE PROJECT REPORT

SUBMITTED TO THE INDIAN SCHOOL OF BUSINESS FOR ADVANCED MANAGEMENT PROGRAMME IN BUSINESS ANALYTICS

HEADING	DETAILS
Title of the Capstone project	**Implementation of AI for understanding Supplier Disruption**
Name of the Candidate	**Aadrika V [12110055]** **Manjari Shrivastava [12110082]** **Shashank Jha [12110023]** **Vinayak C Burji [12110056]**
Name and Affiliation of the Faculty Mentor	**Prof. Peeyush Taori, Assistant Professor, HKU Business School, PhD (LBS), M.Res (LBS)**
Name and Affiliation of the Industry Mentor	**Dr. Sachin Agarwal, Principal Engineer Scientist, Head of AI, Sony Research India Private Limited**
Batch	**AMPBA 2022 (Summer)**
Date/Month of submission of Capstone Project Report	**DEC 2022**

Acknowledgement

First and foremost, we would like to express our sincere thanks to ISB and IIDS for providing us with an opportunity to work on this capstone project. We are extremely thankful and indebted to Dr. Sachin Agarwal for sharing his expertise, valuable guidance, and continuous encouragement with our team. We would also like to thank Prof. Peeyush Taori for his dedicated involvement and attention to our team during the feasibility study. It wouldn’t have been simpler without his guidance and support.

We are thankful to Shreya Singireddy for her prompt response and guidance throughout the project.

Finally, we must express our profound gratitude to our family for giving us support and continuous encouragement through the process of the capstone project. This accomplishment would not have been possible without them.

Aadrika V [12110055]
Manjari Shrivastava [12110082]
Shashank Jha [12110023]
Vinayak C Burji [12110056]
DEC 2022

Table of Contents

List of Tables

List of Figures

Abstract

Due to the global economic uncertainty that occurred during the 2008 financial crisis and the 2020 COVID pandemic, the increasing number of analysts and researchers focused on the management of supply chain risks has become more prevalent.

Despite the increasing popularity of data analytics in the industry, there is still a lack of case studies that show how it can be used to predict the impact of supply chain disruptions. Artificial Intelligence (AI) has been gaining popularity in the past couple of years, and its potential to improve the efficiency and effectiveness of supply chain management has led to the development of new machine learning techniques. This study aims to provide a comprehensive overview of the various aspects of data analytics using machine learning and how it can be used to predict the impact of supply chain disruptions such as late delivery risk.

In our study, we discuss the implementation of Artificial Intelligence (AI) to understand the impact of late delivery risk on different customer segments. Along with it, we will also assess the statistical significance of different AI models using hypothesis testing. To measure the impact of sales, we will explore time series analysis.

There will also be a theoretical assessment to understand the practical business application of the chosen AI model using machine learning.

Keywords:
Machine learning, AI, supply chain risk, data analytics, supply chain disruptions, time series analysis, and hypothesis testing.

Executive Summary

The complexity of today's supply chains has become more apparent. The outbreak of the coronavirus pandemic has highlighted the inadequacies of the global supply-chain system. The rapid emergence and evolution of new products and the increasing number of linked physical flows have created a need for more flexible and agile management. Market volatility and the COVID-19 pandemic have also raised concerns about the environment. This has prompted companies and stakeholders to focus on the importance of supply-chain resilience.[7]

It also highlighted the potential of artificial intelligence to improve the efficiency of operations. The business objective of this problem is to increase the benefit per order and to decrease the variance between supply and demand through Predictive Analytics. Here in our scope of the project, we have analyzed the data from an e-commerce company called DataCo Global. We have started with exploratory data analysis to understand the impact of the global supply chain and the countries getting affected. From EDA we can see great insights, which further has been used while building a model. We have looked at the effect of shipping modes on delays, Type of payments on delays, Customer country vs order country effect on delays along with the year and month of shipping also been considered as the weather affects the marine supply chain.

Post regular data cleaning methods, we have split the data into train and test for validation purposes. The train data set has been fed into multiple models such as SVM, Linear Regression, XB-Booster, Decision Tree classifier, etc. and with Evaluation parameters of Accuracy, F1 Score, and Recall, we have chosen decision tree classifier as the final model for predicting the delays. Further, we have done Statistical analysis to understand the confidence and interval and model error (P-value) and Hyperparameter tuning to understand the depth, nodes, and criterion used. We have used Gradio for deployment purposes as it has a user-friendly interface and suffices the model requirement.

Recommendations:

1: Setting up a Supply chain Control tower which will help the planners to predict delays for Products based on Historical Sales data and Delay frequencies. Also, Opportunity analysis is to be done based on Lane level data (Customer Country- order Country) to understand if the product can be Rerouted using other closer Shipping Locations.

2: Inventory management using ABC analysis should be done to understand which products contribute to a larger portion of our revenue and therefore try to prioritize those products in our warehouses to reduce customer churn as well as reduce the overall cost of production.

Future Scope:

1: Customer Segmentation based on RFM analysis could be done. We could do recency, frequency, and monetary-based segmentation of the customers that order from DataCo Global to strategize according to these customers' buying patterns.

2: Forecasting for the coming years based on the current demand and buying patterns.

3: Simulation of data for the coming years to predict late deliveries.

Chapter 1: Introduction

Due to the rise of e-commerce sites, the demand for supply chain logistics has increased. To keep up with the changes, the entire supply chain network must be continuously monitored. The increasing complexity of today's global supply chains and the increasing number of factors that can affect them are forcing companies to rethink their operations and strategies. Therefore, they must adopt effective measures to minimize the impact of their disruptions.

One of the main goals of a shipping company is to ensure that the cargo is in good condition before it is delivered to its destination. Using Artificial Intelligence (AI), businesses can prevent supply chain disruptions.

With AI, businesses can improve their efficiency and prevent supply chain disruptions. It can help them predict delays, manage their production costs, and respond quickly to any issues related to the distribution chain.

Chapter 2: Review of Related Literature

Several studies published in the past couple of years have highlighted the increasing use of data analytics in the supply chain. Wang et al. (2016) and Baryannis et al. (2019) identified the three main types of big data analytics that are commonly used in this field: prescriptive, predictive, and descriptive (Brintrup, Alexandra Melike., Pak, Johnson., Ratiney, David., Pearce, Tim., November 2019) [1,5,6]. Descriptive analytics can help one identify the current situation in a supply chain. Examples of supply chain metrics that can provide historical insight are those related to the company's operational and financial performance. (Brintrup, Alexandra Melike., Pak, Johnson., Ratiney, David., Pearce, Tim., November 2019 [6]). These include the total stock in inventory, the average money spent per customer, and the year-to-year changes in sales (Brintrup, Alexandra Melike., Pak, Johnson., Ratiney, David., Pearce, Tim., November 2019) [6]. Predictive analytics uses advanced statistical techniques to predict the future situation of a company's operations (Brintrup, Alexandra Melike., Pak, Johnson., Ratiney, David., Pearce, Tim., November 2019) [6]. Forecasting customers' purchasing patterns along with social media analysis and forecasts can be good examples of Predictive Analytics (Brintrup, Alexandra Melike., Pak, Johnson., Ratiney, David., Pearce, Tim., November 2019) [6]. Through prescriptive analytics, a company can improve its current situation. For instance, by analyzing the volatility of its products, it can identify areas of its operations where it can improve its efficiency. It can also help prevent its warehouses from experiencing excessive inventory levels by analyzing the data collected from its customers (Brintrup, Alexandra Melike., Pak, Johnson., Ratiney, David., Pearce, Tim., November 2019)[6].

