Capstone Report

Fauzan Pasaribu
Project Title: **Predicting Shipment Outcomes and Inferring Causes of Potential Disruptions**

Problem Statement

Supply chain disruption is a pervasive issue affecting numerous industries today. According to the Interos Annual Global Supply Chain Report, organizations across industries incur a staggering $184 million in lost revenue each year due to supply chain disruptions. This significant financial impact underscores the urgency of addressing this problem.

The nature of supply chain disruptions introduces a high level of uncertainty. These disruptions can occur unexpectedly, anywhere and at any time, impeding the timely delivery or even the arrival of shipments. Compounding the challenge is the fact that these disruptions are predominantly caused by external factors, over which companies have little to no control of.

The lack of control over external causes and the inherent uncertainty of supply chain disruptions necessitate a proactive approach to mitigate their adverse effects. Therefore, the primary objective of this milestone project is to develop a solution that can predict shipment outcomes and identify the potential causes of supply chain or shipment delivery failure. By achieving this, the project aims to reduce the uncertainties plaguing the supply chain and prevent shipment failures from occurring.

The project comprises two key components: 1) predicting shipment outcomes and 2) providing contextual information about potential shipment failures. By accurately predicting the outcomes of shipments, organizations can anticipate and prepare for potential disruptions.

The value-add of this project lies in how it the machine learning model act as an early warning system that empowers companies to identify potential failures in advance. This not only reduces uncertainties in the supply chain but it enables organizations to respond swiftly and effectively, thereby preventing shipment failures and their associated negative consequences.

Data Science in the Supply Chain Management Field

The very nature of the supply chain as we know is synonymous with uncertainty, unpredictability, and volatility. The variability that persists in this field and its impact on many industries makes it an ideal problem to apply data science techniques. Utilizing data science approaches, it is now possible to predict and control risks by analyzing both historical and current data. Machine learning algorithms can predict demand patterns, spot possible interruptions, and suggest proactive mitigation techniques, giving businesses the ability to deal with unforeseen circumstances. Therefore, predictive models and big data analyses are widely implemented and researched in the field. An example of that is a survey paper, "Predictive big data analytics for supply chain demand forecasting" published by Seyedan and Mafakheri. In addition, the supply chain has many moving parts. Most of the time it involves a very complex network structure including suppliers, manufacturers, distributors, retailers, and customers. The connections and dependencies that exist within these networks add to the supply chain's operational complexity. Data science approaches like network analysis and optimization algorithms becomes very relevant to improve coordination, streamline processes, and increase efficiency in the supply chain. A major example of a widely applied data science method on a supply chain network is route optimization which simply determins the most cost-efficient route in a very complex supply chain network.

Dataset

The dataset used in this project is acquired from Kaggle. This dataset was used by DataCo Global, a data analytics consulting company. It contains information about the supply chain of an unidentified company. The data entails information of order items that are delivered from the "customer" (the entity who made purchase of the delivery - the origin of the shipment) to the "buyer" (the entity who made the order from the customer - the destination of the shipment). The information on sales transactions, order quantities, and other order-specific information comes from internal data sources including enterprise resource planning (ERP) systems and customer relationship management (CRM) systems.

Cleaning and Data Preprocessing Summary

The initial dataset for the DataCo Smart Supply Chain project consisted of 180,519 rows and 53 columns. The EDA and data cleaning process refined the dataset to only 37 columns. The cleaning process involved the removal of redundant columns, such as those containing similar information to other features, as well as numerous ID columns that did not contribute substantial insights to the analysis. Additionally, columns containing censored information, including customer names and passwords were dropped. Only a total of 3 rows were excluded due to 3 existing null values in a column.

Approximately 34% of the data points were identified as outliers. However, after conducting thorough exploratory data analysis (EDA), it was determined that these outliers were not anomalies but rather inherent characteristics of the supply chain's unpredictable nature. Consequently, they were retained in the dataset to maintain its integrity.

Categorical feature variables were converted into numerical representations using one-hot encoding, with the first feature in each encoded set dropped to prevent multicollinearity. The target variable underwent two different approaches based on the classification task. For binary classification, the four-class target variable was transformed into two classes, where ‘late delivery’ and ‘shipping canceled’ were mapped to 0, and ‘advanced shipping’ and ‘shipping on time’ were mapped to 1. For multiclass classification, the target variable was label encoded with values assigned to each class (0 for advanced shipping, 1 for late delivery, 2 for canceled shipping, and 3 for shipping on time). Prior to modeling, features that exhibited multicollinearity were dropped. These preprocessing steps resulted in a final dataset with 42 features ready for subsequent modeling and analysis tasks.

