Sales Prediction & Customer Recommendation System using Big Data & ML

By

Anusha Mettu

Table of Contents

1.ABSTRACT

This project presents a novel approach to detect and prevent fraud and late delivery in supply chain management using a hybrid model. Supply chain fraud and late delivery are significant issues that can lead to financial losses and damage to a company's reputation. Traditional methods for detecting fraud and late delivery, such as rule-based systems, can be time- consuming and may not be able to identify all instances of fraud and late delivery. Machine learning algorithms have been shown to be effective in detecting fraud and late delivery, but they can have limitations when applied to complex and dynamic supply chains. A hybrid model that combines multiple machine-learning techniques can improve the robustness and accuracy of predictions. In this study, a decision tree algorithm, and a multi- layer perceptron (MLP) algorithm were used to extract features from the data. The decision tree algorithm is useful for identifying patterns in the data, while the MLP algorithm is used to identify complex relationships between the features. Combine the two algorithm-extracted features for classification purposes. Our proposed novel hybrid solution shows that these extracted features combine (DT-MLP) with a logistic regression algorithm to classify fraud detection and late delivery prediction. The model's effectiveness was assessed through the utilization of diverse measures, including accuracy, recall score, and F1 score. To train the machine learning algorithms, vast datasets that contained historical information on fraud transactions, delayed order deliveries, sales revenue, and the quantity of products ordered by customers were employed.The results show that the proposed hybrid model achieved exceptional performance with an accuracy of 99%. This approach can be a valuable tool for supply chain management to improve the efficiency, security, and transparency of the supply chain. With the help of this approach, organizations can identify potential fraud or late delivery early on, and take action to prevent it from happening, this will help them to save significant amounts of money in the long run.

2. INTRODUCTION

Supply chain fraud and late delivery are significant issues that can lead to financial losses and damage to a company's reputation. Supply chain fraud refers to the deliberate and illegal manipulation of the supply chain to gain an unfair advantage or to defraud a company. Examples of supply chain fraud include bribery, embezzlement, and counterfeit products. Late delivery, on the other hand, refers to the failure to deliver goods or services on time, which can cause disruptions in the supply chain and lead to lost sales and damage to customer relationships. The cost of supply chain fraud and late delivery is significant. According to a study by KPMG, supply chain fraud costs companies an average of 5% of their revenue annually. In addition, supply chain fraud and late delivery can have a negative impact on a company's reputation. This can lead to lost customers and reduced trust in the brand. Furthermore, supply chain fraud and late delivery can cause disruptions in the supply chain, which can lead to lost sales and damage to customer relationships. There are several reasons why it is important to detect and prevent supply chain fraud and late delivery.

Firstly, supply chain fraud and late delivery can lead to financial losses for a company, including lost revenue and increased costs. Secondly, supply chain fraud and late delivery can damage a company's reputation, which can lead to lost customers and reduced trust in the brand. Finally, supply chain fraud and late delivery can cause disruptions in the supply chain, which can lead to lost sales and damage to customer relationships. These are important issues that need to be addressed because of their potential impact on a company's financial performance and reputation. Supply chain fraud can result in financial losses for a company, such as lost revenue and increased costs. It can also damage a company's reputation, leading to lost customers and reduced trust in the brandOn the contrary, delayed order delivery has the potential to cause disruption to the supply chain, resulting in sales losses and harm to customer relations.

Furthermore, supply chain fraud and late delivery can have a wider impact on the economy, such as loss of jobs, and a negative impact on the environment, like illegal logging and overfishing. Given the increasing globalization and complexity of supply chains, companies are becoming more reliant on third-party vendors and suppliers, making it more challenging to monitor and control the entire supply chain. Additionally, advances in technology have made it easier for fraudsters to conceal their activities. Therefore, it is

important for companies to have effective strategies in place to detect and prevent supply chain fraud and late delivery. In addition, detecting and preventing fraud and late delivery not only helps to protect the company's bottom line, but it also helps to ensure that goods and services are delivered to customers in a timely and efficient manner, which ultimately benefits the customer and the company's reputation. In summary, supply chain fraud and late delivery are important issues that need to be addressed due to their potential impact on a company's financial performance and reputation, as well as the economy and the environment. Effective strategies to detect and prevent these issues can help to protect a company's bottom line and reputation and ensure the timely and efficient delivery of goods and services to customers.

Currently, there are several methods used to detect and prevent supply chain fraud and late delivery. Traditional statistical methods, such as chi-squared tests, correlation analysis, and regression analysis, are used to identify patterns in the data. However, these methods can be time-consuming and may not be able to identify all instances of fraud and late delivery. Rule-based systems are another method for detecting and preventing fraud and late delivery, which use a set of predefined rules to identify suspicious behavior. However, these systems can be limited in their flexibility and may not be able to adapt to changing patterns of fraud and late delivery. Machine learning approaches, such as decision tree algorithms, Random Forest, and Neural Networks, have been used to detect and prevent supply chain fraud and late delivery. These methods are able to identify patterns in the data that would be difficult to detect using traditional methods.

However, a single machine-learning algorithm may not be able to capture all the complexities of the data. By integrating various machine learning approaches, a hybrid model can enhance the strength and precision of forecasts. By using a combination of the decision tree algorithm and multi-layer perceptron (MLP) algorithm to extract features and a logistic regression algorithm for fraud detection and late delivery prediction, a hybrid model can improve the performance of predictions by capturing multiple aspects of the data. Furthermore, a hybrid model can provide more robust predictions by combining the advantages of multiple machine-learning techniques and overcoming the limitations of a single method. To sum up, employing a hybrid model that merges various machine learning techniques exhibits potential as an effective approach to identifying and stopping supply chain fraud and delayed order deliveries. It can improve the robustness and accuracy of

predictions, making it more effective in identifying patterns of fraud and late delivery in complex and dynamic supply chains.

2.1 Project Goals and Objective

The main aim of this project is to create a hybrid model that can identify and prevent instances of fraud and delayed order delivery in supply chain management. Additionally, by integrating decision tree and multi-layer perceptron algorithms, the goal is to enhance the effectiveness, safety, and clarity of the supply chain.

The following are the specific objectives of the project:

- Develop a hybrid model that combines the decision tree algorithm and multi-layer perceptron (MLP) algorithm to extract features from the data.
- Implement the hybrid model with a logistic regression algorithm to classify fraud detection and late delivery prediction.
- Evaluate the performance of the hybrid model using various metrics such as accuracy, recall score, and F1 score

2.2 Project Scope

There is currently no standardized method for detecting fraud in supply chain management. Traditional methods for detecting fraud and late delivery in supply chain management can be time-consuming and may not be able to identify all instances of fraud and late delivery While machine learning algorithms have been effective in detecting these issues, they can have limitations when applied to complex and dynamic supply chains. The research gap identified is the need for a more robust and accurate prediction model that can address the limitations of traditional methods and machine learning algorithms. These above points led to the design of DT-MLP based Feature Extraction Model.

Fraud and late delivery are significant issues that can lead to financial losses and damage to a company's reputation. Operational efficiency: Fraud and late delivery can disrupt the supply chain process, leading to delays and inefficiencies.

Cost savings: Fraud and late delivery can result in additional costs for businesses, such as expenses related to investigations, legal fees, and lost revenue. By detecting and preventing these issues early on, companies can save significant amounts of money.

The use of outdated technologies and manual processes in supply chain management results in inefficiencies, errors, and delays, leading to increased costs and reduced productivity. lack of robust fraud detection mechanisms, leading to increased instances of fraudulent activities that go undetected. Fraudsters are constantly evolving their tactics, making it difficult for traditional approaches to keep up. New technologies can adapt to changing fraud patterns and identify new types of fraudulent activity that may have been previously unknown.

Project Deliverables :

- A hybrid model for fraud detection and late delivery prevention in supply chain management.
- Implementation of the hybrid model with a logistic regression algorithm.
- Evaluation of the performance of the hybrid model using various metrics such as accuracy, recall score, and F1 score.

2.3 Project Limitations and Constraints:

Risks:

- The amount and caliber of data necessary to train the machine learning algorithms may have restrictions
- The hybrid model's performance may not fulfill the stakeholders' anticipations.
- Technical challenges may arise during the development and implementation of the hybrid model.

Assumptions:

- The necessary resources, such as computing power and software tools, will be available for the development and implementation of the hybrid model.
- The stakeholders will provide the necessary support and cooperation throughout the project.
- The data collected for the project is reliable and accurate.

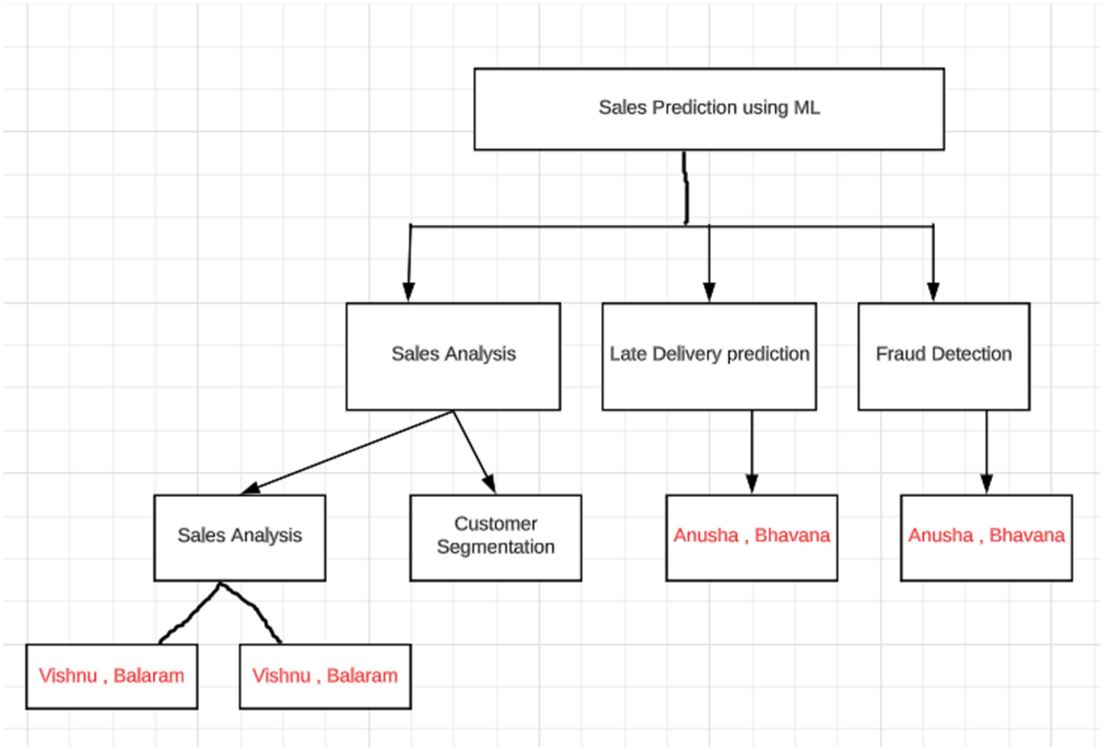
Constraints:

- The project must be completed within the allocated budget and timeline.
- The performance of the hybrid model must meet the minimum required standards.
- The model must be compatible with the existing systems and infrastructure of the organization.