Due to the increasing number of AI applications being used in the supply chain, more companies are now turning to these tools to address the challenges of their operations. According to the 2022 Warehousing Education and Research Council (WERC) Report, the top 12 metrics reported by warehouse experts are Average warehouse capacity used, Order-picking accuracy, on-time shipments, On-time ready to-ship, Peak warehouse capacity used, Dock-to-stock cycle time(in hours), percent of supplier orders received damaged free, orders with on-time delivery, Fill rate-order, Shipped damaged free (outbound), Fill rate-line and Shipped complete per customer order (Hyster-Yale Group, Inc. 2022[4]).

The likelihood of experiencing moderate or high-impact disruptions is usually associated with various factors such as unexpected demand, supply constraints, and company buyouts (Scheibe and Blackhurst 2018). Big data has the potential to help predict supply chain disruptions, though there have been no reported cases of its use.

This report aims to provide a case study that demonstrates how machine learning and data analytics can be used to predict disruptions in the supply chain. The first step in the process is to identify the potential features that can be used to improve the accuracy of the predictions. Then, a performance metric is developed to measure the effectiveness of the methods. The next step involves analyzing the performance of the various algorithms and their parameter sets. This process is carried out through experimental and statistical processes using hypothesis testing. It then highlights the importance of domain knowledge in developing effective engineering features.

Chapter 3: Project Description

3.1 Business/Research Problem

3.1.1 Business Problem

Global Supply chains are currently under the strain of unprecedented demand and constricted effective logistics capacity. On average, global container shipping rates have more than quadrupled since 2019, and schedule delays have risen. In some key trading routes, such as Asia–Europe and Asia–North America, the rate spikes are even higher, and the delays are more frequent.

One of the main factors that have caused the complexity of today's global supply chains is the implementation of just-in-time manufacturing (JIT). This process has allowed companies to reduce their inventory levels and improve their efficiency (Melendez, Carlos., 2020, Oct 28 [8]).

Unfortunately, implementing just-in-time manufacturing has been very challenging when a natural disaster or pandemic occurs, which can result in a spike in demand. Therefore, companies must adopt a new model that enables them to manage their supply chains in real-time (Melendez, Carlos., 2020, Oct 28 [8]).

The major issues causing the disruptions include labor shortages, equipment availability, unpredictable climate situation, and the ripple effect of global bottlenecks. Therefore, we need to Understand and develop an optimized Logistics Strategy in terms of people, processes, and Technology to curb the effect of Global Supply chain Disruptions. (Katsaliaki,K, Galetsi,P., Kumar, S. Jan 2021 [3])

From the analysis of Madenas et al., (2014), it is clear that the manufacturing phase, product development, and product lifecycle are largely separated. The service phase also requires a different level of information from the supply chain. Nevertheless, due to a lack of information flow within the system of supply chain management, the overall operation is not able to flow smoothly. A different and separate level of information is required in each phase of the product lifecycle, which is mainly due to the rising complexities of modern business. Hence, lack of information flow can be a major problem within the operation department and the benefit of predictive modeling in the Supply chain can make a significant positive difference in smoother information flow and help businesses to take productive decisions.

3.1.2 Area of Impact

Supply chain leaders in different industries, including the technology and manufacturing sector, conduct business with thousands of worldwide suppliers. As a result, different suppliers' problems ranging from material shortages to legal investigations could crop up, leading to disruptions in the supply chain.

Especially, Maritime transport is crucial for international trade, as it carries over 90% of the world's goods. As per World Economic Forum statistics, approximately 60% of this is transported in containers.

The average reliability of container shipping lines has historically hovered around the 66% mark, implying that only 2 in 3 vessels arrive as per schedule; a number that would be considered unacceptable in most other industries. Over the past year, in the Covid and post-Covid scenario, even this low benchmark has become a pipe dream, with the container shipping industry's overall schedule reliability levels having fallen through the floor, recording an abysmal 33% in August 2021.

Several instances cause the delays such as Weather, Peak Season, Blank Sailing, Port call Omissions, Labour Shortage/unrest, Congestion at Ports, Holdup Delays/customs, etc. But an accurate information flow within the supply chain can reduce huge financial leakages within the business. A few of the major business consequences can be seen below

1) Increase in TCO: The delays increase the overall cost of ownership, on account of the higher charges incurred due to delays

2) Delayed sales and manufacturing: if the container contains raw material destined for a manufacturing plant, production can be held up if delivery is delayed. Likewise, in the case of a container carrying finished goods, the sale to the importer or final consumer will be delayed.

3) Money tied up in inventory and lower cash flows: The more the container is delayed, the longer it will take to liquefy the cargo, which means more money tied up in inventory and slower cash flows.

4) Increased risk of damage, pilferage, spoilage, and obsolescence: The longer cargo stays in transit, the more it is exposed to the risk of damage or pilferage. In the case of time-sensitive cargo, such as perishables, delays could reduce the value of the cargo and hence its selling price.

5) Lower asset turnover: The longer it takes to deliver a container, the fewer times it can be turned over, indicating suboptimal usage of assets. The same holds for the cargo carried in the container.

3.1.3 Business Objective

The business objective of this problem is to increase the benefit per order and to decrease the variance between supply and demand.

Benefits of using Predictive Analytics in the Supply Chain:

- Demand Forecasting
- Predictive Maintenance
- Supply Chain Visibility
- Price Optimization
- Reduce Downtime and Prevent Defects
- Increase in NPS score

3.2 Data Available

In our study, the dataset used for this problem is downloaded from Kaggle whose time range of data spans from ***2015 – 2018***. It contains ***180,519*** records. The source of this data is an e-commerce company called DataCo Global. The features of the dataset include information about Customers, Stores, Products, Sales, Shipping details, Orders, and late delivery risk. A complete list of features and dimensions available in the dataset along with a description is mentioned in Table 3.2 below.

Table 3.2: Overview of Data

Fields	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
Benefit 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	The categorical variable that indicates if sending is late (1), it is not late (0).
Category Id	Product category code
Category Name	Description of the product category
Customer City	City where the customer made the purchase
Customer Country	Country where the customer made the purchase
Customer Email	Customer's email
Customer Fname	Customer name
Customer Id	Customer ID
Customer Lname	Customer last name
Customer Password	Masked customer key
Customer Segment	Types of Customers: Consumer, Corporate, Home Office
Customer State	State to which the store where the purchase is registered belongs
Customer Street	Street to which the store where the purchase is registered belongs
Customer Zipcode	Customer Zipcode
Department Id	Department code of the store
Department Name	Department name of the store
Latitude	Latitude corresponds to the location of the store
Longitude	Longitude corresponding to the location of the store
Market	Market to where the order is delivered: Africa, Europe, LATAM, Pacific Asia, USCA
Order City	Destination city of the order
Order Country	Destination country of the order
Order Customer Id	Customer order code
order date (DateOrders)	The date on which the order is made
Order Id	Order code
Order Item Cardprod Id	Product code generated through the RFID reader
Order Item Discount	Order item discount value
Order Item Discount Rate	Order item discount percentage
Order Item Id	Order item code
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	The total amount per order
Order Profit Per Order	Order Profit Per Order
Order Region	Region of the world where the order is delivered: Southeast Asia, South Asia, Oceania, Eastern Asia, West Asia, West of USA, US Centre, West Africa, Central Africa, North Africa, Western Europe, Northern, Caribbean, South America, East Africa, Southern Europe, East of USA, Canada, Southern Africa, Central Asia, Europe, Central America, Eastern Europe, South of USA
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 Card Id	Product code
Product Category Id	Product category code
Product Description	Product Description
Product Image	Link of visit and purchase of the product
Product Name	Product Name
Product Price	Product Price
Product Status	Status of the product stock: If it is 1 not available, 0 the product is available
Shipping date (Date Orders)	Exact date and time of shipment
Shipping Mode	The following shipping modes are presented: Standard Class, First Class, Second Class, Same Day