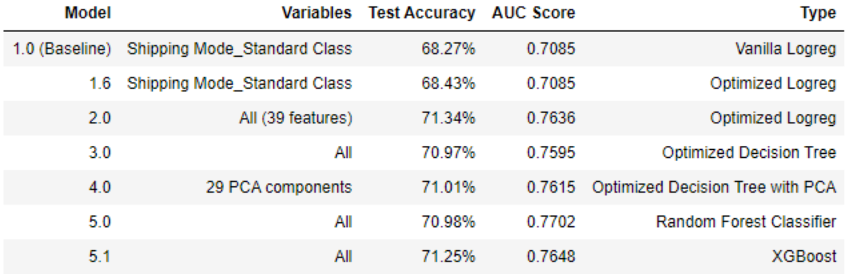
Insights, Modeling, and Results

Initially, a baseline model was created using a binary classification vanilla logistic regression. Subsequently, various models were experimented with binary, then moved forward with multiclass classification tasks. The model evaluation process focuses not only on accuracy but also on metrics such as AUC score and F-1 score. This emphasis was driven by the objective of prioritizing the prediction of successful deliveries with high confidence, considering the significant costs associated with misclassifying failed deliveries as successful ones.

The evaluation framework recognized that the cost of false positive predictions, i.e., classifying order items as successfully delivered when they will actually fail, outweighed the cost of false negative predictions. Misclassifying failed deliveries could result in lost potential revenue and diminished customer satisfaction. In contrast, misclassifying successful deliveries resulted in the additional effort and resources invested in ensuring their success. Therefore, AUC scores and F-1 scores played a crucial role in assessing the models' performance.

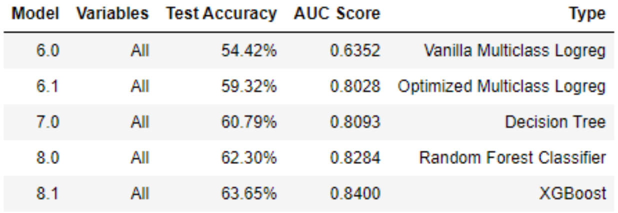
For both binary and multiclass classification, a range of machine learning models was employed, including logistic regression, decision trees, principal component analysis (PCA), neural networks, and ensemble learning methods such as Random Forest and XGBoost. Here are the model results. **Model 5.0, 8.0, and 8.1** are chosen to be deployed in a simulation due to their great accuracies and AUC scores.

BINARY CLASSIFICATION MODELS



Model	Variables	Test Accuracy	AUC Score	Type
1.0 (Baseline)	Shipping Mode_Standard Class	68.27%	0.7085	Vanilla Logreg
1.6	Shipping Mode_Standard Class	68.43%	0.7085	Optimized Logreg
2.0	All (39 features)	71.34%	0.7636	Optimized Logreg
3.0	All	70.97%	0.7595	Optimized Decision Tree
4.0	29 PCA components	71.01%	0.7615	Optimized Decision Tree with PCA
5.0	All	70.98%	0.7702	Random Forest Classifier
5.1	All	71.25%	0.7648	XGBoost

MULTICLASS CLASSIFICATION MODELS



Model	Variables	Test Accuracy	AUC Score	Type
6.0	All	54.42%	0.6352	Vanilla Multiclass Logreg
6.1	All	59.32%	0.8028	Optimized Multiclass Logreg
7.0	All	60.79%	0.8093	Decision Tree
8.0	All	62.30%	0.8284	Random Forest Classifier
8.1	All	63.65%	0.8400	XGBoost

Findings and conclusions

Despite the inherent challenges in predicting supply chain outcomes, including the presence of uncertainties and a significant number of outliers in the data, the predictive models achieved better accuracy than anticipated. However, in terms of the prescriptive aspect of the project, the goals were not fully realized. While the models were able to provide which features are most contributory to shipment failures, the project fell short of the desired comprehensive inference regarding the precise causes behind specific shipment failures. This limitation indicates opportunities for further exploration and refinement in future iterations of the project.

The practical value of the project lies in its ability to predict shipment outcomes and provide very basic insight into potential delivery failures. However, moving forward, potential next steps and future directions could involve incorporating more granular data, such as historical weather patterns, transportation network information, and supplier performance metrics. This would allow more predictive features that can give more context to the supply chain ecosystem. Moreover, integrating real-time data feeds and developing a dynamic predictive model that adapts to changing supply chain dynamics would be valuable for proactive decision-making.