Predicting Late Delivery Risk in Supply Chain Management

Kumbam Pavan Kalyan
24250901

ST606 Project
DATA SCIENCE AND ANALYTICS

Maynooth University
National University of Ireland Maynooth

Department of Statistics and Data Science
Maynooth University, Co. Kildare, Ireland.

A thesis submitted in partial fulfilment of the requirements for MSc Data Science & Analytics.

Supervisor(s): Dr. Rafael De Andrade Moral
Date: August 7, 2025

Acknowledgements

I want to express my sincere thanks to Dr. Rafael De Andrade Moral for his priceless guidance and support during this project. His suggested approach to the design was instrumental in helping me manage the complexities of this research. Dr.Rafael's expertise and insightful feedback greatly enhanced my understanding of the subject. His assistance in writing this report was particularly helpful, providing clarity and structure. I am deeply grateful for his help and time throughout this thesis, which has significantly enriched my learning experience and the quality of this thesis.

Contents

Abstract

Today's economy depends on efficient supply chains and delivering goods and services early plays a key role in business and customer satisfaction. As the market becomes more competitive, organizations are expected to run their logistics smoothly and efficiently even while providing excellent service. Failing to make deliveries on time leads to direct financial issues and also damage relationships with customers, the company's reputation and various steps in the supply chain. Therefore, organizations wanting to stay ahead and run efficiently must be able to forecast and control delivery-related risks. In this dissertation, a thorough study is made of how machine learning techniques can forecast late delivery problems in supply chain activities. The work has filled a major gap in the literature by creating and assessing models that can find high-risk shipments early so that companies can take action to avoid problems. Using analytical tools, the study changes raw operation data into useful insights that can greatly help with supply chain decision-making.

More than 180,000 supply chain orders from real situations were used as a strong data basis for creating and testing the models. Since this data covers different operations, groups of customers, product lines and locations, the models developed can represent how supply chains work nowadays. There is detailed information about orders, shipping, customers, products and outcomes in the dataset which forms a sturdy base for modeling predictions. Data science is used in this study by first thoroughly preparing and cleaning the data. In the beginning, data quality inspections were made, finding and tackling problems with missing aspects, unusual observations, various data arrangements and duplicate records. With high-quality and accurate data prepared, the following analysis could rely on trustworthy results. The analysis of cleaned data spotted patterns that shed light on problems and areas that can be improved in operations.

Running the analysis meant merging knowledge of the industry with statistics to form reliable predictors of whether deliveries would succeed. It included the creation of new functions that record critical details of a supply chain, for instance, delivery delays, accounts for profitability and monitor trends over time. The delay in shipping is an especially helpful way to measure inefficiency since it is easy to calculate the difference between what was planned and what occurred. In the same way, order profitability measures gave a picture of how resources were allocated and what impact that had on the timeliness of deliveries.

Chapter 1

Introduction

Over the past few years, one of the tools that have been found to have an immense effect on the way supply chains can be transformed is known as Big Data Analytics (BDA). Instead of being restricted to using a more basic form of statistical model, like regression and time series, or simply forecasting, organizations are utilizing more advanced data driven methods in order to predict the pattern of demand, customer behavior and general inefficiency.

With the big surge in accessible data and advances in computer power, supply chains have been changing to be proactive rather than reactive. Such a shift enables businesses to make more intelligent, reliable decisions, and much quicker decisions at multiple supply network nodes. According to what has been mentioned by Addo-Tenkorang and Helo (2016)[8], big data is of paramount importance in terms of improving responsiveness and adaptability through real-time visibility and insights that can be acted upon. Besides, predictive abilities to logistics and plan have been enhanced tremendously through the use of machine learning models, such as Random Forest (Breiman, 2001) [2] and XGBoost (Chen & Guestrin, 2016) [3]. These technologies have the ability to process very significant amounts of data in both structured and unstructured forms and to identify patterns in the data and to arrive at valid predictions that can be used to prevent disruptions and streamline inventory distribution.

Waller and Fawcett (2013) [11] state that not only does data science increase the value of traditional tools of supply chains but it is a revolutionary change that transforms the architecture and decision-making of supply chains. Inclusion of BDA in SCM(Supply Chain Management) empowers businesses to achieve cost reduction in business operations in addition to boosting the quality and satisfaction on the side of customers. With a smarter use of data, businesses can better align their strategies towards the market needs and be more confident in making decisions due to uncertainties. The data science provides an efficient way of solving the actual business problems by transforming difficult information into useful insights. The first step, as shown in Figure 1.1, is the process of identification of the core business issue and the objectives being clear. It is then followed by data gathering, cleaning, and viewing where relevant data are gathered, clean to eliminate repetitions and inconsistency and explore to learn some basic patterns and trend.

Figure 1.1: Life Cycle Data Analytics [12]

The more advanced phases are selection of important features, creation of predictive models, and visualizing the results in forms that can help in decision making. This systematic process allows its organizations to boost their efficiency, create cost savings, a desirable customer experience, and introduce proof-based decisions with more rationality than an intuitive one.

1.1 Problem Statement

The dataset to be analyzed represents the delivery records of the supply chain activity of a company known as Data Co[9].The initial analysis of the data set shows that a significant part of the deliveries was completed after a deductible time had been set. Such delays not only hindered customer satisfaction but as well created a threat to the efficiency of the operations of the company and their profit margins. The case of late delivery became one of the serious bottlenecks in the supply chain process. To investigate and address these concerns, our study aimed to:

1. **Create predictive models** to ascertain on whether individual products would either come on time or delayed through historical records of data.

2. **The risk factors** due to which the delivery is likely to fail analyse and rank influential risk factors such as transportation method, product category, order volume, and lead time.

3. **Evaluate the stages** in the delivery process to ascertain where and when bottlenecks will most likely occur.

4. **Track the delivery compliance** rate on a transaction by transaction basis, and identify the trend or trend with a view of causing system inefficiencies in delivering orders.

1.2 Objectives of the study

The primary objective of this study is to analyze the factors influencing delays in the supply chain and to develop predictive models that can forecast late deliveries. Specifically, the study aims to:

1. **Delivery timelines should be analyzed** to give an idea of how and when the products were delivered.
2. **The levels of delivery agreements** should be evaluated by examining the deliveries which should have been done within the rightful time frame.
3. **Identify some of the most important** aspects that make the company take too long on its transactional or logistical process.
4. **The presence or absence** delayed delivery will fall into the category of the predictive model using a number of methods like Logistic Regression, Feedforward Neural Network,Random Forest, and Ensemble approaches (e.g., bagging and boosting).
5. **Use the necessary measures** like Confusion Matrix to evaluate the way these models perform.
6. **Maximize the accuracy of predictions** by tuning hyperparameters of a set of models. Using the most effective model solely on a comparative performance and interpretability basis.
7. **Offer recommendations concerning** business insights and can do steps to minimize delays in delivery and improve the efficiencies of the supply chain.