Below are the features given in the data:

After removing duplicated and missing data, it was revealed that the dataset contained over 180,519 orders of 118 distinct products, with 55% of them being delivered late. The data was then analyzed, and it was revealed that the average expected shipment is 3 days, whereas the average shipment days for delivery is 3.5. There are ***51 product categories***, the maximum ordered categories being cleats, men's footwear, women's apparel, indoor/outdoor games, and fishing. Under these categories, there are ***118 distinct products*** ordered.The average expected shipment is 3 days, whereas the average shipment days for delivery is 3.5. There are 4 different shipping modes, same day, first class, second class, and standard. The costs associated with these are $10, $7, $4, and $3, respectively.

Within customer segments, ~52% is the Consumer Segment, 30% is the Corporate Segment and the rest ~18% is the home Office segment. In actual terms, 99% of the deliveries are shipped using standard class whereas only 0.05% of the deliveries were same-day deliveries.

In relative terms, in both the First and Second classes, the late delivery risk is more likely. The late delivery risk is 95% in the first class and 77% in the second class. Whereas in the standard class, the risk of late delivery is ~38%. This is also lower than the Same Day shipping mode where the risk is ~46%.

There are only 2 countries from where the orders are placed, the United States and Puerto Rico.

Chapter 4: Assumptions, Approach & Process

4.1 Assumptions

Post research and doing a review of related literature, we did a feasibility study to understand the scope of the project.

4.1.1 Scope of the project

Our business scope will be limited to assessing late delivery risk and categorizing delivery as late delivery. This will help the end user to flow accurate information through the supply chain and help in reducing unexpected leakages and inventory loss.

4.1.2 Business exclusions of the project

After going through various datasets, we concluded that adding the other data columns related to product damage would be a difficult task, given the timeframe of this project. Hence, we limited the scope to only assessing late delivery risk.

Along with it, creating data manually for columns such as product damage would lead to biases in the dataset which we were suggested to avoid.

4.2 Approach

In this project, the Agile methodology is followed to track each deliverable, breaking it into user stories, features, and tasks.

Business Understanding was done by determining business objectives by doing an in-depth literature review. This was done by analyzing research papers, articles, blogs, journals, reports, case studies, etc. using the bibliometric technique [3]. The situation was assessed by doing a feasibility study and determining analytical outcomes using the reverse engineering process. Along with it, IT infrastructure was also assessed, and its feasibility was considered as well.

Data understanding will be implemented by doing data collecting relevant data from trusted sources. The data will then be cleaned by doing data imputation and noise reduction.

Normalization and one hot encoding will be done as part of data standardization. Feature engineering will be done to introduce derived variables from the existing data set. Missing data and different levels of granularity within the data will be identified. Finding such as pattern detection and correlations can be derived from Exploratory Data Analysis.

The dataset will then be split into training and test datasets so that suitable modeling techniques can be selected, and the model can be assessed based on the application of the selected modeling technique.

Post regular data cleaning methods, we have split the data into train and test for validation purposes. The train data set has been fed into multiple models such as SVM, Linear Regression, XB-Booster, Decision Tree classifier, etc. and with Evaluation parameters of Accuracy, F1 Score, and Recall, we have chosen decision tree classifier as the final model for predicting the delays. Further, we have done Statistical analysis to understand the confidence and interval and model error (P-value) and Hyperparameter tuning to understand the depth, nodes, and criterion used. We have used Gradio for deployment purposes as it has a user-friendly interface and suffices the model requirement.

Chapter 5: Exploratory Data Analysis

Exploratory Data Analysis is done using Seaborn and Matplotlib libraries along with getting insights using Tableau visualizations. Below are notable insights.

5.1 Notable Insights

5.1.1 Delivery Status by Shipping Mode Analysis

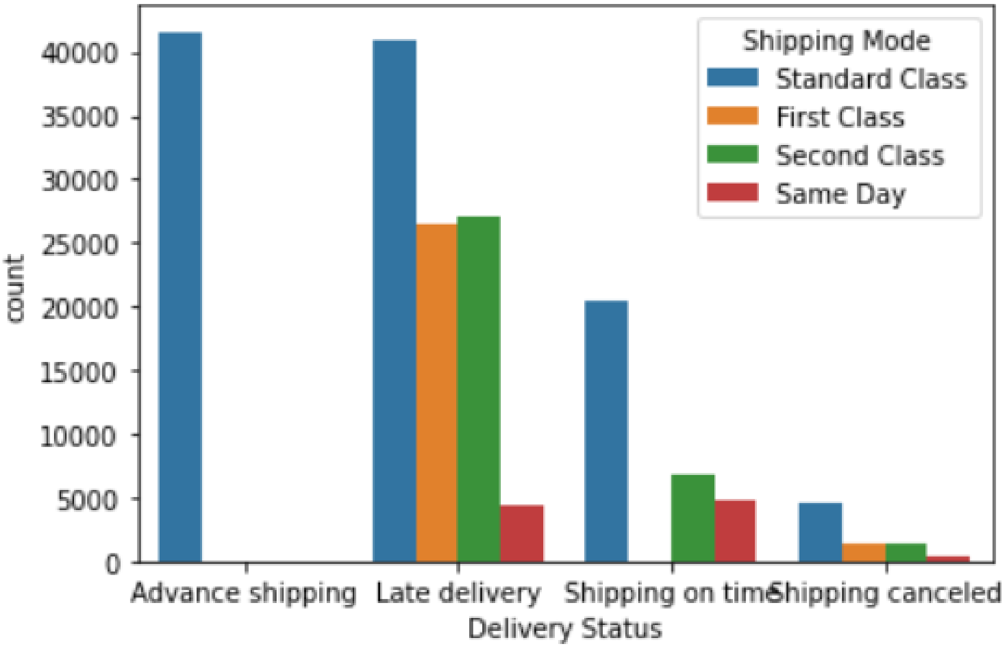


Fig 5.1.1 Delivery Status by Shipping Mode

It is observed that shipping for Standard Class has generally delivery status as Advance Shipping and late delivery. Most of the same-day deliveries are canceled. In general, all shipping modes have late delivery risks.