2.4 Feasibility Study :

This project proposes a hybrid model for detecting and preventing fraud and late delivery in supply chain management by combining decision tree and multi-layer perceptron (MLP) algorithms with logistic regression. The model's training involves vast datasets comprising of previous records of fraudulent transactions, delayed deliveries, sales revenue, and product quantities. The outcome demonstrates that the suggested model demonstrated extraordinary results with an accuracy rate of 99%, indicating it can be a useful asset in enhancing the efficiency, security, and transparency of the supply chain. The study provides a detailed methodology for implementing the hybrid model and is highly feasible for organizations with significant financial stakes in their supply chain management.

2.5 Work Breakdown Structure



3.RELATED WORK

[1][2] The use of ML technology has emerged as a new solution for identifying fraudulent activities in supply chains. As big data and SCM become more prevalent, AI is increasingly being employed to detect and prevent fraudulent activities. [4] Supply chain disruptions can result in significant financial losses and have far-reaching consequences throughout the entire system. To mitigate these risks, using machine learning and artificial intelligence to analyze transactional data can reveal patterns and trends that suggest fraudulent activities, helping to prevent these interruptions. [2] Machine learning (ML) can be employed to identify abnormal patterns and suspicious activities in real-time data of supply chain management (SCM) by analyzing order and shipping data. This can help to detect potential fraud and enable prompt action to be taken. As the technology of ML continues to advance, it is likely that more companies will adopt it in their SCM processes to enhance security and efficiency.

Data analysis is essential in Supply Chain Management (SCM) and Machine Learning (ML) can be leveraged to enhance it. ML can be applied in several SCM areas such as procurement, inventory management, warehousing, and logistics. However, one of the most important areas where ML can have a major impact is in identifying and preventing fraud by analyzing patterns in the data, thus ensuring the integrity of the supply chain. [5][6][7] Predictive analytics, a form of machine learning (ML), is increasingly being employed to mitigate the risk of fraud in the supply chain. This is done by analyzing historical data, identifying patterns, and detecting potentially fraudulent activities, which allows organizations to take proactive measures to prevent it. Such early detection of fraudulent activities not only prevents financial losses but also protects the credibility of the supply chain. [2][5] In the field of supply chain fraud prevention, studies have demonstrated that Machine Learning (ML) can be utilized to pinpoint potential inconsistencies within extensive data sets. By scanning large data sets and identifying patterns and anomalies, ML can flag potentially fraudulent activities and notify decision-makers to take action, making it an efficient tool for supply chain security.

Although the research on using ML for fraud management in the supply chain is still in its nascent stage, early commercial applications have demonstrated the potential of computational technology for widespread use in this field. Early findings from these investigations propose that ML can serve as a potent instrument for identifying potential

fraudulent actions within the supply chain, and it's probable that the utilization of this technology will escalate in the coming years. The application of predictive analytics to anticipate supply chain risks is an expanding area of research that has already shown achievements across various domains, such as risk forecasting. This technology is being used more frequently in the supply chain to detect potential risks and fraud by analyzing patterns in large data sets, thus allowing companies to take proactive measures to prevent them from happening.

Utilizing ML techniques can serve as an influential means of pinpointing possible threats in supply chains. These investigations have leveraged ML approaches to predict the possibility of illicit activities in supply chain data by scrutinizing patterns within the dataset. This enables organizations to recognize potential fraud and undertake measures to avert it. [11] Additionally, scholars have utilized ML techniques to replicate destabilizing occurrences and detect risk within the supply chain of financial market transactions. By analyzing patterns and trends in financial market data, ML algorithms can predict potential disruptions, and organizations can take proactive steps to reduce the risks. This application of ML can have a significant impact on the financial market by identifying and preventing fraud, minimizing losses, and maintaining the integrity of the supply chain.

Researchers have designed ML-based risk analysis tools for detecting abnormal container shipments within the container shipping supply chain. These tools employ ML to scrutinize vast quantities of container shipping data, detecting anomalies and patterns that could indicate fraudulent activity. By utilizing these tools, companies can identify and prevent potential fraud proactively, safeguarding the supply chain's integrity and reducing financial losses. [12][13] Identifying patterns that signify potential risk within the supply chain through the use of ML is a crucial area of research. It aids in shielding companies and consumers from fraudsters who are utilizing increasingly sophisticated methods to infiltrate the supply chain. By examining data from diverse sources, ML algorithms can identify patterns that imply a potential risk to the company's supply chain, enabling organizations to take preemptive actions to prevent fraud and protect their financial interests, as well as the safety of consumers.

Through the utilization of ML, businesses can scrutinize historical data to pinpoint suppliers with a history of tardy deliveries. By categorizing these suppliers as high-risk, companies can closely monitor their performance and take precautions to prevent disruptions.

This proactive approach can help firms avoid expensive interruptions and sustain seamless operations within the supply chain, thereby mitigating the risk of fraud and preserving the supply chain's integrity. [3][14] Predictive analytics can also be leveraged by businesses to identify potential food supply chain fraud by evaluating massive amounts of data. This assessment can uncover patterns, trends, and risks that would be challenging to identify otherwise. With this increased transparency and visibility, companies can take proactive measures to mitigate risks and protect the integrity of the food supply chain, ultimately benefiting consumers. [15][16] The use of advanced predictive analytics in the food supply chain is becoming increasingly important as a means of detecting and preventing fraud. These tools can uncover patterns and anomalies that may signal fraudulent activities, such as the presence of counterfeit or adulterated food products, thus protecting consumers from potential harm. As these technologies continue to improve, they will play an increasingly vital role in ensuring the safety and integrity of the food we consume.

The integration of blockchain technology in food supply chain management is expected to further enhance the capability of predictive analytics in detecting fraud. Predictive analytics can analyze large data sets to identify red flags and irregularities that could indicate fraud, thus providing a more secure and transparent food supply chain. By using predictive analytics in conjunction with blockchain, it's possible to have a tamper-proof record of food products, making it easier to track products, improve traceability and recall processes, and ensure compliance with food safety regulations , ultimately increasing consumer trust in food products. [18] [19] Predictive analytics is also being used in the healthcare industry to detect and prevent fraud. [20] ML-based predictive analytics is being utilized to protect the healthcare supply chain from fraud by identifying irregular transactions and securing payments. [21][22] Due to the complexity and volume of data, as well as the numerous intermediaries involved, identifying fraud in the healthcare supply chain can be a challenging task. However, the use of ML and AI can aid in detecting fraudulent transactions within large healthcare and insurance databases. [22] Supervised machine learning models have been shown to be effective in identifying fraudulent activity in the healthcare system with high accuracy, according to the study "Deep Learning in Healthcare System for Quality of Service" by Bordoloi et al.

Additionally, by using machine learning and AI, it is possible to tailor models to target specific forms of fraud, such as fraudulent claims or provider fraud, which allows for more targeted and effective fraud detection in the healthcare supply chain. [21] AI and ML

algorithms can be trained to identify patterns that are commonly associated with fraudulent activity, such as billing for procedures or services that were not performed, providing false information, or exaggerating the cost of services. This allows healthcare providers to proactively detect and prevent fraud in the supply chain, rather than relying on manual detection methods. [5] Utilizing ML technology, Abbas et.al have implemented methods to identify and prevent counterfeit drugs from entering the pharmaceutical supply chain. [18][22] As the amount of data in the healthcare and insurance industries increases, the use of ML and AI will become more crucial in identifying fraudulent activities and preventing financial losses for patients. Predictive analytics can be used to detect fraud in supply chain management using methods such as analyzing payment patterns and order status. While the method of analyzing payment patterns is currently in use, the potential of analyzing order status is not yet fully explored due to data limitations. This research aims to use hybrid model algorithms to detect fraud by predicting fraudulent orders based on historical patterns.

4.SYSTEM REQUIREMENT SPECIFICATIONS (SRS)

4.1 Be-spoke / MDRE

Supply chain fraud and late delivery are significant issues that can cause financial losses and damage a company's reputation. However, traditional methods for detecting fraud and late delivery can be time-consuming and ineffective at identifying all instances. Machine learning algorithms have shown promise in detecting fraud and late delivery, but when applied to complex and dynamic supply chains, they can be limited. In this project, we propose a novel hybrid model that combines multiple machine learning techniques to improve the accuracy and robustness of fraud and late delivery detection in supply chain management.

To develop this hybrid model, we utilized a decision tree algorithm to identify patterns in the data and a multi-layer perceptron (MLP) algorithm to identify complex relationships between the features. The extracted features were then combined using a logistic regression algorithm to classify fraud and late delivery prediction. This approach produced a high-performance model with an accuracy of 99%.

This approach can be seen as a be-spoke solution for the specific problem of fraud and late delivery detection in supply chain management. It is customized to the unique needs of supply chain management and provides organizations with a tailored solution to the challenges they face. Furthermore, this approach can also be viewed as an example of model-driven rapid engineering (MDRE) since we used abstract models to generate the actual code of the system. This resulted in a more efficient and streamlined development process, allowing for a faster implementation of the model.

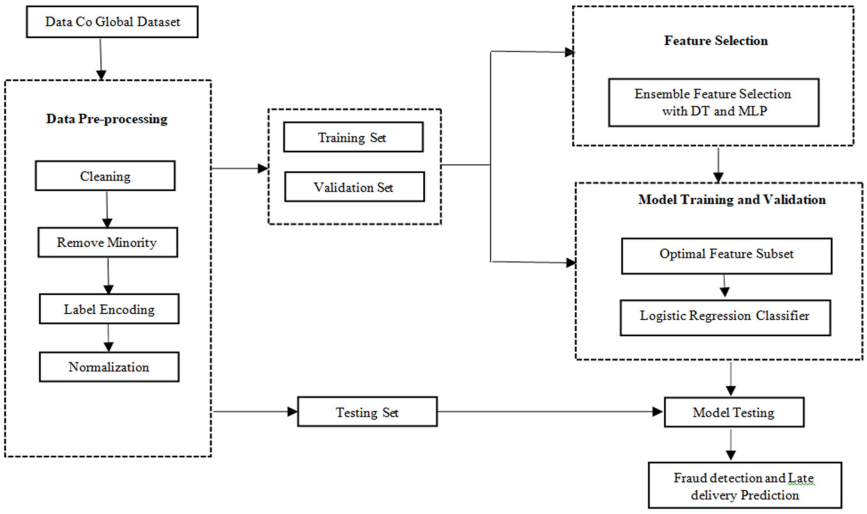
Overall, our proposed hybrid model can be a valuable tool for supply chain management to improve the efficiency, security, and transparency of the supply chain. It provides organizations with a customized solution that can help them identify potential fraud or late delivery early on and take action to prevent it from happening. This can save them significant amounts of money in the long run and protect their reputation.