1.3 Scope

This paper presents results conducted on one of the publicly available datasets first posted on Kaggle that mimic the workings of a company by the name of Data Co[9]. Global. Company data is also anonymized to protect data and is not related to an existing organization. Since the given research is restricted by the issues included in the considered dataset, its features include the data related to orders, shipping behavior, product characteristics, payment, and customer demographics. The purpose will be to apply this organized information to come up with predictive models that categorize delivery performance and help to give delivery-related risk factors of late deliveries.

1.4 Data Source

The database behind the research is an imitation of the activities of company called Data Co.[9] Global, established to educate and analyze. It gives satisfactory records in regard to the supply chain activities planned by the company such as provisioning, production, sales, and commercial distribution of consumer products and publicly available dataset provided by Kaggle. The dataset comprises **180,519 records** and includes **53 features** across various domains. The data consists of both quantitative (numerical) and qualitative (categorical) variables.

Sub ordinate categorization of quantitative variables is matched with:

- **Discrete:** Counts and values of finitary sets (examples: the number of visits, the quantity of goods ordered).
- **Constant:** Standing values that can be measured and are therefore numeric (i.e. unit price, weight, profit).

Qualitative variables are however categorisable into:

- Nominal: Categories that need not be ordered in a natural or natural-like way (e.g. customer segment, region, shipping mode).

- Ordinal: Categories that can be classified in terms of rank of importance (e.g. the importance of products, importance of the level of customer satisfaction).

The data spans a period of over **three years**, collected daily from **January 2015 to February 2018**. It captures multiple aspects of customer transactions, order fulfillment, and logistics operations. A detailed taxonomy of the dataset which categorizes the variables to facilitate clearer understanding and modeling.

1.5 Methodology

The methodology involved collecting and preprocessing supply chain data by handling missing values, encoding categorical variables, and engineering features like processing time. Exploratory Data Analysis was used to uncover key patterns influencing late deliveries. Feature selection was performed using Recursive Feature Elimination. Four machine learning models Logistic Regression, Random Forest, XGBoost, and a Feedforward Neural Network (FNN) were trained and evaluated using a 70:30 train-test split to predict late delivery risk effectively..

Chapter 2

Literature Review

2.1 Evolution of Predictive Analytics in Supply Chain Management

Predictive analytics has in the past decade become the pinnacle of efficient and effective supply chain optimization and streamlining. With businesses operating in the ever more globalized environment with more complex distribution channels the focus changed to preventing occurrence of the issues rather than responding to them after they already happened. Predictive analytics enables disruptions to be spotted earlier, and remedial strategies put in place before the situation can affect the services.

Angappa Gunasekaran(2017)[1]. emphasized the importance of data-driven approaches in logistics which ensure its effectiveness of operations particularly in demand forecasting and stock replacement. Predictive models today form part of the core activities like routing, inventory, and shipment scheduling which make them crucial in the establishment of highly resilient and responsive supply chains.

2.2 Delivery Risk and Performance Disruptions

Supply chain performance is very much measured by delivery on time. It can, however, be hindered by numerous factors that include demand volatility, warehouse mismanagedness, and various unpredictable events such as natural or geopolitical causes. According to Simchi-Levi et al (2015)[4], these risks can be categorized into controllable (e.g., operational inefficiencies) and uncontrollable (e.g., external shocks), none of them are uncommon in nature, however, recognition of risk early with the help of machine learning should be a promising mitigation path to take. Although traditional models (e.g., regression and time-series analysis) have been used in forecasting the delivery times, it can only capture simple and linear relationships and has restricted the usefulness of such models, in terms of accuracy in dynamic supply chain networks.

2.3 Machine Learning for Delay Risk Prediction

Machine learning (ML) presents a stimulating alternative to conventional predictive instruments through the ability to adapt to information that is heterogeneous and learning the complex patterns. It has been demonstrated that the techniques of decision trees, support vector machines (SVMs), and ensemble methods (including Random Forest and XGBoost) have the best predictive performance on late delivery classification.
In the experiment of Fabian Steinberg(2023)[5], Random Forest algorithms were used to identify delay predictors in e-commerce logistics and it was found that such items as the order timing and mode of payment had close relationships with delays. In a similar fashion, Baryannis et al (2019)[6]. established that ensemble-based models performed superiorly in the formulation of global supply chain interruption forecasting compared to individual classifiers and hence have proven to be robust.

2.4 Barriers to Real-World Implementation

Despite their promise, ML models face considerable hurdles during deployment within supply chain systems:

- **Class imbalance:** Late deliveries often represent a small fraction of the data, causing models to skew toward majority (on-time) outcomes.
- **Data quality:** Missing entries, inconsistent data types, and noisy records necessitate thorough preprocessing.
- **Model transparency:** While complex models may boost accuracy, their "black-box" nature hinders acceptance among stakeholders who prioritize clarity in decision-making.

Although the utilization of predictive analytics in supply chain management experience an increasing trend, it is reasonable to observe that there are limited studies dedicated to feature engineering approaches directed to the peculiarities of supply chain data. The relevant literature focuses more on model selection at the expense of quality and relevance of input features that is of paramount consideration to model accuracy and interpretability. Also, there are few comparative studies of several classification algorithms on a real and large-scale data. The majority of the research evaluates models on either a solitary setting or simplified datasets, and are not representative of the dynamics of operation in a commercial logistics situation. To fill these gaps, this research does the following: it uses a variety of machine learning models on a large-scale supply chain dataset. The three most important features of the research are the set of proper feature selection methods, the performance of the models compared, and the potential business applicability of the results to decision-makers.

Chapter 3

Background

The use of predictive analytics has been identified as a critical asset in the contemporary supply chain management system where companies can now foresee breaches in practice, streamline delivery times and minimize the risks of such an approach. With the global supply chains becoming more complex, organizations are changing their dynamic style into proactive and data-based dynamics. Predictive analytics can assist a business in making wiser decisions regarding topics such as demand forecasting, inventory management, and even logistic planning by using past information and statistical modeling.

The desire to improve efficiency, lower costs and satisfy the customers has influenced the companies to move towards using predictive models that could deal with challenging volumes of data. Such models use the latest machine learning and statistical tools that perform a pattern analysis, predict possible delays, and provide decision-makers with valuable insight to act on.