5.1.2 Sales by Market Analysis

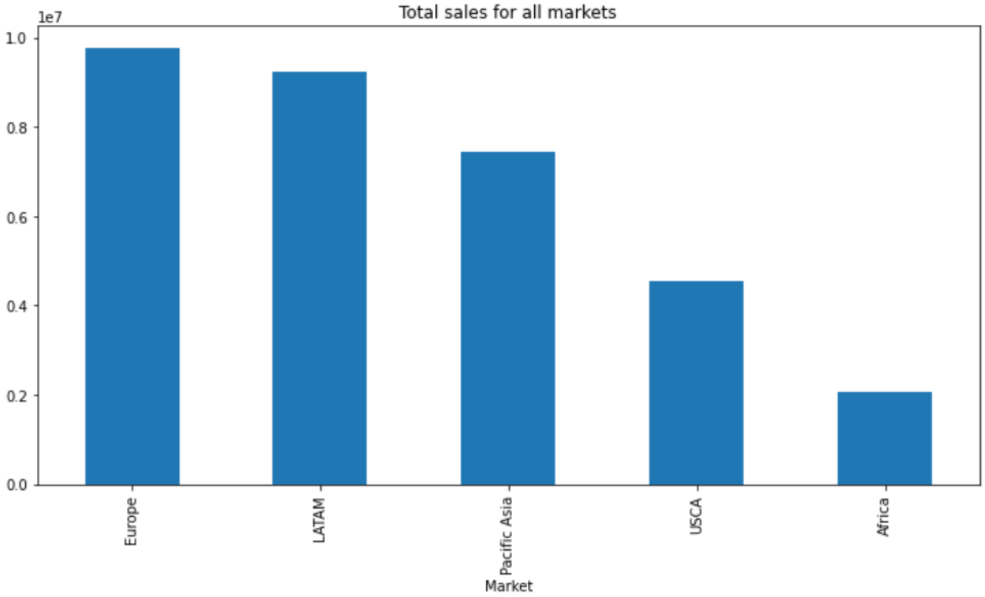


Fig 5.1.2 Sales by Market

From the above graph, we observe that Europe followed by Latin America has the highest sales among all the markets with Africa having the lowest sales.

5.1.3 Top 10 Products with the most late deliveries

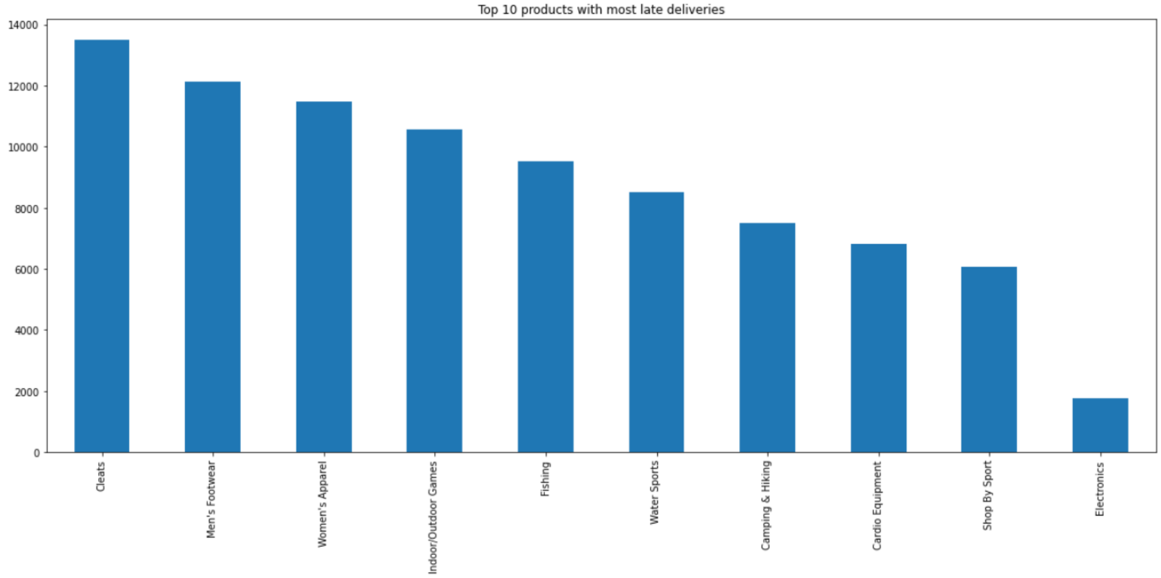


Fig 5.1.3 Top 10 Products with the most late deliveries

Cleats and Men’s footwear are mostly delivered late. Electronics have the least delivery risk.

5.1.4 Sales by Region Analysis

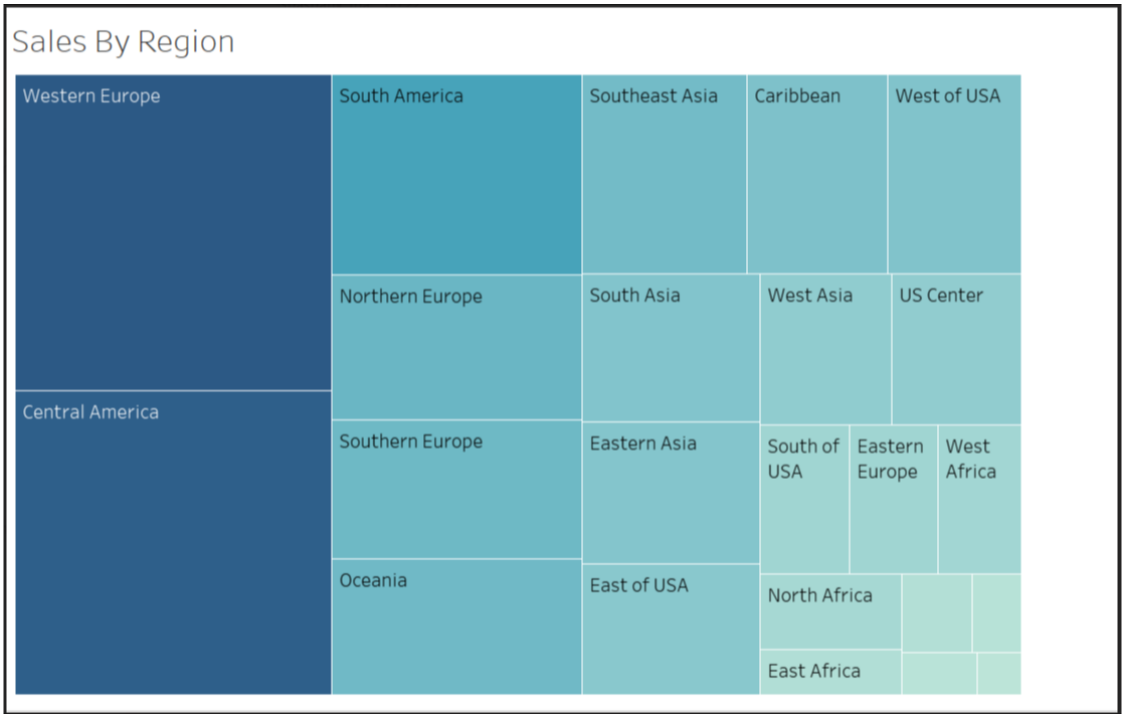


Fig 5.1.4 Sales by Region

Region-wise, Western Europe and Central America have the highest sales. African regions such as North Africa and East Africa have the least sales.