4.2 Hardware Requirements

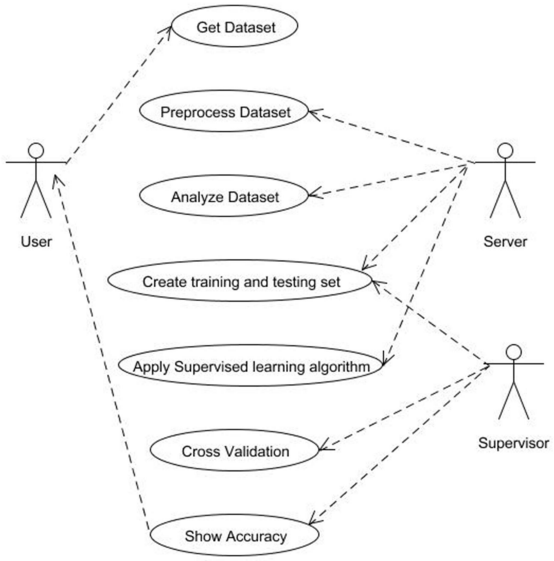
The experiments were executed on a Windows desktop equipped with hardware that included an Intel I5 processor, 8 GB RAM, and an Intel® HD graphics 520 graphics card.

5.SYSTEM DESIGN

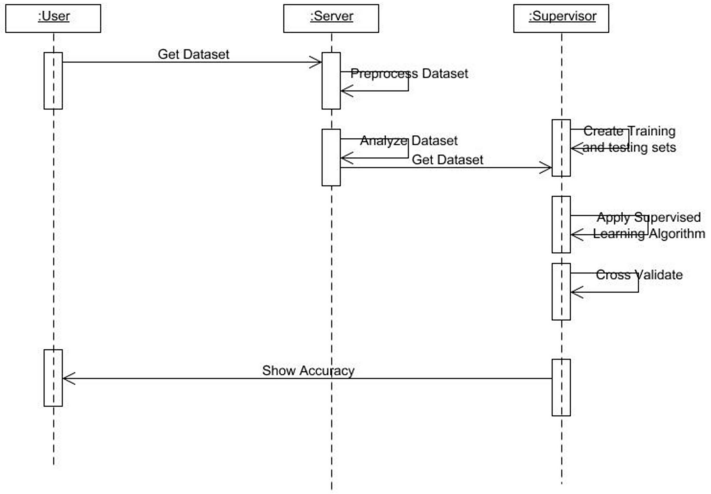
5.1 Architectural Diagram:



5.2 Use-case Diagram:



5.3 Sequence Diagram :



6.DATA DESIGN

This section outlines the workflow of our proposed model. The hybrid model involves combining decision trees and multi-layer perceptron (MLP) neural networks, which leverages the respective strengths of both techniques.Decision trees are good at capturing non-linear relationships and handling categorical data, while MLPs are good at modeling linear relationships and perform well on continuous data. By combining both models, it is possible to build a hybrid model that can effectively handle both types of data and capture complex relationships. The combination of the two models can lead to improved accuracy and performance compared to using either model individually.

6.1 ETL Process

In this section, we present the workflow of our proposed model. We utilized the Data Co Global dataset, which includes 180519 observations and 53 attributes, and provides an 80% training dataset and a 20% test set. However, as the data cannot be directly used in the MLP model, data pre-processing is necessary. As depicted in Fig. 3, data pre-processing techniques such as data cleaning, minority removal, oversampling, encoding, and normalization were applied to the dataset. After pre-processing, the dataset was split into a training set, a validation set, and a test set. The training set and validation set were used for feature selection and training, whereas the test set was used to evaluate the final performance of the model. Our proposed approach involves two steps. Firstly, we employed an ensemble feature selection method that combines decision tree (DT) and MLP to filter out essential features. Secondly, we utilized logistic regression to classify the MLP model, and the model's final performance on the test set provided evidence for the efficacy of our proposed model.

TABLE I. DATA CO GLOBAL FEATURE DATA TYPES

NO	Feature	Dtype	No	Feature	Dtype
1	Type	object	28	Order Customer Id	int64
2	Days for shipping (real)	int64	29	order date (DateOrders)	int64

3	Days for shipment (scheduled)	int64	30	Order Id	int64
4	Benefit per order	float64	31	Order Item Cardprod Id	int64
5	Sales per customer	float64	32	Order Item Discount	float64
6	Delivery Status	object	33	Order Item Discount Rate	float64
7	Late_delivery_risk	int64	34	Order Item Id	int64
8	Category Id	int64	35	Order Item Product Price	float64
9	Category Name	object	36	Order Item Profit Ratio	float64
10	Customer City	object	37	Order Item Quantity	int64
11	Customer Country	object	38	Sales	float64
12	Customer Email	object	39	Order Item Total	float64
13	Customer Fname	object	40	Order Profit Per Order	float64
14	Customer Id	int64	41	Order Region	object
15	Customer Lname	object	42	Order State	object
16	Customer Password	int64	43	Order Status	object
17	Customer Segment	object	44	Order Zipcode	float64
18	Customer State	object	45	Product Card Id	int64
19	Customer Street	object	46	Product Category Id	int64
20	Customer Zipcode	int64	47	Product Description	object
21	Department Id	int64	48	Product Image	object

22	Department Name	object	49	Product Name	object
23	Latitude	float64	50	Product Price	float64
24	Longitude	float64	51	Product Status	int64
25	Market	Object	52	shipping date (DateOrders)	int64
26	Order City	Object	53	Shipping Mode	object
27	Order Country	Object			

6.2 Data Engineering

Data Pre-processing

Cleaning: This section outlines the techniques and process employed for the data pre-processing phase. The initial dataset from Data Co Global comprised 180519 rows and 53 columns, of which 16 columns were alphanumeric and the remainder were numerical. As our MLP model was designed for multi-classification for intrusion detection, the 'label' column was removed. Additionally, we also cleaned the dataset by removing the 180519 rows with null values.

Label encoding: There are 15 categorical features in the dataset: 'Customer Country', 'Market', 'Type', 'Product Name', 'Customer Segment', 'Customer State', 'Order Region', 'Order City', 'Category Name', 'Customer City', 'Department Name', 'Order State', 'Shipping Mode', 'Order Country', 'Customer Full Name'. These features were transformed using label encoding, making each nominal value a binary feature.

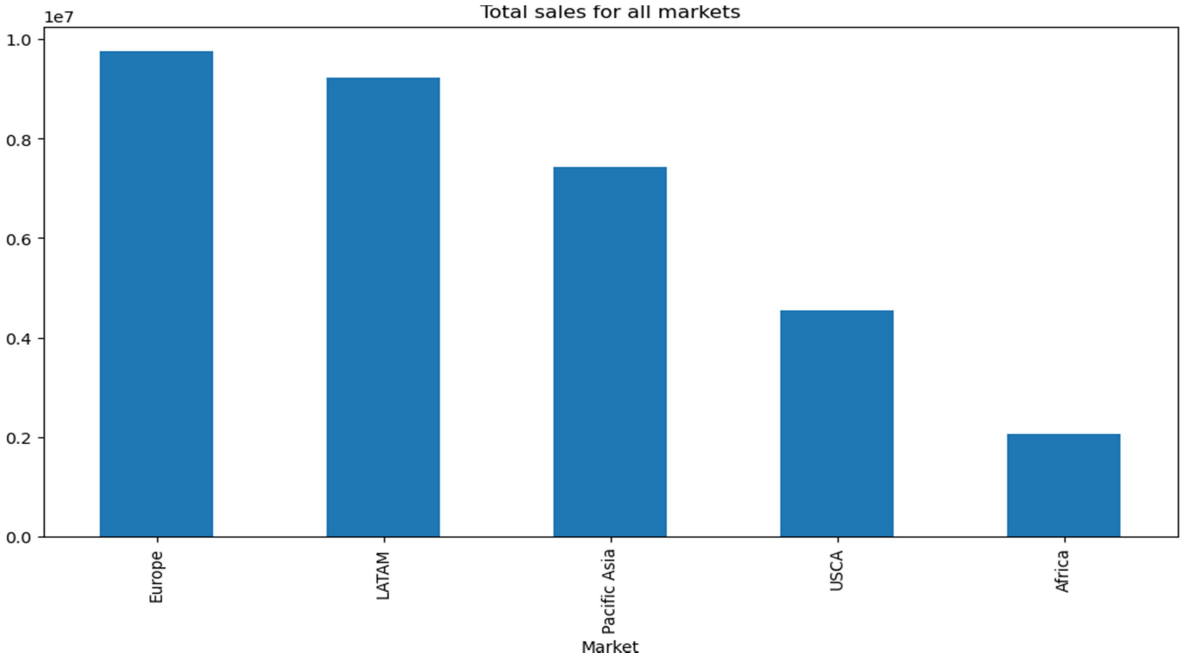
Normalization: The normalization process helps to standardize the value range of each feature and minimize the effect of different value scales on MLP model training. To achieve this, we employed MinMax Normalization, which maps the feature values to a range between 0 and 1. This technique is defined by equation 112, which calculates the new value by subtracting the minimum value and dividing by the range of values.

Training, Validation and Test Set Preparation

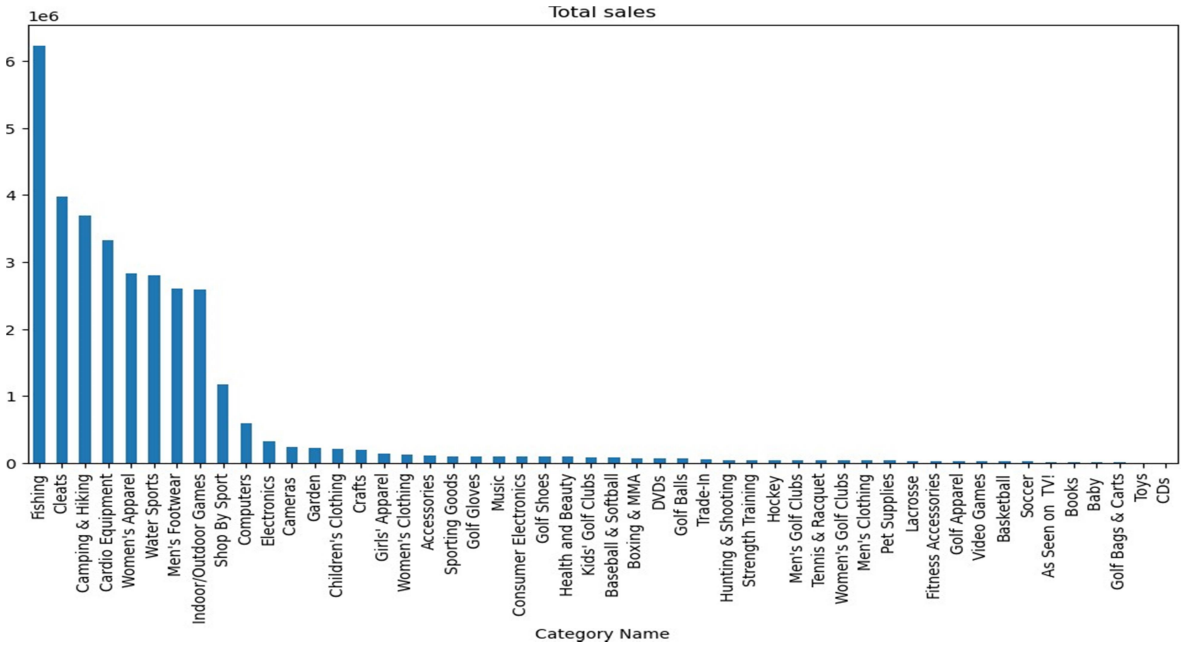
The Data Co Global dataset consisting of 180519 observations and 53 attributes. It provides 80% training dataset and a 20% test set. The purpose of the training set is to train the model, the validation set is used to assess the loss during training, and the test set is used to evaluate the model's performance.