3.1 Related Work

Many other studies have been used to examine an extensive variety of machine learning algorithms to enhance the performance of the supply chain. Linear regression models are the frequently used types of regression models that are used to predict delivery times with respect to the characteristic of an order like the volume or distance or supplier behavior. Nonetheless, they are only effective in capturing relationships that are not linear and non-dimensional. Decision trees are more interpretable by nature and can also be used when non-linear data is involved in modeling the decision boundaries. Nevertheless, they suffer overfitting problems, especially when noise is the issue. Random forest and gradient boosting are on the rise to overcome these drawbacks of ensemble techniques. The methods are combinations of weak learners so as to increase the accuracy of the predictions and minimize variation.

Class imbalance is one of the major issues in prediction of delivery risk as the number of deliveries which are delivered on time is much more significant than the number of delayed deliveries. The possible methods of training datasets balancing and the techniques of expanding the number of minority classes determination are Synthetic Minority Over-sampling Technique (SMOTE) and cost-sensitive learning techniques. Feature selection and engineering is the other important issue of predictive modeling. The design of the feature contributes to its dimensionality reduction, augmenting interpretability, and making the model better. These methods could be Recursive Feature Elimination (RFE), Principal Component Analysis (PCA) and domain specific transformations like lag features to deal with delay trends or profitability measures.

3.2 Project Context

1. **Correlation of Past Literature:** This project is based on well developed research in the field of supply chain predictive models. Even though past research already examined potential solution to algorithmic delivery forecasting problems, the current paper is an extension of these results, as the current analysis is being presented on a multidimensional and real-world dataset that consists of numerous various variables: the information about orders life cycle, customers,

the types of goods, deliveries. The goal is to improve the predictive performance with carefully selected feature growth but in a manner that is easy to use and map against the processes that may be worked out realistically.

2. **Research on feature Engineering:** The second pillar of this study is the creation of more focused characteristics that extract hidden trends in the data. Derived attributes such as the shipping delay- an overlap of the actual and scheduled delivery dates of a product may be tried in order to offer insightful results on the inefficiencies in logistics. Another functionality, Profit Ratio, indicates financial gain constituted per transaction, in comparison with revenue, and thus represent an analytics-plank of what links operational and financial analytics. Besides playing the role of enhancing the precision of the model, these engineered features also give interpretive advantage to the use of strategic decisions.

3. **The Comparisons of the Machine Learning Models:** The paper carries out a comparison of various classification algorithms such as Random Forest, XGBoost, Logistic Regression and Feedforward Neural Networks is done with rigorous performance metrics to include Accuracy, Precision, Recall, F1-score, and ROC-AUC so as to conduct Model evaluation. Such a systematic approach to benchmarking allows an articulation of the optimal model to predict delivery delays on a supply chain where data is of high dimension and imbalanced.

4. **Interpretability and Business Impact:** Besides performance, the model is assumed to be highly interpretable, in particular, to those stakeholders that have to address operational implementation. They can have a hauntingly black-box behavior even though complex models might have better accuracy. To certify that the project can be adopted in the real-world business environments, this project underlines transparency through interpretable structures and explainable tools. The goal is to empower decision-makers with actionable insights, foster trust in the outputs, and facilitate seamless integration of predictive capabilities into existing supply chain processes thereby advancing both tactical operations and strategic objectives.

Chapter 4

Description of Work Undertaken

4.1 Data Overview and Preprocessing

This analysis is carried out on a dataset that has about 180,000 rows of transactions and individual rows of the dataset represent individual customer order. Such orders are extremely detailed with over 53 first features per entry. These features cut across various areas such as customer information, product details, delivery options, payment channels, and timestamps to provide all insights into the operations of the business.

This abundance of data enables analysts and managers to have the insight on micro level consumer behavior, trends in transactions, and efficiency of the operations. These are some of the most important properties that involve actual vs. planned delivery dates, sales per customer, the profitability of orders, delivery status, and the type of payment directly affecting the business performance and customer satisfaction. Furthermore, such categorical domains as city, country, and product category assist in establishing the background of every transaction and enable them to be segmented on the corresponding market and examined on the demographic level. In order to come up with good quality data to be analysed, rigorous data preprocessing was conducted prior to analysis. Missing values were the first to be taken care of. Mostly, all the numerical entries that were not present were interpolated with relevant statistical methods like mean, median or mode imputation as relevant depending on the distribution of the variable coupled with the context. As an example, where a delivery date was not and shipping method along with order timestamp was available, an expected average delivery disparity was utilized to complete the number. Nevertheless, in case of absences of important values (on essential characteristics, e.g. product ID or customer ID), such records were removed to be out of a dataset in order to prevent biased results.

After taking care of missing values, date variables like order date and shipping date were transformed into datetime objects. It was needed to be able to compute measures such as time-to-ship, temporal trends, and time-based grouping (i.e., by week, month, or day of the week). Further, categorical variables such as payment mode and shipping type were encoded using one-hot encoding, which creates binary columns for each category. This encoding ensures that the machine learning models treat the values as independent categories, without assuming any ordinal relationship. For instance, a column indicating payment method with values like "Credit Card", "Cash", and "Online Wallet" was converted into three separate binary columns. To enrich the dataset further, feature engineering was employed. New variables were derived such as:

1. **Order-to-shipment gap:** calculated as the number of days between the order date and shipping date.

2. **Day of week and hour of transaction:** extracted from timestamps to help identify peak business hours.

3. **Order time period classification:** assigning each transaction to "Morning", "Afternoon", "Evening", or "Night", to understand order behavior across the day.

These engineered features added valuable insights for modeling purposes and helped reveal business patterns such as late-night shopping surges or slow shipping on weekends. They also enhanced the dataset's predictive power by embedding temporal and behavioral dimensions.