5.1.5 Payment Status with Shipping on Time Analysis

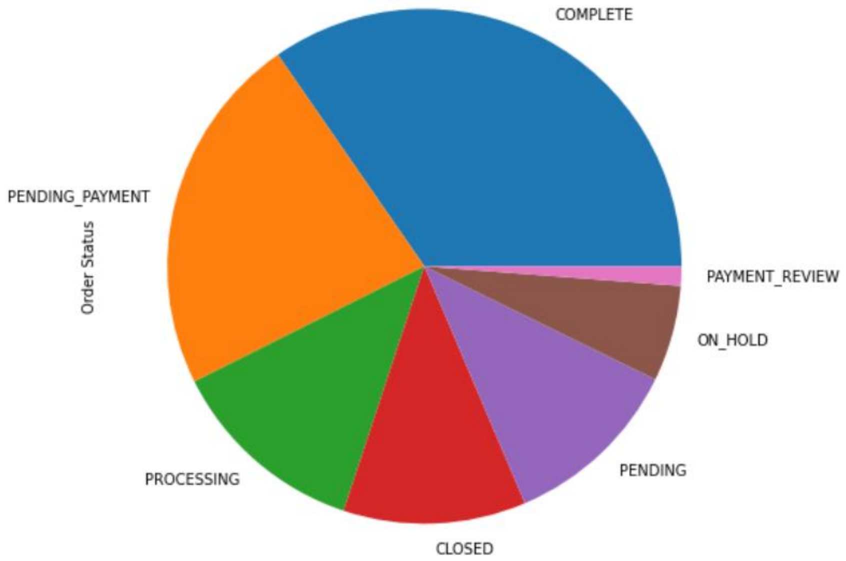


Fig 5.1.5 Payment Status with Shipping on Time Analysis

It is observed that when shipping is on time, most of the payment status is complete. But, still, a portion of the payment is in pending and processing status. This should be taken up with the finance team to see the reason behind it.

5.1.6 Customer Segment Analysis

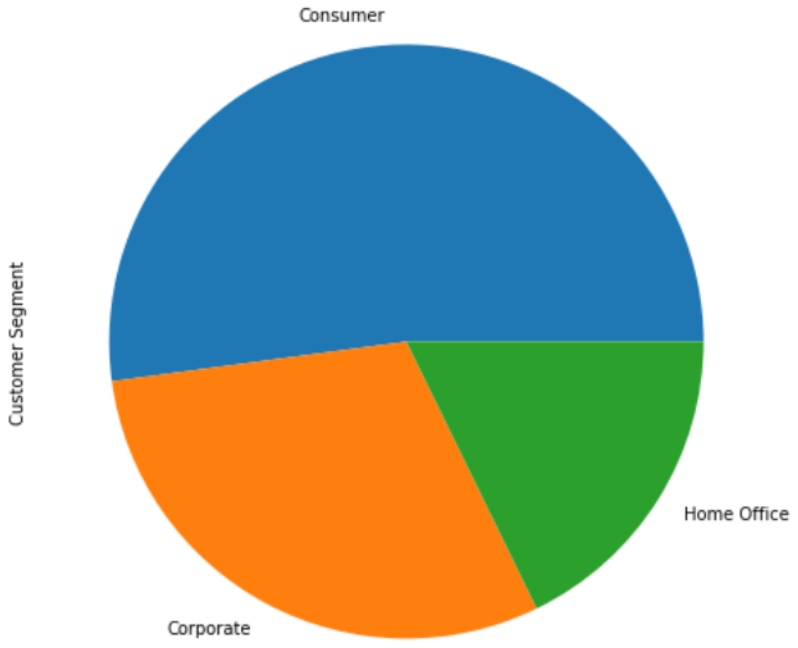


Fig 5.1.6 Customer Segment Analysis

It is observed that most of the customer segment is Consumer followed by Corporate. Further customer segmentation using ABC analysis is done in Section 5.1.7.

5.1.7 Inventory Management using ABC Analysis

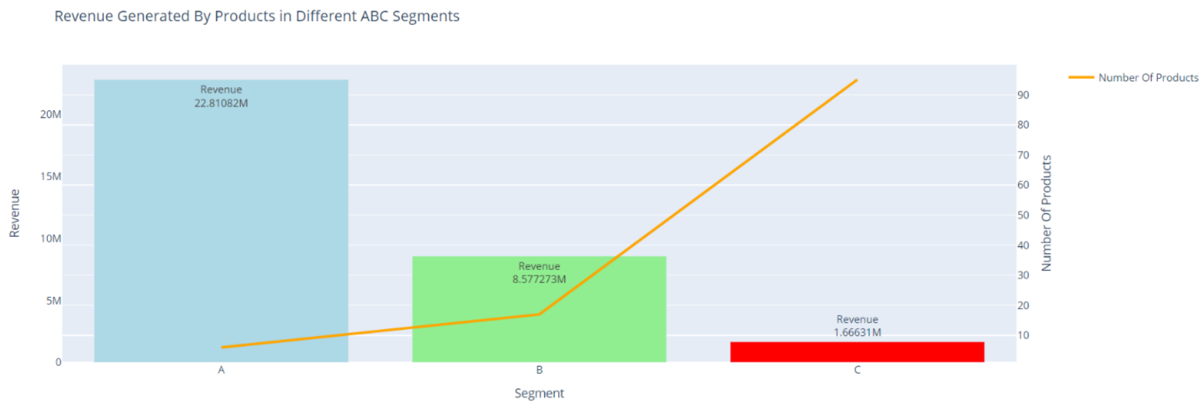


Fig 5.1.8 Revenue Generated by Products in Different ABC Segments

The total revenue is ~$33.07M.

1. Category A is where we generate maximum revenue. From the graph, it is evident that out of 118, only 6 products contribute to ~69% of revenue. DataCo Global must keep these on priority
2. Category B contributes 25% of the revenue. There are 17 products in this category, and DataCo must manage the optimum levels of these products as well.
3. Category C contains the rest of the 95 products. To reduce holding costs, cuts could be made in this category.

5.1.8 Total Order Quantity by Different Shipping Modes

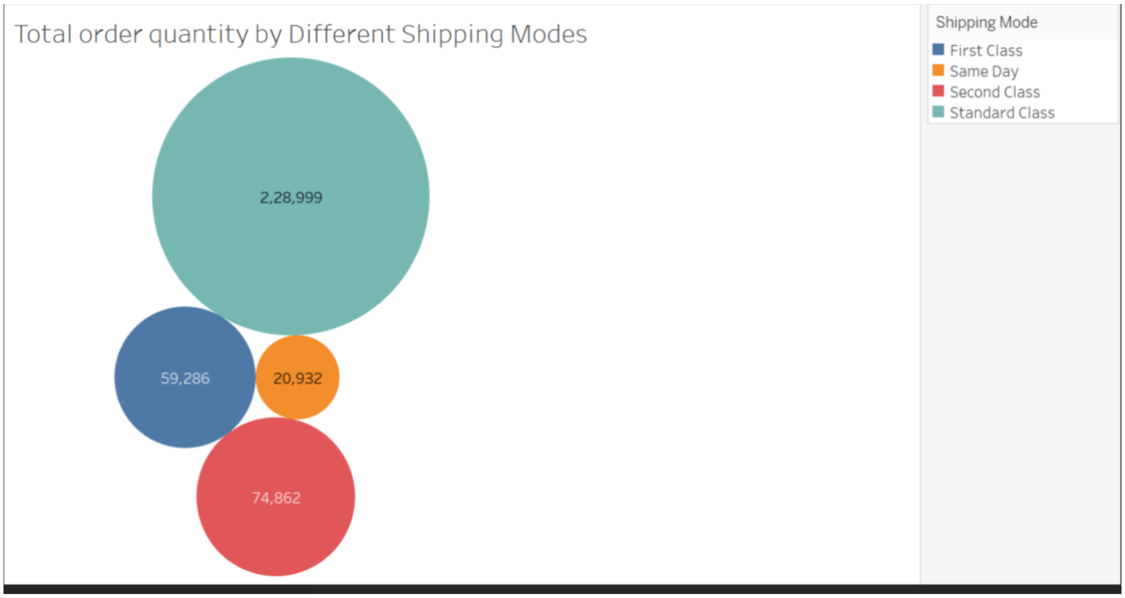


Fig 5.1.8 Total Order Quantity by Different Shipping Modes

With the above Tableau visualization, it is observed that Standard Class is preferred by customers and most of the orders go through this shipping mode. The least preferred shipping mode is the same day and has the least order quantity.