6.3 Data Visualization

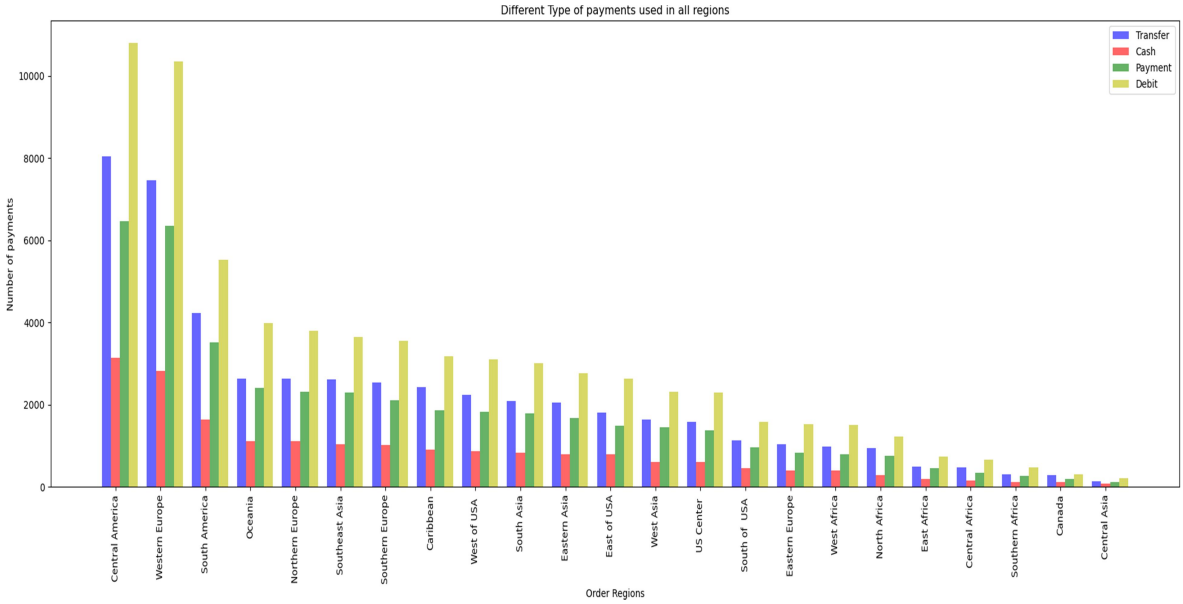
Sales Analysis:



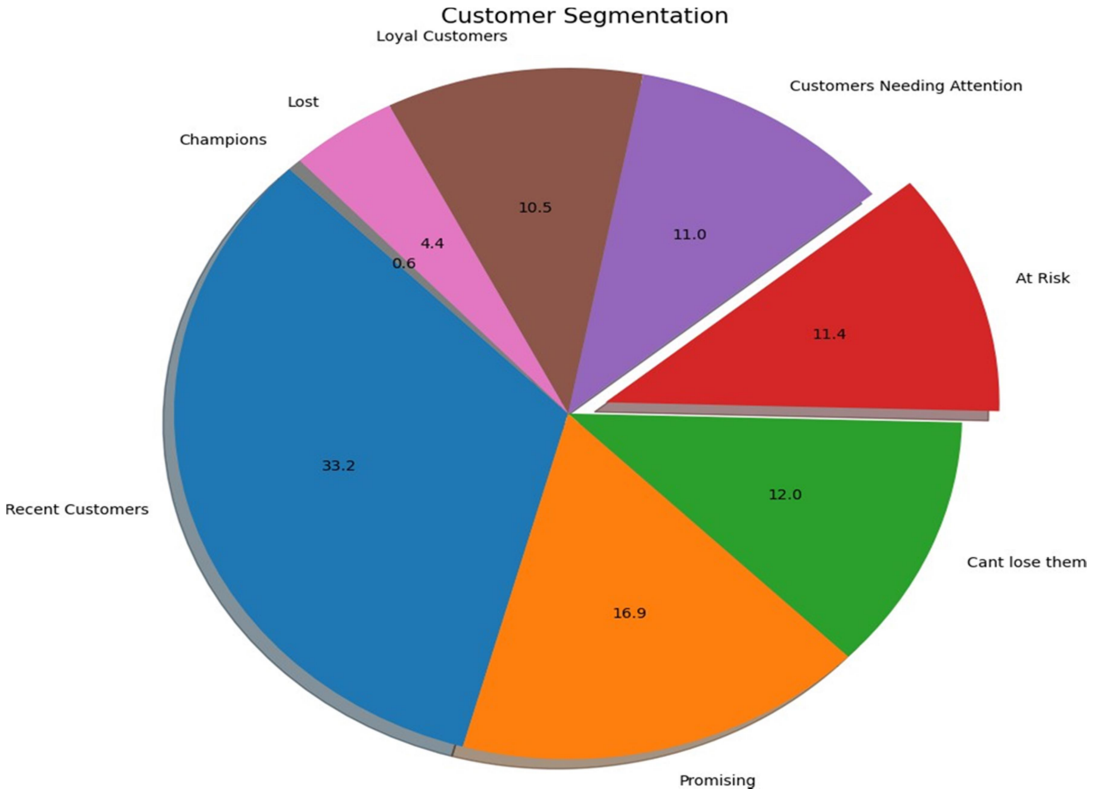
The above graph shows the total sales for all markets



The above graph shows the total sales in each category



The above graph shows the different types of payments used in all regions



The above graph shows the customer segementation

6.4. Data Analysis and Modelling

Feature Extraction

Feature extraction refers to the process of selecting essential and meaningful information from raw data to present it in a compact and more manageable format. The main objective is to convert the data into a set of features that can be utilized by machine learning algorithms to make predictions or conduct other analytical tasks. This is frequently achieved by applying techniques such as PCA, LDA, and filters to decrease the dimensionality of the data while retaining the most significant information. The outcome is a more precise and effective model.

Decision Tree Algorithm;

Decision Trees are a powerful machine learning approach that can handle a variety of supervised learning tasks, such as classification and regression. Although they can be applied to both tasks, they are typically used for classification problems. These classifiers have a tree-like structure, with nodes that represent the features of the dataset, branches that indicate decision rules, and leaves that represent the final classification outcomes.

A Decision Tree is a type of supervised learning algorithm used for classification and regression problems. It is composed of two types of nodes: Decision Nodes and Leaf Nodes. Decision Nodes make decisions based on the dataset's features and have multiple branches, while Leaf Nodes represent the final output of those decisions and do not have any further branches. Decision Trees offer a visual representation of all potential solutions to a problem or decision based on given conditions. Its name comes from its tree-like structure that begins with a root node and extends into branches. We use the Classification and Regression Tree (CART) algorithm to construct Decision Trees. This algorithm poses a question and partitions the tree into subtrees based on the response (Yes/No).

Root Node : The root node initiates the decision tree and represents the entire dataset, which is then divided into homogeneous subsets.

Leaf Node: It represents the final output of the decision tree, and no further segregation is possible **after getting a leaf node.**

Splitting: Splitting involves dividing the decision or root node into sub-nodes based on specific conditions.

Branch/Sub Tree: A tree formed by splitting the tree.

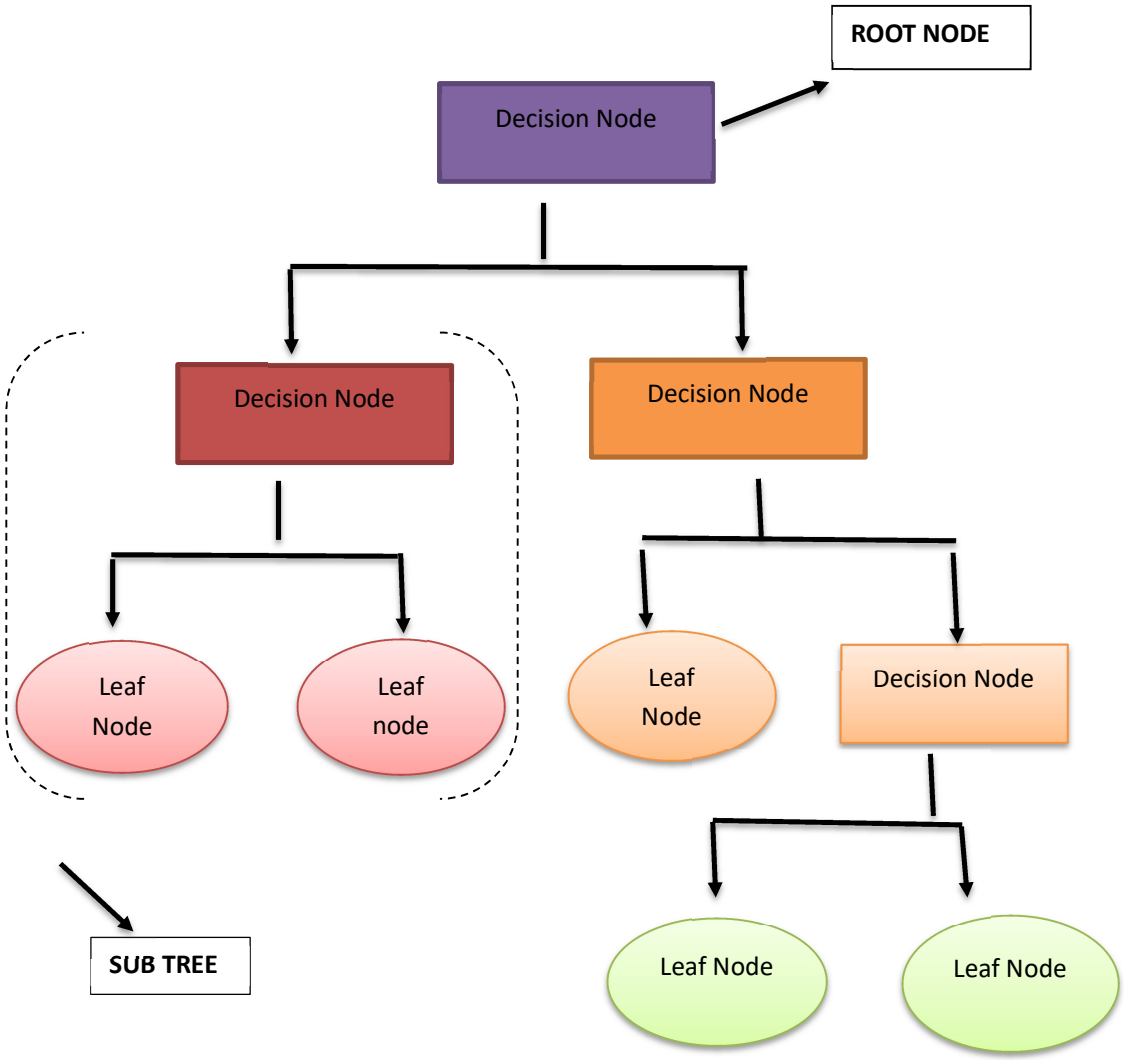
Pruning: Pruning involves removing unnecessary branches from the decision tree.