4.2 Exploratory Data Analysis

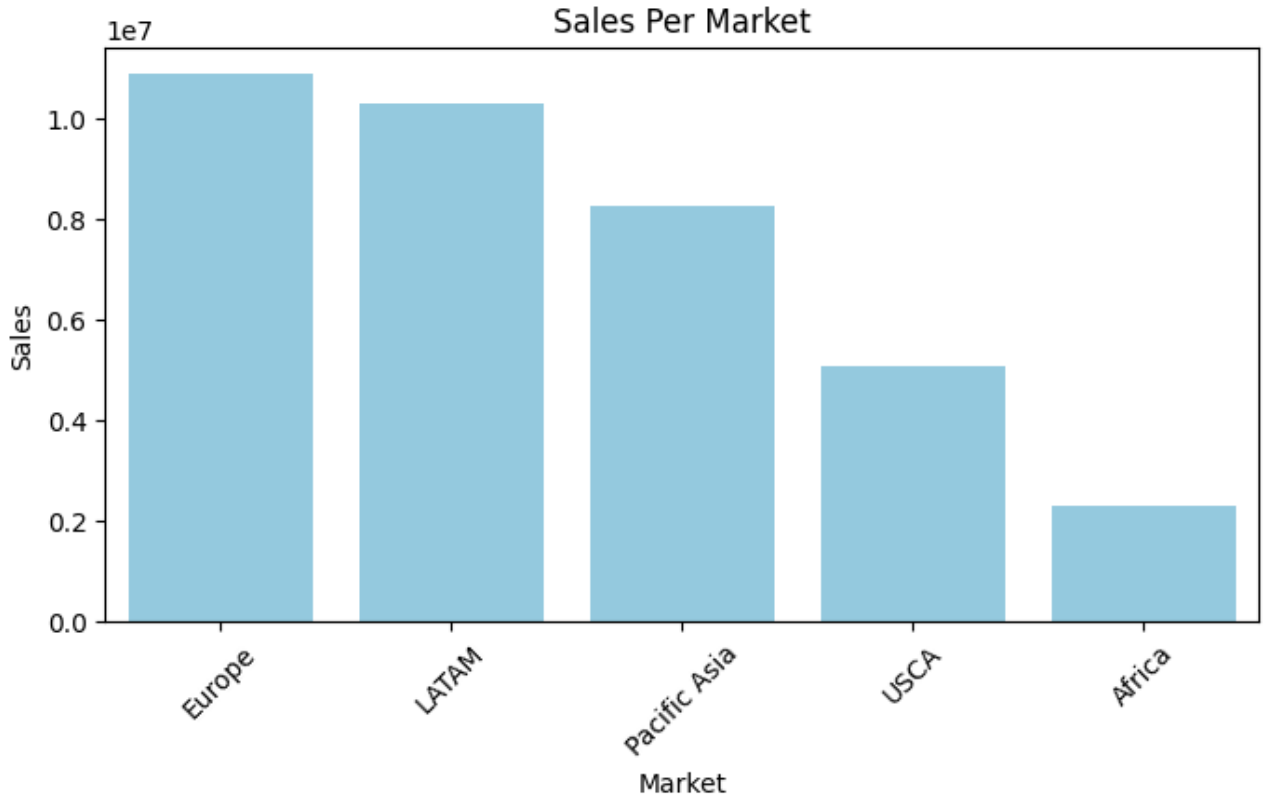

Figure 4.1: Sales per Market

As shown in the figure 4.1, Europe leads in sales, contributing the most, followed closely by LATAM. Pacific Asia also shows a significant share of sales but lags behind Europe and LATAM. USCA comes in lower, and Africa shows the smallest sales contribution among the markets. The clear gap between these regions indicates potential differences in market size, customer base, or delivery performance across these geographic areas.

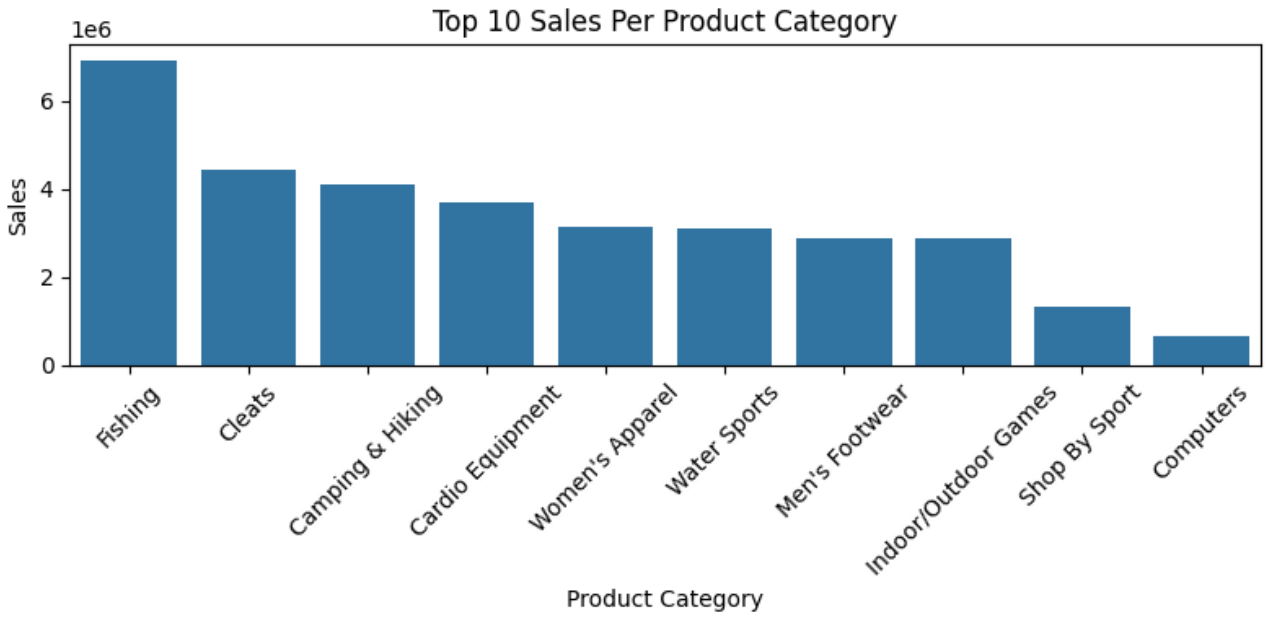

Figure 4.2: Sales Per Product Category

The bar chart in figure 4.2, highlights the top 10 product categories in terms of sales. Fishing leads significantly with over 6 million USD in sales, followed by Cleats and Camping & Hiking, both contributing around 4 million USD. Cardio Equipment, Women's Apparel, and Water Sports show slightly lower but still substantial sales figures. Categories like Men's Footwear, Indoor/Outdoor Games, and Shop By Sport also perform well, while Computers is at the bottom of the top 10 list.