5.1.9 Total Benefits per order by Country

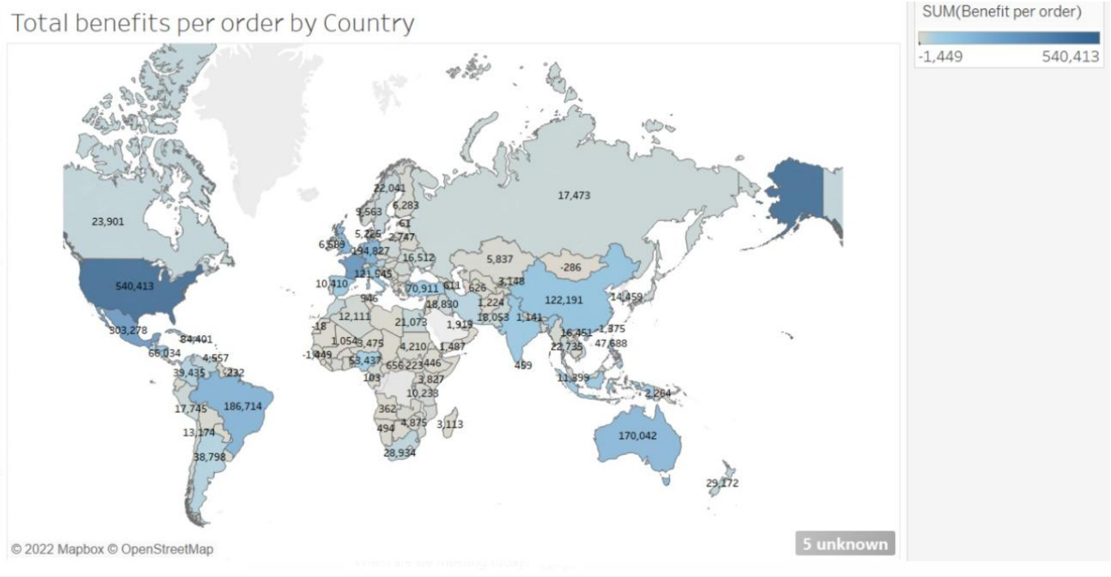


Fig 5.1.9 Total Benefits per order by Country

From the above geographical visualization, we observe that total earnings per order placed are highest in Estados Unidos and Francia. African and Central American countries such as Nigeria, and Guatemala have the least earnings per order placed.

5.1.10 Total Order Items by Country

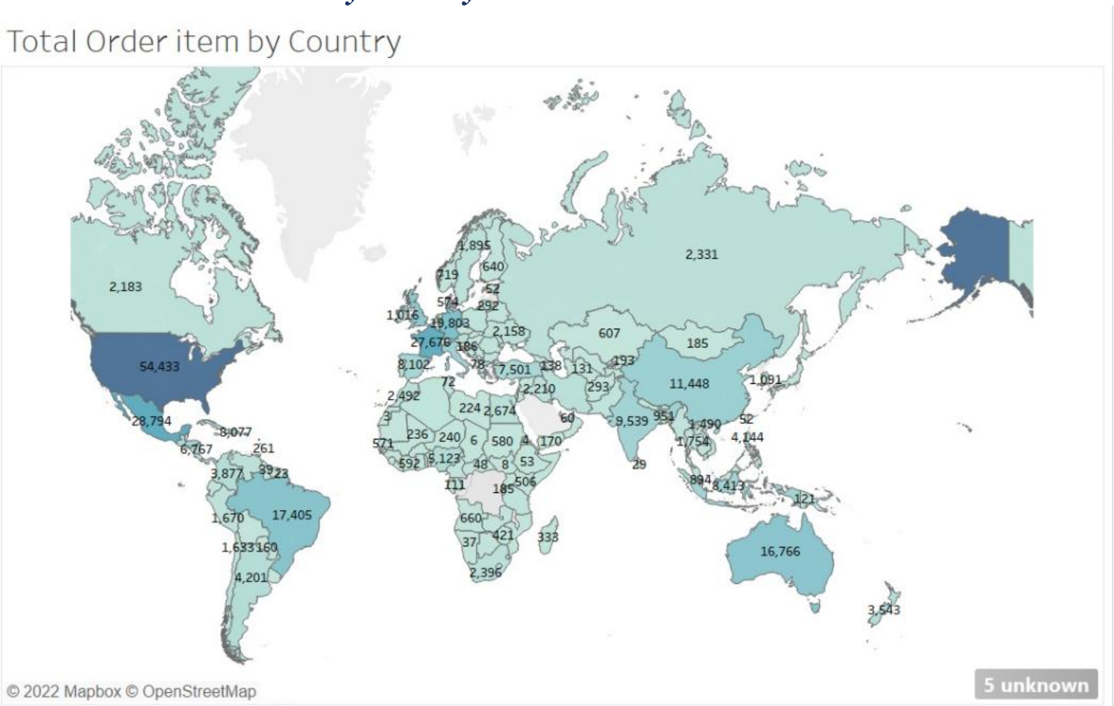


Fig 5.1.10 Total Order Items by Country

Estados Unidos and Mexico followed by Mexico has the highest total order item. The above geographical visualization helps us to get a birds-eye view of how the orders are distributed across the globe.

5.1.11 Countries with no stock available

With EDA, it is observed that countries such as Estados Unidos and Mexico followed by Francia have no stock available and are at late delivery risk. These countries also have a lot of delivery with advanced shipping. This indicates that there is a lot of demand in this market. Hence, supply chain management needs to be more efficient in the above countries.

Chapter 6: Feature Selection and Engineering

For model building, we have selected the following features

- Type
- Customer Country
- Market
- Order Country
- Order weekday
- Order month
- Order year
- Shipping weekday
- Shipping month
- Shipping year
- Shipping Mode
- Fraud
- Late delivery (Target variable)

The features that have been selected are used to understand customer demographics, based on the customer's country and market, order date-based features, type of payment to assess if payment has any impact on delivery time, and estimated delivery parameters that are related to shipping date details.

Chapter 7: Models Used

7.1 Models and Metrics Considered

Table 7.1: Models and Metrics Considered

Model	**Brief Approach**	**Challenges**
Logistic Regression	This classification model is used for time series analysis. In DataCo Global, the late delivery status is considered a predictor of supplier disruption. Using Logistic Regression, the binary outcome of late delivery is predicted and solver “lbfgs” (**Limited-memory Broyden–Fletcher–Goldfarb–Shanno)** is used.	The accuracy score is very low at 70.54%. The Recall Score is 76.45% and the F1 Score is 70.92%.
Gaussian Naïve Bayes	This classifier technique is considered for the DataCo Global dataset as late delivery status is independent of other features such as product damage.	This modeling approach gave the least accuracy score among all the classification models with an accuracy score= of 56.52%, Recall Score = of 55.54%, F1 Score = of 71.41%
Support Vector Machines	With this classification modeling approach, Linear Support Vector Machine is considered as late delivery status can be classified into two classes, and hence, data is linearly separable.	This approach gave a poor accuracy score of 70.54%, a recall score of 76.45%, F1 score of 70.92% which was the same as logistic regression.