Parent/Child node: The tree's root node is the parent node, and any other nodes that derive from it are considered child nodes.

Working of the algorithm :

To determine the class of a given dataset using a decision tree, the algorithm starts at the root node and compares the attribute value of the record to the root attribute value. Then, based on the comparison, the algorithm follows the corresponding branch to the next node. This process is repeated for the sub-nodes of the next node until it reaches the leaf node of the tree. The complete process of the algorithm is demonstrated in the following algorithm

1. The decision tree algorithm for predicting the class of a dataset involves the following steps:
2. First, start at the root node S, which represents the entire dataset.
3. Next, select the best attribute in the dataset using an Attribute Selection Measure (ASM).
4. Then, create subsets of S based on the possible values for the best attribute.
5. Generate a decision tree node that includes the best attribute.
6. Finally, repeat steps 2-4 recursively for each subset of the dataset until a leaf node is reached, and no further classification can be made.



The decision tree algorithm is based on a recursive process that divides the data into subsets according to the feature values. The aim is to obtain subsets that are as homogeneous as possible, with only data points belonging to the same class. The algorithm can handle numerical and categorical data and missing values. In this project, the decision tree algorithm was utilized for feature extraction. It was applied to detect patterns in the data that suggest the occurrence of fraud or late delivery. The algorithm was used to identify the essential features that contribute to the prediction of these outcomes. The resulting tree-like model represents

the data, with internal nodes denoting features and leaf nodes representing decisions or class labels.

Multilayer perceptron Classifier (MLP):

Multi-layer perceptron (MLP) is a type of feed-forward artificial neural network that contains multiple hidden layers. MLP is often used for classification problems, where the number of neurons in the input layer is determined by the number of features, and the number of neurons in the output layer is equal to the number of classes to be classified. The hidden layers in between the input and output layers are fully connected and are trained using backpropagation. During forward propagation, the network calculates the output of each layer based on an activation function and corresponding weight and bias values.

MLP-based feature extraction has been successfully used in various domains such as image recognition, speech recognition, and natural language processing, and it has the advantage of automatically learning and extracting discriminative features without requiring handcrafted feature engineering. However, one potential drawback of MLP-based feature extraction is its sensitivity to hyperparameters, such as the number of hidden layers, number of nodes per layer, and activation functions.It is important to carefully tune these hyperparameters to achieve optimal performance

$$[k] = W[k]A[k-1] + b[k] \quad \ldots\ldots\ldots\ldots (1)$$

where $Z[k]$ is the output matrix of layer, $W[k]$ is weight matrix that connects layer k to $k-1$, $A[k-1]$ is the activation matrix of the previous layer $k-1$ and $b[k]$ is the bias vector of layer k

$$A[k] = g(Z[k]) \quad \ldots\ldots\ldots\ldots\ldots (2)$$

where $A[k]$ is the activated output matrix.

The method we suggest for use involves the Rectified Linear Unit (ReLU) activation function for the hidden layer and Softmax activation function for the final output layer. The ReLU activation function, as outlined in equation 3, is a simple function that sets negative values to zero. Conversely, the Softmax activation function, as defined in equation 4, is frequently utilized for multi-class classification problems.

It overcomes the limitations of the sigmoid function for such problems and ensures that the probabilities in the output layer add up to 1. The Softmax function assists in determining the most likely prediction.

$$a_{relu} = \max(0, z) \ldots\ldots\ldots (3)$$

$$a_{soft} = \frac{e^{z_i}}{\sigma_{j=1}^{J} e^{z_j}} \; for \; i \in [1,j] \qquad (4)$$

where j refers to the class number, $z\,i$ denotes the $i\,t\,h$ output value.

The loss function is employed to compute the difference between the predicted and actual values. This difference is then used in the backpropagation algorithm to adjust the weights.

Specified MLP Classifier

This study employs a Multi-Layer Perceptron (MLP) classifier with two hidden layers, each containing 128 neurons, and uses the ReLU activation function. Batch normalization is added after each hidden layer for regularization. The preprocessed data and selected features are fed into the input layer, and the model is trained through forward and backward propagation. The output layer uses the Softmax activation function to produce the probability of each class. During the prediction stage, the argmax function is used to find the highest probability and return its index. Adam's optimization algorithm with a learning rate of 0.0003, batch size of 64, and 300 epochs is used for training, and early-stopping is applied to prevent overfitting. The early-stopping parameter is set to 30, meaning that training is stopped when the validation set's loss has not decreased for 30 consecutive epochs to avoid any overfitting, and the best model parameters are restored.

Ensemble Feature Selection with DT and MLP with Logistics regression for classification

The first step of our feature extraction method involves an ensemble feature selection technique that combines decision tree and MLP with logistic regression classifier. This technique is applied only to 53 attributes, and the input training set is pre-processed to remove duplicate data before applying the filter or ensemble feature selection methods. To retain the most important features, different thresholds are chosen based on the importance ranking by each method. Features with importance values greater than the threshold are kept, while those with lower values are removed. We believe that there may be significant

features in both reduced feature subsets based on the Decision tree and MLP metrics, so their union set is used to optimize the features further.

Model Training and Validation

Optimal Feature Subset:

An optimal feature subset is a set of features that accurately represents the data and improves the performance of a machine learning model. Including irrelevant or redundant features may lead to overfitting, while selecting too few features may result in underfitting. Selecting the optimal feature subset involves evaluating the performance of the model with different feature combinations.

Various techniques are used for feature extraction, including filter methods, wrapper methods, and embedded methods. Filter methods evaluate the importance of each feature independently of the model, and select the top-ranked features based on certain criteria, such as correlation or mutual information. Wrapper methods evaluate the performance of the model with different subsets of features using a specific machine learning algorithm. Embedded methods incorporate feature extraction into the model training process, for example by adding regularization terms to penalize the inclusion of unnecessary features. After selecting a feature subset, it is important to validate the model performance using a separate test set. Model evaluation metrics like accuracy, precision, recall, and F1-score can be used to assess the quality of the model's predictions and determine the need for further feature extraction.

Logistic Regression Classifier

Logistic Regression is a statistical technique that can be applied to datasets with one or more independent variables that determine an outcome. The outcome is binary, meaning it has only two possible values. This method is commonly used to predict a binary outcome, such as Yes/No or True/False, based on a set of independent variables. In Supply Chain, Logistic Regression can be used to predict the likelihood of delayed delivery or supply chain fraud by using input features such as order history, shipping details, and vendor information, among others. By predicting the probability of a delay or fraud event, high-risk orders can be flagged for further investigation. Fraud is identified when the probability exceeds 0.5, while

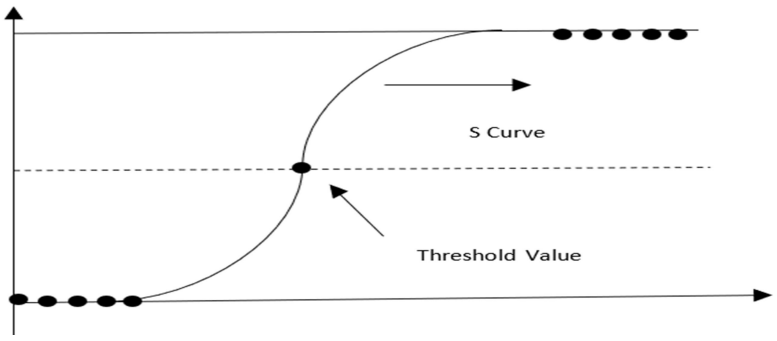
probabilities below 0.5 indicate no fraud. Logistic Regression can be represented by the following equation:

$$(x)^n = {}_{(e)} + b'X / \ 1 + (e)^{b+b\ X}$$

P is the Probability , is the Input data set , b and b' are the corresponding coefficients calculated using Maximum-likelihood estimation while training.

Logistic regression is a type of linear model used for classification problems. In model training, the logistic regression classifier learns the relationship between input features and output labels from a labeled dataset. The logistic regression classifier uses a sigmoid function to transform the output into a probability score, which represents the likelihood of each label. During model validation, the logistic regression classifier's performance is evaluated on an independent dataset that it has not seen before.

The logistic regression classifier's accuracy is measured using metrics such as precision, recall, F1 score, and ROC curve. To improve the logistic regression classifier's performance, various techniques such as regularization, feature extraction, and hyperparameter tuning can be used. Regularization helps prevent overfitting by adding a penalty term to the loss function, which reduces the complexity of the model. Feature extraction involves choosing the most relevant input features that contribute the most to the classifier's prediction The input features include variables such as order history, shipping details, vendor information, and others. Logistic Regression is used to predict the probability of a delay or fraud event, which can then be used to flag high-risk orders for further investigation. Fraud occurs if the probability exceeds 0.5; Otherwise, the transaction is considered as not being fraudulent



HYBRID CLASSIFIER

Class Definition:

This is a class definition for a Hybrid Classifier that inherits from two scikit-learn base classes, BaseEstimator and ClassifierMixin. The init method initializes two classifiers, a decision tree classifier (dt) and a multi-layer perceptron classifier (mlp), which can be passed as arguments to the constructor or default to None.

```
class HybridClassifier(BaseEstimator, ClassifierMixin):
    def __init__(self, dt=None, mlp=None):

        self.dt = DecisionTreeClassifier() if dt is None else dt
        self.mlp = MLPClassifier() if mlp is None else mlp
        self.final_classifier = None
```

The final_classifier attribute is set to None and will be used later to store the final trained classifier. This class can be used to create a hybrid classifier that combines the strengths of both decision tree and MLP classifiers. The idea behind this approach is to use the decision tree classifier to make simple decisions and then use the MLP classifier to make more complex decisions based on the decision tree's output. The final prediction is made by combining the output of both classifiers.