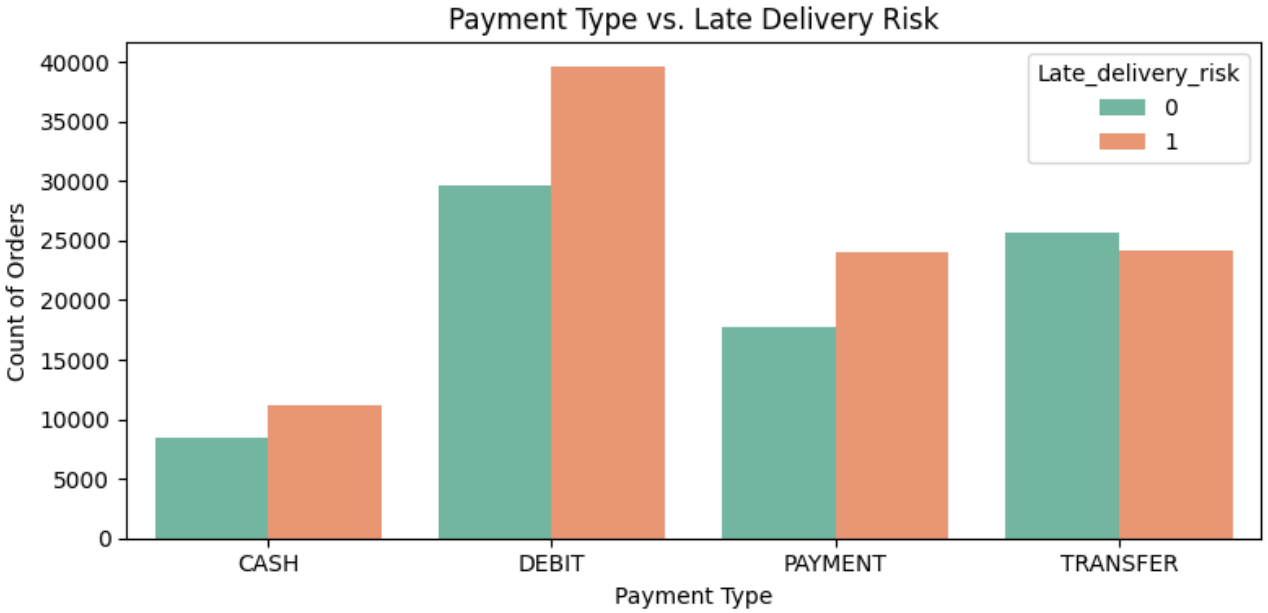

Figure 4.3: Payment Type Vs. Late Delivery Risk

The bar chart in figure 4.3, illustrates the relationship between different payment types and late delivery risk. The DEBIT payment type shows the highest count of late deliveries (in orange), significantly exceeding orders delivered on time. CASH payments have the lowest number of orders overall, with fewer late deliveries than other payment types. PAYMENT and TRANSFER types have a relatively balanced split between orders delivered on time (in green) and late deliveries. The chart suggests that certain payment types, especially DEBIT, are associated with a higher late delivery risk, indicating a potential operational inefficiency or delay related to these transactions.

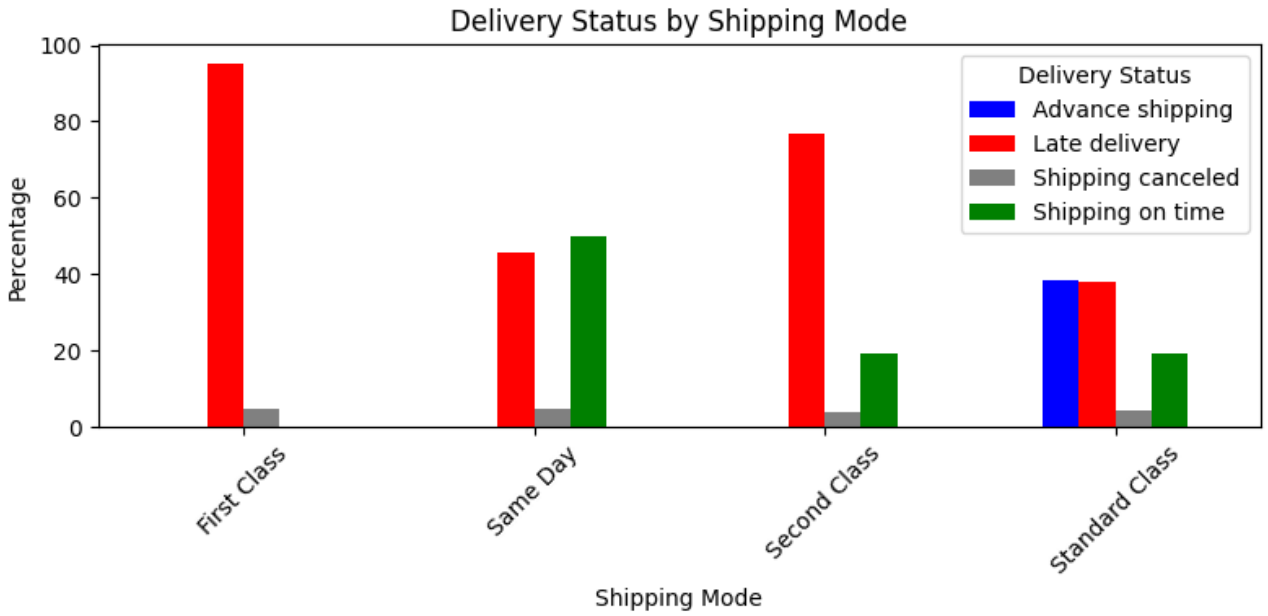

Figure 4.4: Delivery Status By Shipping Mode

The figure 4.4, shows that the First Class and Second Class modes have a high proportion of late deliveries (in red), with First Class reaching nearly 100% late deliveries. Same Day shipping has a more balanced distribution, with a significant portion of on-time deliveries (in green) and some late deliveries. Standard Class shows a mix of delivery statuses, with a notable percentage of advance shipping (in blue) and on-time deliveries. However, it also has a notable number of late deliveries. This chart emphasizes that First Class and Second Class are more prone to late deliveries, whereas Same Day and Standard Class offer a better balance of timely deliveries.

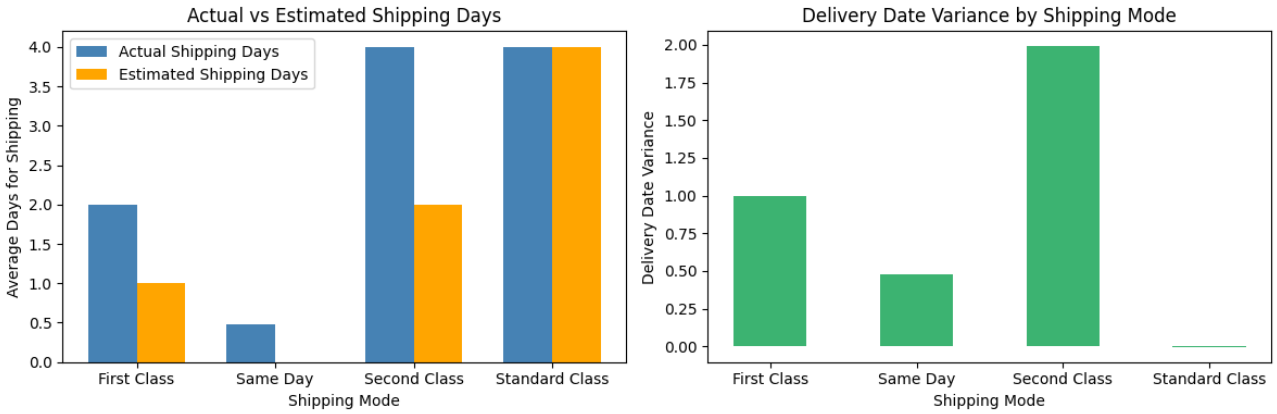

Figure 4.5: Shipping Mode Comparison: Actual vs Estimated Delivery Time and Variance

