In the complex world of supply chains, where products and services move from suppliers to customers, there's a growing problem: fraud. Fraud in supply chains can mean anything from fake orders to scams that cost businesses a lot of money and harm their reputation. In the past, companies used basic methods to catch these frauds, but as fraudsters get smarter, these old ways aren't working as well anymore. This is a big risk for businesses, as it can lead to losing money and trust.

BUSINESS PROBLEM

The main issue we're facing is finding and stopping fake orders in the supply chain. Our current system isn't good at spotting the little signs that something might be a scam. This means we could be losing money without even knowing it. Also, if we can't catch these frauds, people might start to think our supply chain isn't safe or reliable, which is bad for our business's image. To tackle this problem, our team is dedicated to developing a more sophisticated approach. We're focusing on implementing advanced data analytics and machine learning techniques to enhance our ability to detect fraudulent activities. By analyzing patterns and inconsistencies in the data, we aim to identify these deceptive transactions more effectively. This proactive strategy is not just about safeguarding our finances; it's also about ensuring the reliability and integrity of our supply chain. By strengthening our fraud detection capabilities, we're working towards maintaining a secure and trusted business environment.

KEY OBJECTIVES

Our primary objective is to significantly improve the detection and prevention of fraudulent orders within our supply chain. To achieve this, we are focusing on several key areas:

Integration of Advanced Data Analytics: We plan to integrate state-of-the-art data analytics into our system. This involves using sophisticated algorithms to analyze large sets of supply chain data. The goal is to identify unusual patterns and anomalies that could indicate fraudulent activities. By analyzing trends, transaction histories, and behavioral patterns, we can detect inconsistencies that traditional methods might miss.

Employment of Machine Learning Techniques: Machine learning will play a crucial role in our strategy. We aim to develop models that learn from historical data and continuously improve their accuracy in fraud detection. These models will be trained to recognize the characteristics of fraudulent transactions and alert the system when similar patterns emerge. This approach ensures that the system evolves and stays effective against new and sophisticated types of fraud.

Real-Time Fraud Detection and Alert System: A critical aspect of our objective is to establish a real-time monitoring system. This system will continuously scan transactions as they occur, enabling immediate detection and response to potential frauds. Quick identification allows us to take prompt action, minimizing financial loss and disruption to the supply chain.

Enhancing System Reliability and Security: By improving our fraud detection capabilities, we also aim to enhance the overall reliability and security of our supply chain. A secure supply chain fosters trust among all stakeholders, including suppliers, customers, and partners.

Data Privacy and Compliance: In implementing these advanced technologies, we are also committed to ensuring data privacy and compliance with relevant regulations. Safeguarding sensitive information is paramount, and our approach will include robust data protection measures.

Continuous Improvement and Adaptation: Finally, our objective includes a commitment to continuous improvement. The landscape of fraud is ever-changing, and our system must adapt accordingly. We plan to regularly review and update our approaches, incorporating feedback, and learning from industry trends.

EXPECTED OUTCOMES

Better Fraud Detection: We want to get better at finding possible scams, which will help us avoid losing money to fraud.

Safer Supply Chain: By quickly finding and dealing with fake orders, we can make our entire supply chain more secure.

More Efficient Operations: If we can spot frauds without having to check everything manually, we can do our work faster and more efficiently.

Better Reputation: By showing that we can handle fraud well, we'll build trust with our partners, like the people who supply us and our customers.

ABOUT THE CLASSIFIERS

We will use three classification models to make predictions about the supply chain data. After doing the analysis, we will compare the accuracy and performances of these models and choose the most suitable model for the task.

K_NN classification

The K-Nearest Neighbor (KNN) algorithm is a popular machine learning technique used for classification and regression tasks.

KNN classifier operates by finding the k nearest neighbors to a given data point, and it takes the majority vote to classify the data point. The value of k is crucial, and one needs to choose it wisely to prevent overfitting or underfitting the model. (*What Is the K-nearest Neighbors Algorithm?* | *IBM*, n.d.)

Guassian Naïve Bayes Classification

Gaussian Naive Bayes is a classification algorithm based on Bayes' theorem with an assumption of independence among predictors. It's particularly suitable for continuous data and assumes that the features follow a Gaussian (normal) distribution. The algorithm calculates the probability of a given data point belonging to a particular class based on the feature values. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Despite its simplicity and the assumption of independence, Gaussian Naive Bayes often performs well in practice and can be quite efficient, especially on small datasets.