Fit Method:

The fit method is used to train the hybrid classifier on the input data X and target labels y. First, both the decision tree and MLP classifiers are fit to the data using the fit method of each classifier.

```
def fit(self, X, y):

        # Fit both models
        self.mlp.fit(X, y)
        self.dt.fit(X, y)

        # Extract predictions from both models
        dt_pred = self.dt.predict(X)
        mlp_pred = self.mlp.predict(X)

        # Concatenate the predictions
        X_new = np.concatenate([dt_pred.reshape(-1, 1), mlp_pred.reshape(-1, 1)], axis=1)

        # Train a final classifier
        self.final_classifier = LogisticRegression().fit(X_new, y)
        return self
```

Then, the predictions from both classifiers are extracted for each input sample using the predict method. The predictions from the decision tree and MLP classifiers are concatenated horizontally using np.concatenate to form a new input feature matrix X new with two columns. The reshape method is used to ensure that the predictions have the same shape for concatenation. Finally, a logistic regression classifier is trained on the concatenated feature matrix X new and target labels y using the fit method of LogisticRegression(). The trained logistic regression classifier is stored in the final classifier attribute of the hybrid classifier instance. The fit method returns the hybrid classifier instance, allowing method chaining with other methods of the class.

Predict Method:

The predict method is used to make predictions on new input data X. First, the predictions from both the decision tree and MLP classifiers are extracted for each input sample using the predict method of each classifier.

```
def predict(self, X):
        # Extract predictions from both models
        dt_pred = self.dt.predict(X)
        mlp_pred = self.mlp.predict(X)

        # Concatenate the predictions
        X_new = np.concatenate([dt_pred.reshape(-1, 1), mlp_pred.reshape(-1, 1)], axis=1)

        # Make final predictions
        return self.final_classifier.predict(X_new)
```

Then, the predictions from the decision tree and MLP classifiers are concatenated horizontally using np.concatenate to form a new input feature matrix X_new with two columns. The reshape method is used to ensure that the predictions have the same shape for concatenation. Finally, the predict method of the trained logistic regression classifier stored in self.final_classifier is used to make final predictions on the concatenated feature matrix X_new. The predicted target labels are returned by the predict method.

14 Hybrid Classifier

In [54]:
```
from sklearn.tree import DecisionTreeClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.datasets import make_classification
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.base import BaseEstimator, ClassifierMixin
from sklearn.linear_model import LogisticRegression


class HybridClassifier(BaseEstimator, ClassifierMixin):
    def __init__(self, dt=None, mlp=None):

        self.dt = DecisionTreeClassifier() if dt is None else dt
        self.mlp = MLPClassifier() if mlp is None else mlp
        self.final_classifier = None

    def fit(self, X, y):

        # Fit both models
        self.mlp.fit(X, y)
        self.dt.fit(X, y)

        # Extract predictions from both models
        dt_pred = self.dt.predict(X)
        mlp_pred = self.mlp.predict(X)

        # Concatenate the predictions
        X_new = np.concatenate([dt_pred.reshape(-1, 1), mlp_pred.reshape(-1, 1)], axis=1)

        # Train a final classifier
        self.final_classifier = LogisticRegression().fit(X_new, y)
        return self

    def predict(self, X):
        # Extract predictions from both models
        dt_pred = self.dt.predict(X)
        mlp_pred = self.mlp.predict(X)

        # Concatenate the predictions
        X_new = np.concatenate([dt_pred.reshape(-1, 1), mlp_pred.reshape(-1, 1)], axis=1)

        # Make final predictions
        return self.final_classifier.predict(X_new)
```

Evaluation Metrics

In machine learning classification tasks, there are typically four cases based on the "one versus all" principle, where:

True Positive (TP): represents instances that are correctly classified positive samples.

False Negative (FN): represents instances that are incorrectly classified positive samples.

False Positive (FP): represents instances that are incorrectly classified negative samples

True Negative (TN): represents instances that are correctly classified negative samples.

Precision is a performance metric that measures the proportion of true positives to all positive predictions, and it is calculated using the following formula:

$$Precision = \mathbf{TP} / TP + FP$$

The recall measures the proportion of accurately classified positive samples to all the samples that are actually positive and predicted as positive, and is expressed by the equation below.

$$Recall\ (True\ Positive\ Rate) = TP / T\ P + F\ N$$

The F1 score, as shown in the equation below, is a metric that measures a model's accuracy by considering both precision and recall. It represents the harmonic mean of precision and recall and serves as a useful performance indicator when dealing with imbalanced multi-class data, addressing the limitations of precision and recall.

7.CODE STRUCTURE

The entire project is divided into 3 Modules for our convenience and better understanding.

A. Sales Analysis
B. Late delivery prediction
C. Fraud Dectection

A.SALES ANALYSIS:

The purpose of this module is to analyze data obtained from the Kaggle resource. Data cleaning and structuring are conducted to facilitate better analysis. This process entails identifying and correcting mistakes, discrepancies, and inaccuracies in the initial data. The primary goal of data cleaning is to ensure the precision, completeness, and consistency of the data, enabling it to be utilized for analysis or other purposes. During the data cleaning process, several tasks may be performed, such as:

This section is focused on how to effectively handle various data quality issues that may arise during data analysis. These issues include missing data, duplicates, inconsistencies, outliers, and formatting problems. To handle missing data, the first step is to identify the missing values and then decide whether to delete them or use statistical methods to fill them in. Removing duplicates requires identifying and removing records that are exact copies of each other. Resolving inconsistencies requires identifying and fixing any errors or inconsistencies in the data, such as spelling mistakes, date format discrepancies, or invalid values. Removing outliers involves identifying and removing data points that are significantly different from the majority of the data. Finally, addressing formatting issues requires standardizing the format of the data, such as converting text to lowercase or uppercase, to ensure consistency and accuracy. Handling data type errors: This involves identifying and correcting data type errors, such as converting numerical data stored as text to a numeric format.

This Module has below code structure

1. Importing all the required libraries

2. Importing the data set
3. Removing the unwanted data
4. Removing the anomalies
5. Displaying the Heat map for correlation matrix to know which feature is more important for the data prediction
6. Performing/visualizing sales analysis
    - Total sales for all the markets
    - Total sales for all the regions
    - Total sales , average price , average sales based on category name
    - Product price vs sales per customer
    - Different types of payments used in all regions
    - Products with most loss
    - Regions with most loss
    - Regions with highest frauds
    - Top 10 products with highest fraud detection
    - Top 10 highest fraud customers
    - Top 10 products with most late deliveries
    - Late delivered products used in all regions
    - Different types of shipping methods used in all regions
7. Customer Segmentation – we are doing this by using RFM score
    - Finding the RFM values and scores for segmentation
    - Based on that dividing the customers in to loyal, lost, champions, customers needing attention, at risk and recent customers.

B.LATE DELIVERY PREDICTION

The purpose of a late delivery prediction module is to estimate the likelihood of a package or shipment being delivered past its expected arrival time. In order to create such a module, relevant historical data on delivery times must be collected and preprocessed to make it suitable for analysis. Important predictive features for late delivery are then identified, and a machine learning model is trained to predict the probability of a late delivery using these selected features. The performance of the model is evaluated and then deployed, where it can be utilized to make predictions on new data. The late delivery prediction module is intended to support logistics firms, shipping carriers, or e-commerce retailers in enhancing

their delivery procedures, decreasing costs, and improving customer satisfaction by anticipating and handling possible late deliveries before they occur.

This Module has below code structure

1. Importing all the required libraries
2. Importing data set
3. Cleaning the data
4. Displaying the features in the given data set
5. Displaying the statistical summary of the dataset after structuring
6. Encoding categorical Data
7. Handling the Missing Data
8. Feature Scaling
9. Splitting the dataset in to testing set and training set
10. Machine learning Algorithms training
    - Logistic regression algorithm
    - K nearest neighbors Algorithm
    - Random Forest Algorithm
11. Making predictions on test data
12. Model Evaluation for all the algorithms using accuracy score metric
13. Visualizing the confusion matrix for all the algorithms
14. Defining the HYBRID Classifier
15. Performing the Late Delivery Prediction and visualizing using confusion matrix

C.FRAUD DECTECTION

A fraud detection machine learning model uses machine learning algorithms to identify fraudulent activities in financial transactions or other domains. identify fraudulent activities in financial transactions or other domains. The model is designed to analyze large amounts of data, identify patterns and anomalies, and provide predictions or alerts when fraudulent behavior is detected. It involves collecting and preprocessing data, creating new features, training a machine learning model, validating its performance, and deploying it in a production environment. The goal is to accurately and timely detect fraudulent activities, prevent financial losses, and protect individuals and businesses from harm.

This Module has below code structure

1. Importing all the required libraries
2. Importing data set
3. Cleaning the data
4. Displaying the features in the given data set
5. Displaying the statistical summary of the dataset after structuring
6. Encoding categorical Data
7. Handling the Missing Data
8. Feature Scaling
9. Splitting the dataset in to testing set and training set
10. Machine learning Algorithms training
    - Logistic regression algorithm
    - K nearest neighbors Algorithm
    - Random Forest Algorithm
11. Making predictions on test data
12. Model Evaluation for all the algorithms using accuracy score metric
13. Visualizing the confusion matrix for all the algorithms
14. Defining the HYBRID Classifier
15. Performing the Fraud Detection and visualizing using confusion matrix

8. ANALYTICAL OUTCOMES

To prevent overfitting of features, we removed duplicate samples from the training set before using ensemble feature selection with DT and MLP. We applied this approach to the training set and ranked the importance of 53 numeric features. Table II presents the performance metrics of our hybrid model. As the Data Co Global dataset has several imbalanced classes, the f1 score is a more appropriate measure to assess the performance of each class.

	precision	recall	f1-score	support
0	0.99	0.99	0.99	16151
1	0.99	0.99	0.99	19953
accuracy			0.99	36104
macro avg	0.99	0.99	0.99	36104
weighted avg	0.99	0.99	0.99	36104

The higher F1 score indicates that the model is performing better in predicting fraud (see Table II and III). It indicates a good balance between precision and recall, meaning that the model is making relatively few false positive predictions (i.e., non-fraudulent transactions that are incorrectly classified as fraud) while also correctly identifying a high proportion of actual fraudulent transactions (high recall). Therefore, our hybrid model shows higher F1 score is considered to be an indicator of better model performance in fraud detection, as it indicates that the model is making fewer incorrect predictions and is more effective at identifying fraud transactions and late delivery prediction.

	precision	recall	f1-score	support
0	0.99	0.99	0.99	35274
1	0.66	0.68	0.67	830
accuracy			0.98	36104
macro avg	0.82	0.83	0.83	36104
weighted avg	0.98	0.98	0.98	36104

Figure 4 and 5 depict the confusion matrix for multi-classification, wherein the predicted label is represented along the horizontal axis, and the true label is represented along

the vertical axis. The confusion matrix reveals some misclassifications among different classes. The matrix's rows correspond to the actual class labels, while the columns correspond to the predicted class labels.