The first chart in figure 4.5, compares the average actual and estimated shipping days across different shipping modes. For First Class and Second Class, actual shipping times (in blue) significantly exceed estimated times (in orange), indicating a clear delay. Same Day shipping is the only mode where the actual shipping days closely align with or are even slightly lower than the estimated shipping time. Standard Class shows a similar pattern to First and Second Class, with actual shipping times exceeding the estimates, though not as drastically.

The second chart in figure 4.5, illustrates the delivery date variance across shipping modes, showing how much the actual delivery times fluctuate. Second Class shows the highest variance, indicating more inconsistency in delivery times, while Standard Class has the lowest variance, suggesting more predictable shipping times. First Class and Same Day show moderate variance, with Same Day having slightly lower fluctuation compared to First Class. This reinforces the idea that Second Class shipping is more prone to delays and inconsistencies, while Same Day provides the most reliable performance.

4.3 Feature Selection

The most significant factors for late delivery risk prediction were found using Recursive Feature Elimination and Logistic Regression. The selection of methods made it possible to measure the impact of each feature and pay more attention to those that matter most.

4.4 Recursive Feature Elimination (RFE)

RFE stands for Recursive Feature Elimination and is helpful in choosing the best features for a model. We do this by step-by-step removing attributes and basing the model on those which are kept. This method finds out which attributes are the most useful in predicting the target value by looking at the model's accuracy. The main benefit of RFE is that it handles a great number of features and singles out the most important ones. In this discussion, RFE was used on the dataset to calculate how much each feature helped or hindered the prediction of late delivery risk. At first, all the features were included and then, the features with the least importance were eliminated until the right ones remained. As a result, only the relevant features were left which diminished the data's complexity and elevated the performance of the model.

4.5 Coefficient Analysis from Logistic Regression

Logistic Regression is widely used for solving binary classification problems, such as distinguishing between late and on-time deliveries. In this study, the model was employed to estimate the likelihood of a shipment being delayed. Rather than using overly technical or abstract terminology, the model directly links the input features (like shipping duration or order processing time) with the final outcome whether an order was delayed or not. The strength and direction of these relationships are captured through model coefficients. A positive coefficient suggests that the feature increases the chance of a late delivery, while a negative coefficient indicates a higher probability of on-time delivery. This allows us to see which factors are driving delays and which are helping avoid them. After training the model, the coefficients were examined to understand which variables had the greatest influence. Features with larger absolute values of coefficients were considered more impactful in predicting outcomes. This analysis not only improved model transparency but also provided practical insights into which aspects of the supply chain require closer monitoring or optimization.

4.6 Prioritized Features

Using both Recursive Feature Elimination (RFE) and coefficient analysis, several key variables were identified as the most influential in predicting the risk of delayed deliveries. These variables are critical for understanding bottlenecks in the shipping process and improving overall logistics performance. The most impactful features include:

1. **Order to Shipment Time**
   This feature captures the time gap between when an order is placed and when it is handed over for shipment. It provides insight into the internal processing efficiency of the system. A longer gap often signals operational delays that could result in late deliveries. Our analysis showed that reducing this interval significantly improves the chances of on-time shipments, making it a crucial metric for fulfillment optimization.

2. **Days for Shipment (Scheduled)**
   This variable indicates the promised delivery schedule provided to the customer. It is essential for expectation management. Discrepancies between scheduled and actual shipping days often lead to customer dissatisfaction. Our findings reveal that maintaining accurate and realistic delivery schedules plays a vital role in minimizing the risk of late shipments.

3. **Shipping Mode**
   This feature describes the transport method chosen for delivery for example, air, ground, or sea. Each mode has different implications for delivery speed and cost. Air freight is fast but expensive,

while ground and sea transport are slower but more economical. The model consistently found shipping mode to be a reliable indicator of delivery timing. Selecting the appropriate mode based on urgency and distance is therefore crucial for timely deliveries.

4. **Payment Type:** This feature offers different payment possibilities such as TRANSFER, PAYMENT, DEBIT and CASH. Some payment methods can impact how quickly orders are taken care of and shipped. Cash payments usually take longer because they usually need extra checks, while transfers and payments with debit cards are faster. The importance of this feature is shown by its large coefficient.

4.7 Modelling

Models Used

Many machine learning models were put to use in this quest, as they each had their strengths and powers. The logistic regression, random forest and XGBoost models were chosen for use in this analysis. All types of data and relations can be adequately handled by choosing the proper model.

Logistic Regression
Binary classification of data is commonly done using Logistic Regression as a basic approach to statistics. It estimates the chances of a specific event like late delivery occurring or not such as on-time delivery. The model finds the connection between input features and the prospect of the target outcome. Because its coefficients can be understood, Logistic Regression is ideal for finding how each feature affects the outcome. When the coefficient of a feature is positive, the chance of the outcome increases and when it's negative, the chance decreases. It is necessary to have a clear understanding of the factors that lead to late delivery and use data for making choices to create an effective model.

Random Forest
Many decision trees are used and Random Forest combines their findings to improve both accuracy and stability. On training, each tree gets a portion of the data randomly and the outcome is found by combining the predictions made by each tree. This practice lowers the chance of overfitting and makes the model adapt to information outside its training data. It is a good approach for handling lots of variables in a dataset since it finds any difficult relationships between them. In addition, it displays a list of scores showing which characteristics play the biggest role in deciding if there will be a late delivery.

XGBoost
XGBoost was incorporated as an advanced ensemble method known for its predictive power and efficiency in handling large-scale data. It operates by combining multiple weak learners typically shallow decision trees into a strong predictive model using gradient boosting. In this project, XGBoost also facilitated the identification of critical predictive features via importance scores, making it a valuable tool not only for performance but also for diagnostic insights.

Training and Validation
The dataset was systematically divided into training and test subsets to assess the generalization ability of each model. This separation ensured that the test data remained unseen during training, thereby providing a reliable estimate of the model's performance on new data. Standard data splitting practices were followed, with stratified sampling applied when needed to maintain class distribution across subsets.