Random Forest with Cross Validation

Random Forest constructs multiple decision trees by selecting random subsets of data and features. Each tree predicts the class label and the final prediction is determined by aggregating the results (e.g., via voting or averaging). When used with cross-validation (CV), it becomes a powerful tool for model evaluation and hyperparameter tuning. (Sklearn.Ensemble.Random Forest Classifier, n.d.)

Literature Review:

"DataCoSupplyChainDataset.csv," the dataset used in this study, provides an essential basis for tackling the growing problem of supply chain fraud. The data set contains the information about the customer and customer order i.e., customer ID, customer name, customer city, days for shipping, Benefit per order, etc. The limits of existing methods in detecting fraudulent activity highlight the need for advanced analytics and machine learning techniques as the supply chain landscape gets more complicated. The original 180,519 rows and 53 dataset columns were carefully cleaned up to remove redundant and unnecessary information. After going through this procedure, a refined dataset consisting of 180,519 rows and 36 columns was produced, highlighting how crucial high-quality data is to training machine learning models successfully.

The research indicates that there is a rising realization of the need for more sophisticated techniques to combat the widespread problem of fraud within supply chains, especially in

the area of e-commerce. Businesses face considerable hurdles due to the danger of fraudulent actions, such as frauds and bogus orders, as their products and services move through intricate supply networks. As fraudsters use more complex techniques and risk financial losses as well as damage to a company's brand, traditional ways of detecting fraud are becoming less effective. The stated business issue is that the current system has trouble identifying small indicators of possible fraud, making it difficult to identify fraudulent orders within the supply chain. In addition to posing financial concerns, this difficulty

The literature suggests using machine learning (ML) and advanced data analytics techniques to address this issue. Using machine learning (ML) techniques, integrating cutting edge data analytics, and establishing a real-time fraud detection and alarm system are the main goals. Businesses want to better detect fraudulent actions by analyzing trends and abnormalities in data by utilizing these technologies. The main objective is to protect money while also guaranteeing the supply chain's dependability and integrity. The body of research highlights how important it is to use these technologies to improve fraud detection capabilities in order to preserve a safe and reliable corporate environment.

supply chain management is one of the industries where organizations are increasingly using machine learning (ML) approaches to detect fraud. Applying machine learning (ML) is seen to be crucial for improving security procedures, identifying fraudulent activity, and protecting supply networks from possible attacks. Because ML models are trained to identify patterns and traits linked to fraudulent transactions, they may be continuously improved upon and adjusted to suit changing fraud scenarios. A variety of methods, each with advantages and disadvantages, are reflected in the application of techniques like Gaussian Naive Bayes, Random Forest, and K-Nearest Neighbors (KNN). The necessity for more precise, effective, and instantaneous identification of fraudulent transactions is what motivates the use of these models in the context of supply chain fraud detection.

METHODOLOGIES:

DATA LOADING:

The initial part of the code involves loading a dataset named " DataCoSupplyChainDataset.csv" into a data frame using the Pandas library. We use the read csv method for this. We print the data frame head to ensure the data has been loaded correctly. We see that the data consists of 180519 rows and 53 columns.

DATA PREPARATION:

The dataset contains duplicate data and also irrelevant features, there are also features with missing values so we remove the data that won't add any value to our analysis. After the removal of unnecessary elements, we are left with 180519 rows and 44 columns of data.

DATA VISUALIZATION:

To grasp the patterns and nuances within the dataset, aiding in feature selection, identifying outliers, and making informed decisions about preprocessing, ultimately laying the groundwork for a more effective and accurate model

DATA PREPARATION:

We separate the columns related to features (in the features variable) from the target variable (in the target variable) to prepare data for a machine-learning model. The features variable contains all columns except 'SUSPECTED_FRAUD' and 'Order Status', while the target variable holds the 'SUSPECTED_FRAUD' column specifically, presumably for training a model to predict suspected fraud.

Then we look for the null values in the data and we can consider NaN values as a separate class using LabelEncoder

Then we look for features that are highly correlated by generating the correlation matrix. We identify features that correlate a certain threshold (0.8) and then create a new set of features (features1) by removing these highly correlated features from the original set. This process can be beneficial in machine learning tasks where highly correlated features might introduce redundancy and negatively impact model performance.

We select features that have no significant correlation (either positive or negative) with the target variable 'SUSPECTED_FRAUD' based on the defined threshold, which appears to be 0.004 or -0.004. These selected features might potentially have weaker predictive power or relevance to the target variable compared to others in the dataset.

Identify and select features that have significant relationships with the target variable based on their p-values derived from the f_regression test, considering a significance level of 0.05.