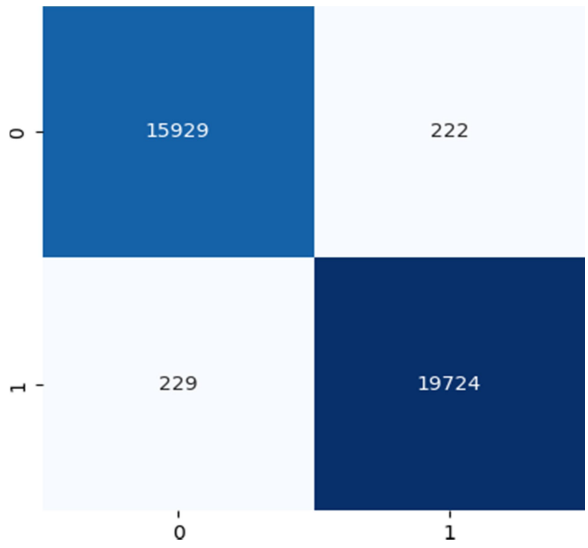


Fig 4. Confusion Matrix for Late Delivery Prediction

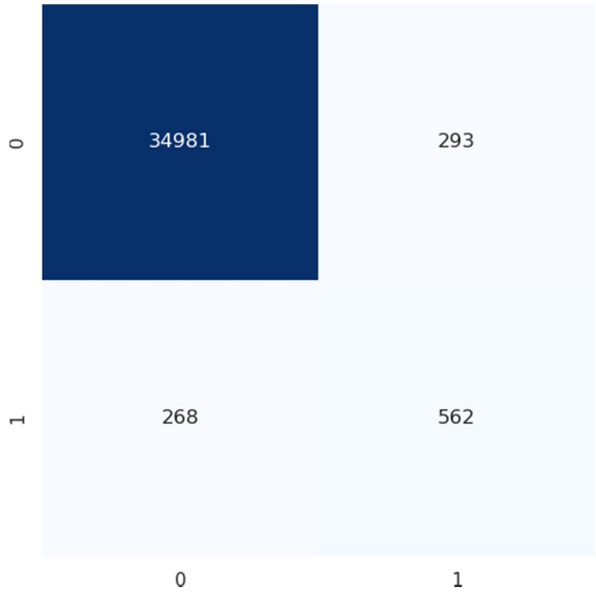
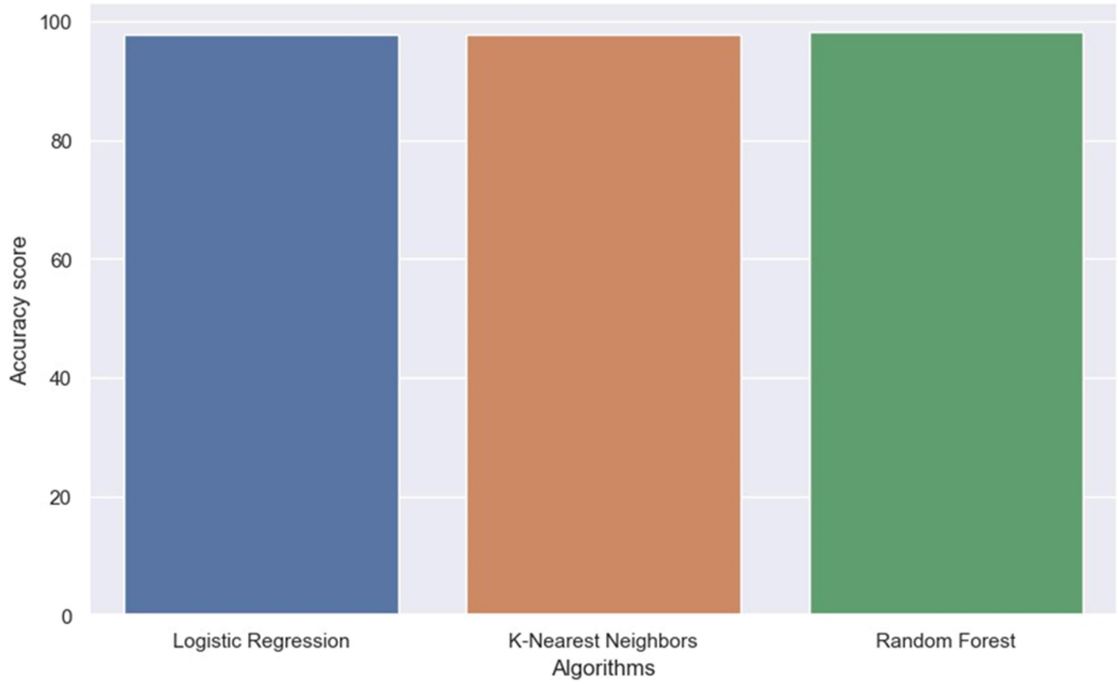
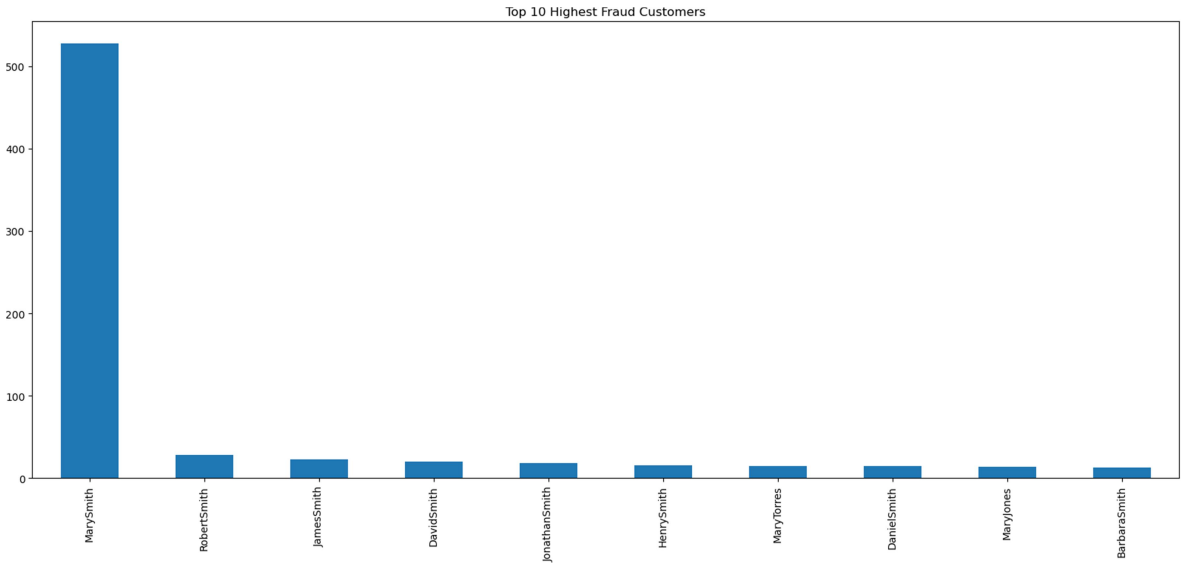


Fig 5. Confusion Matrix for fraud Transactions

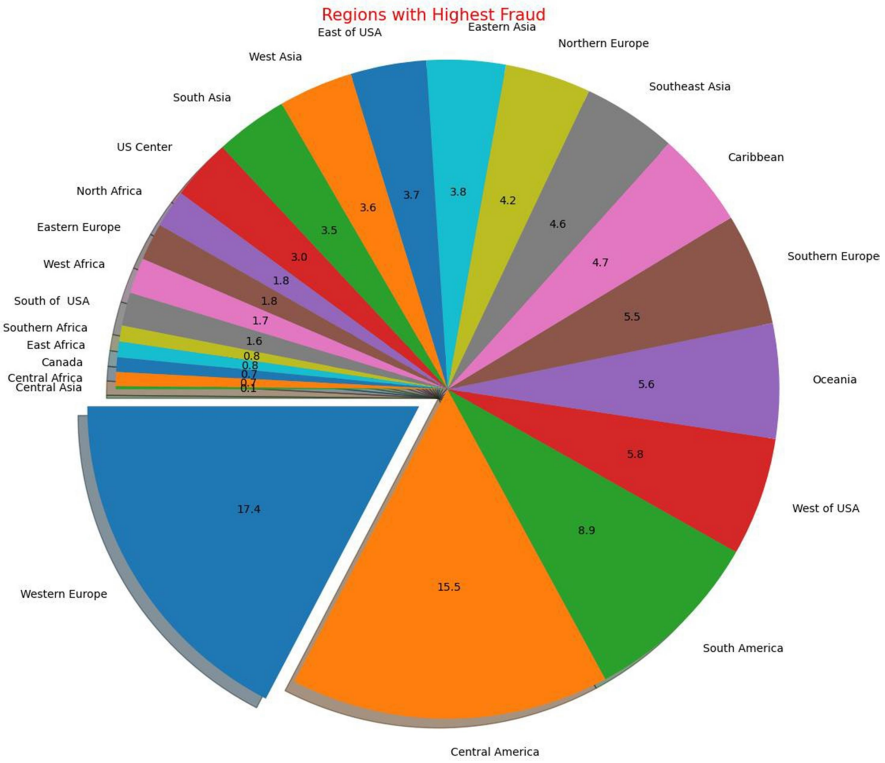
Fraud Prediction Chart :



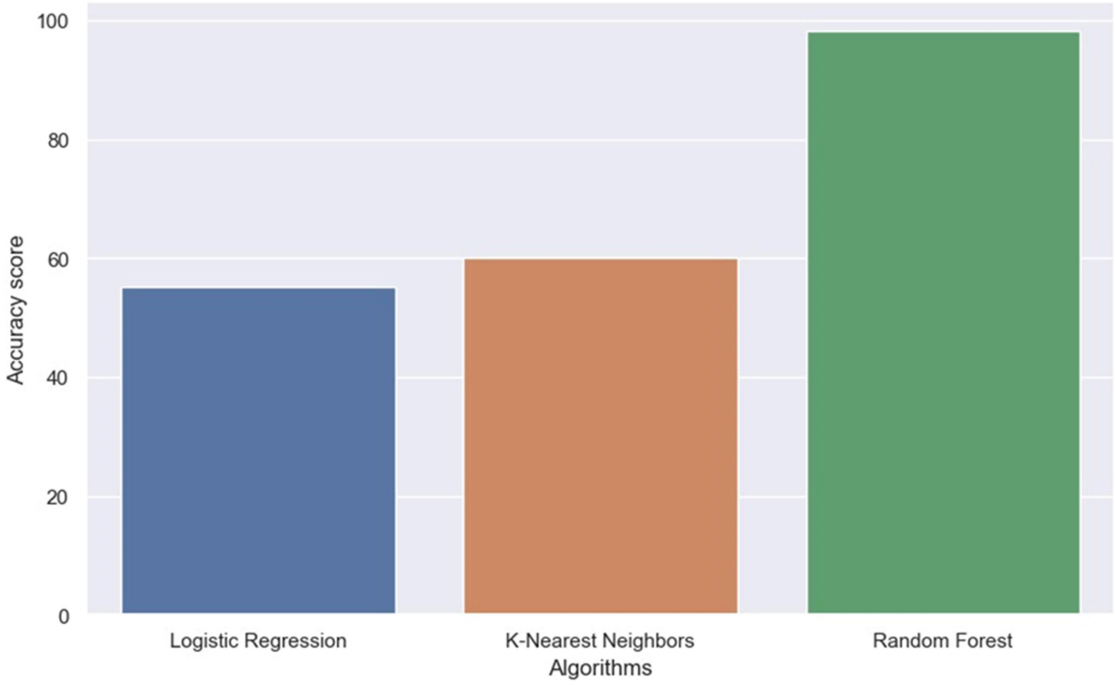
The above graph shows the accuracy of different algorithms in the fraud prediction