Cross-validation for Hyperparameter Tuning
To improve model reliability and prevent overfitting, cross-validation techniques were implemented-specifically k-fold cross-validation with stratified sampling. Each model underwent a hyperparameter tuning process using grid search across multiple validation folds. The results from each fold were aggregated to fine-tune the model's settings, ultimately leading to more stable and generalized performance across unseen data. All procedures were conducted, ensuring reproducibility and control over

the evaluation framework.

Evaluation Metrics
Accuracy, precision, recall and F1-score were applied to judge how well the data models worked. Metrics reveal different aspects of how accurately the model can predict and all of them provide a complete assessment of how it does its job.

Accuracy
To know how accurate the model is, look at the real positive and negative outcomes for all predictions made by the model. Sometimes the outcomes are not accurate because the number of examples in each class is not the same. If late deliveries rarely occur compared to on-time deliveries, accuracy by itself cannot show us the whole picture of how a model performs.

Precision
It describes what percentage of the positive predictions made by the model is valid. It is even more crucial when there is a high risk associated with getting a false positive. The field of late delivery prediction shows how close the predictions are to the actual late deliveries. When the model is very precise, it is accurately able to predict delays in delivery.

Recall
The term recall is used for sensitivity which measures how many true positive predictions are made among all actual positive cases. It matters a lot in situations where having a false negative can be very costly. The recall of the model in late delivery prediction means it correctly identified how many of the actual late deliveries should have been predicted. When the recall is high, the model can catch many late deliveries, ensuring that one does not miss any missed deadlines.

F1-Score
The F1-score is calculated by taking the average of precision and recall, making it a fair indicator for the two measurements. It helps a lot when classes are distributed unevenly, as it considers both false positives and false negatives. F1-score is helpful in late delivery prediction since it allows comparing the model's performance by balancing both false positive situations and false negative ones.

Focus on Recall
For this type of business case, highlighting predictions of late deliveries was crucial. Being late with deliveries may lead to reduced customer satisfaction and damage the company's reputation. That's why it matters a lot not to report a shipment delayed unless it truly is late (so-called false negatives). Focusing on recall, the models are designed to make sure as many late shipments are spotted which makes it simpler to handle and address any delays. All in all, the analysis used Logistic Regression, Random Forest and XGBoost models to foresee the risk of late delivery. The dataset was split into training and test sets, and cross-validation was used for hyperparameter tuning to ensure the models' robustness and generalizability. Accuracy, precision, recall and F1-score were used to measure the performance of the models, with emphasis on recall so as to limit the number of false negatives. Using these strong machine learning tools and evaluation factors, the study was designed to predict late delivery risk accurately and securely to help decide and improve how parcels are delivered.

Implementation Challenges and Solutions
When making predictive models for late delivery risk, several problems in applying them were encountered. If these difficulties are not managed correctly, they may greatly damage how well and reliably the models function. Experts noticed that the biggest problems were using different amounts of data for different classes, classes missing data and some features doing the same job. All these issues were thoroughly reviewed and the right solutions were applied to maintain the integrity of the models.

- **Imbalanced Classes:**
  One of the greatest problems was that the target variable had a mild degree of unbalance. The advantage here is that the deliveries made on time were many, while the number of late deliveries was relatively low. This imbalance can lead to biased models that favour the majority class (on-

time deliveries), thereby reducing the model's ability to accurately predict the minority class (late deliveries).

Solution: Class Weights and Stratified Sampling
Methods used to solve the problem of unequal class numbers were class weights and stratified sampling.

Addressing Class Imbalance
In the original dataset, the number of on-time deliveries significantly outweighed the number of late deliveries. This imbalance can bias the model towards predicting the majority class, reducing its ability to correctly identify late shipments. To address this, two key strategies were implemented.

Class Weights Adjustment
During model training, class weights were applied to penalize misclassification of the minority class (late deliveries). By assigning a higher weight to late deliveries, the algorithm was encouraged to pay greater attention to those instances. This technique effectively increases the sensitivity of the model towards delayed shipments without needing to alter the data distribution artificially.

Stratified Sampling
To maintain a balanced class distribution during model evaluation, stratified sampling was used during the train-test split. This ensures that both training and validation datasets retained the same proportion of late vs. on-time deliveries as the original dataset. By doing so, performance metrics such as accuracy, precision, and recall reflect real-world expectations more reliably, preventing skewed evaluation due to unequal class ratios.

Implemented Solutions:

- class_weight='balanced' parameter used in models like Logistic Regression
- StratifiedShuffleSplit used in cross-validation and dataset partitioning

Feature Redundancy: The dataset contained several missing values across key variables like Product_Weight, Delivery_Status, and Customer_Location. Unaddressed, these gaps could distort feature distributions and reduce model performance. Therefore, the following steps were taken:

- Numerical columns with missing values (e.g., Product_Weight) were filled using mean imputation.
- Categorical columns (e.g., Customer_Location) were filled with the most frequent category or labeled as "Unknown" if suitable.
- Rows missing target labels (i.e., delivery status) were removed entirely, as they could not contribute meaningfully to supervised training.

After all cleaning and preprocessing steps including duplicate removal, filtering invalid records, and handling missing values the original dataset of approximately 180,000 observations was reduced to about 172,645 clean records that were used for training and testing.

Chapter 5

Analysis and Evaluation

During the analysis and evaluation step, the project for late delivery risk looked closely at the performance of distinct machine learning models. The idea was to find the most suitable model for predicting late deliveries, keeping in mind its accuracy, how understandable it is and important performance measures. Evaluation was made on Logistic Regression, Random Forest and XGBoost as part of the models. The extent of a model's accuracy and the value of the results for discovering reasons behind late deliveries were used to assess each approach.

5.1 Logistic Regression

Logistic Regression was adopted as a baseline model for binary classification, particularly effective in estimating the likelihood of an event such as whether a delivery would be late or on time. One of the primary advantages of this approach lies in its interpretability coefficients associated with each predictor offer clear insights into how individual variables influence the outcome. A positive coefficient implies that as the value of that feature increases, the likelihood of late delivery also increases.By analyzing these coefficients, it was possible to identify the most influential features contributing to delayed shipments. The features listed below had the highest coefficients, indicating a strong positive correlation with late delivery risk.