Nearest Neighbour Classification	In this classification approach, the KNN algorithm is used with the value of K=1. It classifies a new data point based on similarity using all the stored available data.	The accuracy score is much better than previous classification models with an accuracy score of 98.23%, a Recall score of 98.43%, and an F1 Score of 98.36% but the computation cost is very high and the value of K needs to be determined.
LDA	This approach helps to maximize the distance between the means of two classes and minimizes the variance within the individual class. In the case of DataCo Global, LDA is used to predict the late delivery risk.	This approach gave a poor accuracy score of 70.54%, a recall score of 76.45%, F1 score of 70.92% which was the same as logistic regression and SVM.
Random Forest Classifier	In this classification technique, a random subset of features is taken for splitting a node and the best feature is searched among a random subset of features. In the case of DataCo Global, n_estimators which is the number of trees the algorithm builds before maximum voting is 100, and the maximum depth of trees is 10.	Challenge is that increased accuracy requires more trees which will slow down the model and the relationships cannot be described within the data. Here, the accuracy score is 87.30%with recall score is 86.48% and an F1 score is 88.55%.
eXtreme Gradient Boosting Classification	XGBoost classification approach is used to sequentially built shallow decision trees to avoid overfitting.	In the case of DataCo Global, we have a low accuracy score of 87.30%, recall score of 86.48%, and F1 score of 88.59% which is similar to the Random Forest Classifier as well.

Decision Tree Classification	This tree-structured classifier is used for classification problem where internal nodes represent the features of the dataset and branches represent the decision rules and each leaf node represent the outcome. In the case of DataCo, the late delivery risk is considered a predictor for this approach.	In the case of DataCo Global, if data extrapolation to include the Covid scenario is done, then the accuracy score goes low for this approach.

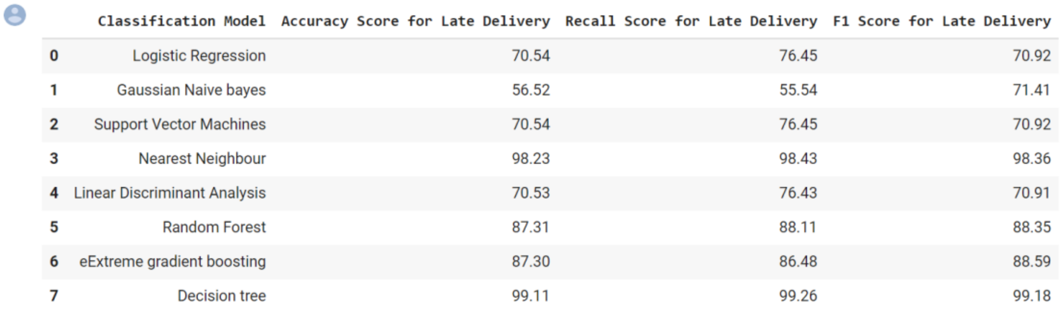
7.2 Models Chosen

Table 7.2

Assumption considered	**Model or Metric Chosen**
Feature values are mostly categorical; use statistical methods to choose the root node internally.	**Decision Tree**; Accuracy, Recall, F1 Score Accuracy: Total number of correct late delivery predictions out of all predictions. Recall: Correctly identifying late deliveries True Positive/ (True Positive + False Negative) F1 Score: This allows us to get a good balance between Precision and Recall. 2* (Precision*Recall)/ (Precision+Recall)

7.3 Results of the models

Comparing all the classification models in the below figure, it is observed that for late delivery risk, the decision tree has the highest accuracy score of 99.11 with a recall score of 99.26 and an F1 Score of 99.18.



	Classification Model	Accuracy Score for Late Delivery	Recall Score for Late Delivery	F1 Score for Late Delivery
0	Logistic Regression	70.54	76.45	70.92
1	Gaussian Naive bayes	56.52	55.54	71.41
2	Support Vector Machines	70.54	76.45	70.92
3	Nearest Neighbour	98.23	98.43	98.36
4	Linear Discriminant Analysis	70.53	76.43	70.91
5	Random Forest	87.31	88.11	88.35
6	eExtreme gradient boosting	87.30	86.48	88.59
7	Decision tree	99.11	99.26	99.18

Fig 7.3.1 Classification Models Comparison

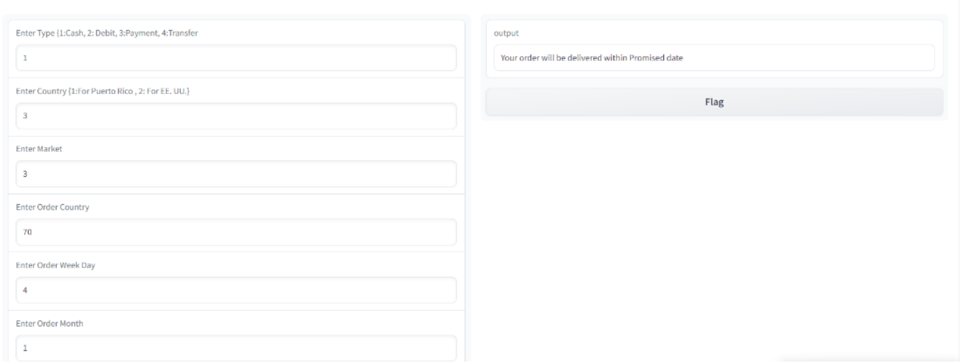
7.4 Model Deployment and Evaluation

Colab : https://colab.research.google.com/drive/14SgT-o3kaviSSt2lq4uT2AaU3BD0LNf7#scrollTo=L3EAZ8NOqzU8

Gradio is the fastest way to demo our machine learning model with a friendly web interface so that anyone can use it without any difficulties.

We have used Gradio to deploy our final decision tree classifier and predict the delays of the products. The Gradio interface is user-friendly and helps the user to predict delays and pass the information through the overall supply chain to reduce unexpected leakages. Gradio also helps in creating a host page which can be seen below.

https://localhost:7860/



Enter Type {1:Cash, 2: Debit, 3:Payment, 4:Transfer

1

Enter Country {1:For Puerto Rico , 2: For EE. UU.}

3

Enter Market

3

Enter Order Country

70

Enter Order Week Day

4

Enter Order Month

1

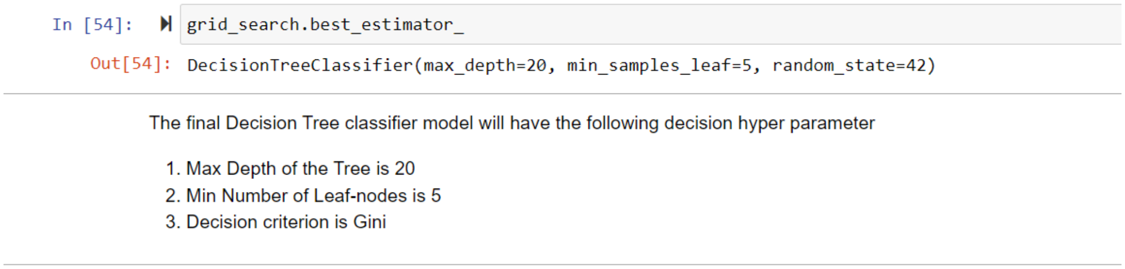
output

Your order will be delivered within Promised date

Fig 7.4.1 Demo for Deployment interface

Post this, hyperparameter tuning is done for the Decision tree and grid search is done with the following observations:

1. Max Depth of the Tree is 20
2. Min Number of Leaf-nodes is 5
3. The decision criterion is Gini



```
In [54]:  grid_search.best_estimator_

Out[54]:  DecisionTreeClassifier(max_depth=20, min_samples_leaf=5, random_state=42)
```

The final Decision Tree classifier model will have the following decision hyper parameter

1. Max Depth of the Tree is 20
2. Min Number of Leaf-nodes is 5
3. Decision criterion is Gini

Fig 7.4.2 Hyper Parameter Tuning

7.5 Model Statistical Evaluation

Model selection involves evaluating a list of machine learning Algorithms and their selection based on performance metrics defined in our business problem.