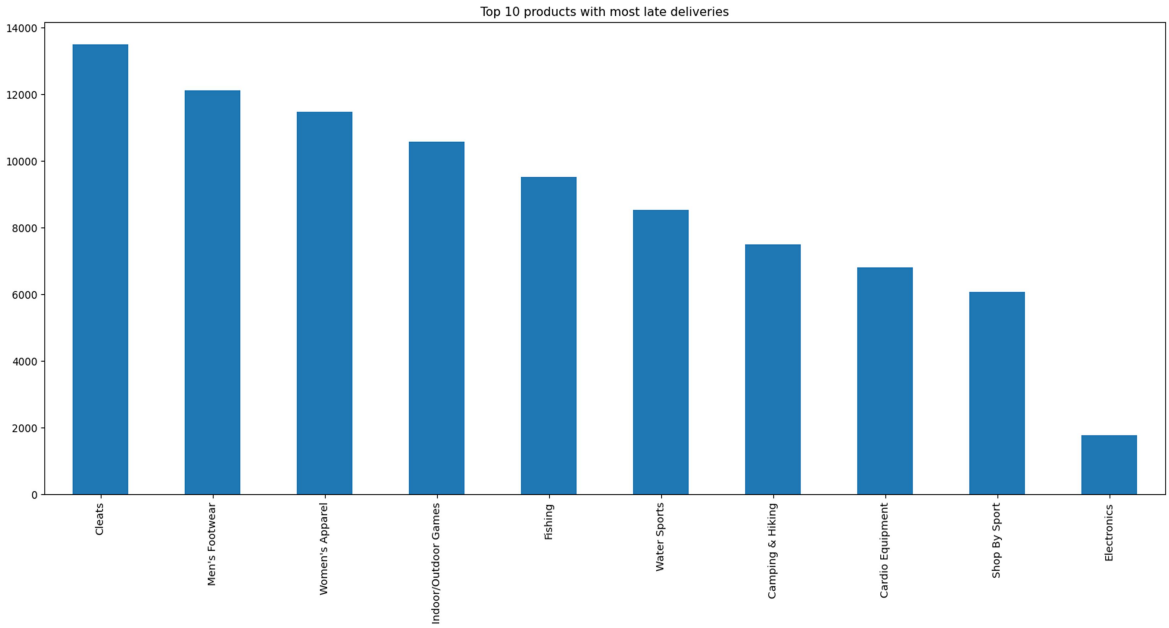
The above graph shows that detected top 10 highest fraud customers



The above graph shows the region with highest fraudLate delivery Prediction Chart



The above graph shows the accuracy of different algorithms in the Late Delivery Prediction



The above graph shows that the top 10 products with most late deliveries

9.CONCLUSION AND FUTURE WORK

This study explores the potential applications of machine learning and artificial intelligence techniques in predictive modeling for supply chain management, with a specific focus on fraud detection. These techniques have been demonstrated to be effective in detecting patterns and anomalies that may indicate fraudulent activity, which can help organizations improve their fraud detection capabilities. The use of machine learning in supply chain management has grown in recent years, owing to its ability to uncover patterns and anomalies in data. This is facilitated by the increased availability of data from a range of sources, such as sensor data from Internet of Things (IoT) devices, and the improved computational power and storage capabilities required to process and analyze this data. Machine learning algorithms such as decision trees, random forests, and neural networks can be employed to identify patterns and anomalies in this data, which can be used to optimize supply chain processes, including inventory management, demand forecasting, fraud detection, and late delivery detection.

It also explores the potential of machine learning (ML) and artificial intelligence (AI) in improving supply chain management, particularly in detecting fraud. ML's ability to adapt and learn from new data makes it an appealing solution for supply chain management, as it can continuously improve its performance. The study demonstrates that businesses can use ML and AI to optimize their supply chains, enhance their efficiency, and reduce risks associated with fraud and human error. The project employed logistic regression as the baseline classifier, with decision tree and multi-layer perceptron (MLP) models for fraud detection and late delivery prediction. All three models were successful in predicting fraudulent transactions, but the hybrid model had the highest accuracy (99%), while the logistic regression model had the lowest (96.6%). These results indicate that ML can be a useful tool in identifying supply chain fraud, but further research is necessary to improve prediction accuracy. As these technologies continue to evolve, they are expected to have a significant impact on the management of supply chains.

Before implementing ML and AI in supply chain management, it is crucial to carefully consider their potential advantages and drawbacks. One of the primary challenges is

the difficulty in obtaining large amounts of data, particularly in an environment where data is often scattered and isolated. Additionally, the implementation and maintenance costs associated with these technologies can be substantial. Moreover, there is a possibility that ML and AI may not always detect fraudulent activities or may yield false positive outcomes. Thus, it is essential to evaluate the potential benefits and risks of these technologies and develop a well-defined strategy before introducing them to the supply chain.

The performance evaluation of the model presented in this study must take into account certain limitations. One of the primary restrictions is the importance of the sequential model's layer order for the model's correct functioning. The sequential model is a type of neural network that comprises multiple layers, each responsible for a specific task. The layer sequence plays a crucial role in the model's ability to learn the relevant features and make accurate predictions. If the layer order is incorrect, the model may not be able to extract the appropriate features and may fail to perform as expected. Additionally, the sequential model may not be the best option for dealing with datasets with large numbers of features. A feature vector is a collection of characteristics or features used to describe the data. In hybrid models, the feature vector's size can significantly impact the computational cost required for model training. As the data size grows, the computational cost also increases exponentially, making it challenging and computationally expensive to classify large datasets using this model. Thus, if the dataset is exceptionally large, alternative models should be considered as this model may not be effective in classification.

In conclusion, due to the rise of digital commerce, there is a need for exploring how ML and AI can improve fraud detection in the supply chain. Traditional methods based on rule-based models are insufficient in adapting to the fast-paced changes in the digital world. However, ML and AI can create adaptable models that continuously update and detect new fraud patterns, while also responding to changes in the supply chain. In addition, predictive models can be developed using these technologies to prevent fraud proactively. Thus, future research should focus on investigating the application of ML and AI to improve fraud detection and late delivery prediction in the supply chain.

10.REFERENCES

[1] George Baryannis, Samir Dani, Grigoris Antonioua, (2019) “Predicting supply chain risks using machine learning: The trade-off between performance and interpretability”

[2] Fabian-Vinicio Constante-Nicolalde, Paulo Guerra-Teran, Jorge-Luis Perez-Medina (2020) “Fraud Prediction in Smart Supply Chains Using Machine Learning Techniques”

[3] [Dianhui Mao, Fan Wang, Zhihao Hao, and Haisheng Li (2018) “Credit Evaluation System Based on Blockchain for Multiple Stakeholders in the Food Supply Chain”

[4] Emmanuel Ileberi, Yanxia Sun & Zenghui Wang, (2022) “A machine learning based credit card fraud detection using the GA algorithm for feature selection”

[5]] Khizar Abbas, Muhammad Afaq,Talha Ahmed Khan and Wang-Cheol Song (2020) “A Blockchain and Machine Learning-Based Drug Supply Chain Management and Recommendation System for Smart Pharmaceutical Industry”

[6] Xiao Z, Lim MK. (2020) “A systematic review of the research trends of machine learning in supply chain management”

[7] Zhou Y, Song X, Zhou M. (2021) “Supply Chain Fraud Prediction Based on XGBoost Method”

[8] Schroeder M, Lodemann S. (2021) “A Systematic Investigation of the Integration of Machine Learning into Supply Chain Risk Management. Logistics”

[9] Wan F. (2021) “XGBoost Based Supply Chain Fraud Detection Model” International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)

[10] Zhang Y, Tong J, Wang Z, Gao F. (2020) “Customer Transaction Fraud Detection Using XGboost Model”

[11] Rodriguez-Aguilar R, Marmolejo-Saucedo JA. (2019) “Structural Dynamics and disruption events in Supply Chains using Fat Tail Distributions”

[12] Camossi E, Dimitrova T, Tsois A. (2012) “Detecting Anomalous Maritime Container Itineraries for Anti-fraud and Supply Chain Security” European Intelligence and Security Informatics Conference

[13] Zhang W, Gao F. (2011) An Improvement to Naive Bayes for Text Classification.

[14] Lo SK, Xu X, Wang C, Weber I, Rimba P, Lu Q, et al. (2019)

"Digital-Physical Parity for Food Fraud Detection".

[15] Shahbazi Z, Byun YC. A (2021) "Procedure for Tracing Supply Chains for Perishable Food Based on Blockchain, Machine Learning and Fuzzy Logic"

[16] Bagga S, Goyal A, Gupta N, Goyal A. (2020) "Credit Card Fraud Detection using Pipeling and Ensemble Learning"

[17] Lezoche M, Hernandez JE, Alemany Díaz M del ME, Panetto H, Kacprzyk J. (2020) "A survey of the supply chains and technologies for the future agriculture"

[18] Herland M, Khoshgoftaar TM, Bauder RA. (2018) "Big Data fraud detection using multiple medicare data sources"

[19] Johnson JM, Khoshgoftaar TM. (2019) "Medicare fraud detection using neural networks"

[20] Zhang G, Zhang X, Bilal M, Dou W, Xu X, Rodrigues JJPC. (2022) "Identifying fraud in medical insurance based on blockchain and deep learning".

[21] Dua P, Bais S. (2014) "Supervised Learning Methods for Fraud Detection in Healthcare Insurance"

[22] Bordoloi D, Singh V, Sanober S, Buhari SM, Ujjan JA, Boddu

R. (2022) "Deep Learning in Healthcare System for Quality of Service"

[23] Li H, Li W, Pan X, Huang J, Gao T, Hu L, et al. (2018) "Correlation and redundancy on machine learning performance for chemical databases: Correlation and Redundancy on Machine Learning Regressions"