After this, we just have 180519 rows and 36 columns.

Customer ZipCode and customer state have a high correlation with Customer Country as they are geographical inputs. we can omit these features and keep only the customer's country.

MODEL TRAINING AND EVALUATION

Data is ready with features finalized and cleaned. We will be training different models to predict fraudulent orders and compare the performance of each.
1) LogisticRegression
2) RandomForestClassifier
3) KNeighborsClassifier
4) GaussianNB
5) DecisionTreeClassifier
Will use CV to increase the robustness of the model.

LOGISTIC REGRESSION:

The dataset is split into training and testing sets using the train_test_split function. Logistic regression () from sci-kit-learn is instantiated and trained using the training set. This step is crucial for enabling the model to learn patterns and relationships between features and labels.
The model achieves a high accuracy of 97.75%, indicating that it correctly predicts the majority of instances. However, the precision, recall, and F1 scores are all zero, suggesting that the model fails to identify any instances of fraudulent orders, indicating poor performance in capturing positive cases. Therefore, despite the high accuracy, the model is not effective in addressing the specific objective of identifying fraudulent orders.

RANDOM FOREST CLASSIFIER:

It first divides the dataset into training and testing subsets, trains the classifier using the training data, and subsequently predicts outcomes on the test set. Evaluation metrics such as accuracy, precision, recall, F1 score, and a confusion matrix are then calculated to gauge the classifier's performance. Finally, these metrics are displayed, offering a comprehensive view of how well the model is performing in classifying the data.
The Random Forest classifier exhibits exceptional performance with an accuracy of 99.57%, characterized by high precision (95.77%) in correctly identifying positive instances and a notable recall (85.29%) capturing a substantial proportion of actual positive cases, resulting in a balanced F1 score of 90.23%. The confusion matrix confirms minimal misclassifications and a significant count of true positives, emphasizing the model's robustness in effectively distinguishing between classes.

KNN CLASSIFIER:

KNN classifier from sci-kit-learn is then instantiated and trained using the training set.

We use k = 1, 2 and 3.
The model achieves an accuracy of 97.71%, demonstrating moderate overall performance, characterized by a precision of 55.73%, indicating the ability to identify positive instances, while a recall of 12.59% suggests challenges in capturing a substantial proportion of actual positive cases. The F1 score of 20.54% reflects a trade-off between precision and recall, and the confusion matrix illustrates a notable number of false negatives, emphasizing the model's struggle in correctly identifying positive instances.

GAUSSIAN NAÏVE BAYES:

For this, we initialize the GaussianNB() classifier from the scikit learn and train it.
This model achieves a high accuracy of 97.64%, but its precision, recall, and F1 score are all 0.0, indicating a failure to correctly identify any instances of the positive class, resulting in a confusion matrix dominated by false negatives. ¶

DECISION TREE CLASSIFIER:

DecisionTreeClassifier() is initialized for this model. Then we define the parameter grid for hyperparameter tuning. Fitting 5 folds for each of 72 candidates, totaling 360 fits. Best Hyperparameters: {'criterion': 'gini', 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5}
The Decision Tree classifier with Cross Validation exhibits outstanding performance with an accuracy of 99.51%, demonstrating a robust ability to accurately identify positive instances (precision of 90.60%) while effectively capturing a substantial proportion of actual positives (recall of 88.47%), resulting in a balanced F1 score of 89.52%. Better than Decision Tree classifier without Cross Validation.

Final comparison of all models:

Among the models evaluated, the Random Forest classifier stands out as the most effective in accurately identifying fraud orders, exhibiting the highest precision (95.77%) and a well-balanced F1 score (90.23%). This model's superior performance suggests that it strikes a strong balance between minimizing false positives and capturing a substantial proportion of actual positive cases.
In a business context, the Random Forest classifier's accuracy and precision are crucial as accurate detection ensures that companies can avoid the fulfillment of orders without legitimate payment, thereby protecting their revenue streams and maintaining profitability. Beyond financial considerations, the efficient identification of fraudulent activities contributes to streamlined operations, reducing the risk of stockouts, overstocking, and other supply chain inefficiencies. Moreover, businesses benefit from the preservation of customer trust and reputation, as legitimate customers feel secure in their transactions, fostering long-term relationships. By optimizing resources and focusing efforts on genuine transactions, businesses can mitigate the legal and compliance risks associated with fraudulent activities. Overall, the proactive identification of fraudulent orders not only

shields businesses from immediate financial harm but also enhances operational efficiency, customer relations, and long-term sustainability.