The problem comes when the difference in the mean performance of the models is real or just a statistical fluke.

Statistical Hypothesis Testing needs to be performed to understand if the difference in the mean performance of different models to predict Delays is based on Actual performance and not just a Statistical fluke.

We Start with the Hypothesis that all Models have the same performance only to prove it wrong if the Hypothesis testing says otherwise.

Null hypothesis: All Models have the same Mean Accuracies

Alternate Hypothesis: Models have Significantly different Mean accuracies and Significantly different performance

We would be using the McNemars Test to understand the Statistical difference in the mean performance of our models.

The MLXtend library will be used via the paired_ttest_5x2cv () function to understand if the Models have significantly different performances or not. We would be using the p-value from the output of this function to determine whether the difference in the performance of the two algorithms is significant or not.

We ran the Hypothesis Test on a different combination of models and Determined that the Decision tree Classifier which has the highest accuracy in terms of the metric we defined in our business Scope matches the result for our Hypothesis Tests. Therefore, the final model chosen based on this test is the **Decision tree Classifier**.

Model 1	**Model 2**	**P-value**	**T stats**	**Difference**
Logistic Regression	Gaussian Naive Bayes	P-value: 0.000	t-Statistic: -42.900	YES
Logistic Regression	Support Vector Machines	P-value: 0.188	t-Statistic: 1.524	NO
Logistic Regression	Nearest Neighbour Classification	P-value: 0.000	t-Statistic: 171.273	YES
Logistic Regression	Random Forest Classifier	P-value: 0.000	t-Statistic: 9.191	YES
Logistic Regression	eXtreme Gradient Boosting Classification	P-value: 0.055	t-Statistic: 2.487	NO
Logistic Regression	Decision tree classification	P-value: 0.000	t-Statistic: 23.855	YES
Decision tree classification	Support Vector Machines	P-value: 0.000	t-Statistic: 22.691	YES

Chapter 8: Conclusion/Recommendations

8.1. Recommendations & Business Impact

Predictive analytics enables supply chain professionals to collect and analyze data and helps management make data-driven decisions. It also recommends solutions for issues like shipping delays, carrier constraints, warehouse inefficiencies, and inventory shortages.

Predictive capabilities enable organizations to monitor trends (e.g., customer market, traffic, labor, weather) and get a peek into the future. Supply chain managers utilize technologies like artificial intelligence and machine learning algorithms to recognize and mitigate risks by identifying patterns in yearly, monthly, weekly, and even daily data.

Major recommendation:

1. Post Exploratory Data Analysis (EDA), it is observed that countries such as the United States and Mexico followed by France have no stock available and are at late delivery risk. These countries also have the highest order of items with advanced shipping. This indicates that there is a lot of demand in this market. Hence, supply chain management needs to be more efficient in these countries.

2. Using ABC analysis, it is observed that Category A is where we generate maximum revenue. Out of 118, only 6 products contribute to ~69% of revenue. DataCo Global must keep these on priority.

   Category B contributes 25% of the revenue. There are 17 products in this category, and DataCo must manage the optimum levels of these products as well.

   Category C contains the rest of the 95 products. To reduce holding costs, cuts could be made in this category.

3. Setting up a Supply chain Control tower which will help the planners to predict delays for Products based on Historical Sales data and Delay frequencies. Also, Opportunity analysis is to be done based on Lane level data (Customer

Country- order Country) to understand if the product can be Rerouted using other closer Shipping Locations.

8.2. User Value from this project

There are two types of users that we have devised the analysis and the model for.

1. DataCo Global: The organization can assess inventory, best-selling products, high-profit products, products with the most loss, etc.

2. End Customer: The end customer would know if their order would be delivered on time based on the model output.

8.3. Future Scope of this project

1. The project can be extended to accommodate the recent COVID pandemic scenario by simulating the data for 2019-2020. As suggested by the sponsor, we did data extrapolation of 2019-20 data by using regression and manual modeling (by taking the average late delivery risk and increasing it considering the pandemic) but this did not give us a good accuracy score for classification models as data was biased. Currently, this extrapolated dataset is not available as open source. So, such data can be simulated to understand the impact of late delivery risk in supply chain management during such unforeseen circumstances.

2. As part of the project for deployment, we have used Gradio which uses a local host. With help of UI experts, the interface can be enhanced with additional features such as a drop-down box for user input and hosting the same on a production environment with adequate software requirements. Further, as per the historical data, we can give an accurate decision to the user on product delay/breach by x days.

3. Customer Segmentation based on RFM analysis could be done. We could do recency, frequency, and monetary-based segmentation of the customers that order from DataCo Global to strategize according to these customers' buying patterns.

4. Forecasting for the coming years based on the current demand and buying patterns.

References

1. Shang, Yang, Dunson, David B, Song., Jing-Sheng. June 2017 “Exploiting Big Data in Logistics Risk Assessment via Bayesian Nonparametrics”, Operations Research 65(6), Research Gate
2. Katsaliaki,K, Galetsi,P., Kumar, S. Jan 2021. “Supply chain disruptions and resilience: a major review and future research agenda”, Annals of Operations Research, Springer.
3. Hyster-Yale Group, Inc. 2022, “Improving DC metrics” https://www.yale.com/en-us/north-america/support-resources/white-papers/improving-dc-metrics/
4. Baryannis, George., Dani, Samir., Antoniou, Grigoris., 2019, July 26, University of Huddersfield, “Predicting Supply Chain Risks Using Machine Learning: The Trade-off Between Performance and Interpretability”
5. Brintrup, Alexandra Melike., Pak, Johnson., Ratiney, David., Pearce, Tim., November 2019, “Supply chain data analytics for predicting supplier disruptions: a case study in complex asset manufacturing”, International Journal of Production Research.
6. Alicke, Knut., Dilda, Valerio., Görner, Pierrick., Reiter, Sebastian., Samek, Robert., 2021, April 30, “Succeeding in the AI supply-chain revolution”, McKinsey
7. Melendez, Carlos., 2020, Oct 28, “Five Ways AI Can Supply -Chain Disruption”, Think Tank, https://www.supplychainbrain.com/blogs/1-think-tank/post/32125-five-ways-ai-can-solve-supply-chain-disruption
8. Carey, Nick., 2022, May 3 “Startups apply artificial intelligence to supply chain disruptions”, Reuters, https://www.reuters.com/technology/startups-apply-artificial-intelligence-supply-chain-disruptions-2022-05-03/
9. https://www.marineinsight.com/maritime-law/causes-and-consequences-of-vessel-delays-in-container-shipping/
10. https://builtin.com/data-science/random-forest-algorithm
11. https://machinelearningmastery.com/hypothesis-test-for-comparing-machine-learning-algorithms/

12. https://www.vertica.com/docs/10.1.x/HTML/Content/Authoring/AnalyzingData/MachineLearning/XGBoost/XGBoostForClassification.htm
13. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
14. Wang, G., A. Gunasekaran, E. W. Ngai, and T. Papadopoulos. 2016. "Big Data Analytics in Logistics and Supply Chain Management: Certain Investigations for Research and Applications." International Journal of Production Economics 176: 98–110.