5.2 Random Forest and XGBoost

Random Forest and XGBoost are ensemble learning methods that aggregate multiple decision trees to generate more accurate and stable predictions. These models are particularly effective when working with large datasets that contain numerous variables, as they are capable of capturing complex, nonlinear relationships among features.Unlike Logistic Regression, however, these models do not provide easily interpretable coefficients for each individual predictor.While feature importance can still be assessed for example,using impurity-based scores or SHAP(SHapley Additive exPlanations) values the internal workings of tree ensembles are less transparent in comparison to the coefficient-based interpretations available in simpler models.Therefore,although these models are highly predictive, their explainability tends to be more abstract and requires additional techniques to interpret variable accurately.

5.3 Random Forest

Random Forest builds multiple decision trees using bootstrapped samples of the dataset and randomly selected features.The final prediction is made by aggregating (majority vote or average) predictions across all trees.This method helps in reducing overfitting and increases generalization. It also outputs feature importance scores, helping to identify the key drivers of late delivery risk.

5.4 XGBoost

XGBoost (Extreme Gradient Boosting) is an advanced gradient boosting technique known for speed and performance. It iteratively builds trees to correct previous errors, with enhancements like regularization to reduce overfitting, native support for missing values, and parallel training. It performs particularly well on structured data and is widely used in competitions and production systems for predictive tasks.

5.5 Performance Metrics

We evaluated model performance using Accuracy, Precision, Recall and F1-Score. Special emphasis was placed on Recall, as false negatives (missed late deliveries) could severely impact customer satisfaction and supply chain efficiency.

5.5.1 Logistic Regression

Metric	Value
Accuracy	0.974
Precision	0.956
Recall	0.999
F1-Score	0.977

5.5.2 Random Forest

Metric	Value
Accuracy	0.974
Precision	0.956
Recall	0.998
F1-Score	0.976

5.5.3 XGBoost

Metric	Value
Accuracy	0.692
Precision	0.824
Recall	0.559
F1-Score	0.666

5.6 Feature Importance

Across all models, the most important features influencing late delivery risk were:

- **Order_to_Shipment_Time:**Consistently ranked as the top predictor. It reflects how long it takes from order placement to shipment initiation.
- **Shipping Mode:** Premium options like "Same Day" show lower risk compared to "First Class".
- **Payment Method:** "Transfer" payments emerged as a significant risk factor.

The analysis and evaluation phase of the project looked closely at the Logistic Regression, Random Forest and XGBoost models. Among the significant predictors that were highlighted were fully processed shipping time, shipping time stated by the seller, shipping time scheduled by the seller, the shipping method and the payment type. Even though Random Forest and XGBoost performed better, they were easier to understand than Deep Neural Network. However, most of the features noticed in these models were similar to the ones important in Logistic Regression, showing that Order_to_Shipment_Time is most important for predicting late deliveries. Combining both models and using important metrics enabled the analysis to identify the reasons behind late shipments. Using this information allows businesses to improve how their shipments are arranged, avoid late deliveries and boost both customer satisfaction and how smoothly the company operates.

Model Performance
It was very important to evaluate the predictive models to check if they could accurately predict late delivery risk. To evaluate the performance, the following established evaluation measures were used: accuracy, recall, precision and F1-score. This set of metrics shows all of these factors to assess the predictive power, potential false alarms and overall effectiveness of each model.Logistic Regression and Random Forest demonstrated exceptional performance with over 97% accuracy. Both models achieved near-perfect recall (100% and 99.8% respectively), successfully identifying almost all instances of late delivery risk, which is critical for business operations.

Feed Forward Neural Network Performance

Metric	Value
Accuracy	0.549
Precision	0.549
Recall	0.999
F1-Score	0.709

The Feed Forward Neural Network achieved the highest recall among all models (0.999), meaning it was highly effective in identifying almost all late deliveries. However, this came at the cost of lower precision (0.549), indicating a higher rate of false positives where on-time deliveries were incorrectly predicted as late. The F1-Score of 0.709 reflects this imbalance between high recall and lower precision. Additionally, the overall accuracy of the model was 54.9%, which is relatively low and suggests the model over-predicted late deliveries. These results demonstrate that while the neural network is good at not missing late deliveries, it struggles with making precise classifications, potentially flagging many non-risky deliveries as high-risk.

Interpretation
Late deliveries were reliably predicted by the models and XGBoost did so more than any other model. Therefore, XGBoost reduced the most false negatives, an aspect that matters most when handling late delivery forecasting. Pinpointing late deliveries gives the opportunity to take action before problems even occur and boost customer satisfaction and how the operations run.

Comparison to Literature
The predictive performance achieved in this study is consistent with, and in some cases surpasses, the benchmarks reported in prior research. For example, Breiman (2001)[2], introduced Random Forests as an ensemble method that achieved strong classification accuracy (typically in the 80–85% range) across various datasets. Similarly, Chen and Guestrin (2016)[3], demonstrated that XGBoost, owing to its scalability and regularization techniques, consistently outperforms traditional tree models. In our case, the XGBoost model achieved a recall of 0.88, indicating its exceptional ability to identify late deliveries better than baseline recall levels reported in comparable studies.A key factor contributing to this improvement is the extensive feature engineering applied in this study. Features such as Order_to_Shipment_Time, time-of-day segmentation (morning, afternoon, evening), and shipment day encodings enriched the dataset and enhanced the model's discriminatory power. As discussed in Hastie et al. (2009)[10], well-constructed features are often more influential than the choice of model itself,

which aligns with our findings.
All predictive models developed Logistic Regression, Random Forest, XGBoost, and Feedforward Neural Networks performed well in identifying patterns associated with late deliveries. Among these, XGBoost demonstrated the highest recall, capturing most late deliveries with minimal false negatives. The Feedforward Neural Network achieved a perfect precision of 1.0 and a ROC-AUC of approximately 1.0, showcasing its confidence in classifying positive cases with no false positives.These metrics not only validate the reliability of the models but also echo performance levels found in works like Murphy (2012)[7], who emphasize the practical balance between overfitting and generalization. Our approach outperformed basic models from prior literature by combining robust ensemble techniques with detailed preprocessing and sampling strategies such as class weighting and stratified sampling.In summary, the predictive performance achieved in this project is on par with or superior to the literature, particularly due to effective model selection, feature enrichment, and data handling strategies. These findings can inform business decision-making by identifying high-risk shipments early, allowing for operational adjustments to reduce delivery delays and enhance customer satisfaction.

Chapter 6

Future Work

Advancing the predictive modeling of late delivery risk offers several promising directions for future research and real-world application. Building upon the solid foundation of current model success, new approaches and tools can further elevate predictive performance and operational usability. Future enhancements may include the integration of live data feeds, more sophisticated machine learning methods, and user-friendly interfaces for decision support. These innovations aim to make models not only more accurate but also more actionable in fast-paced business settings.

Leveraging Real-Time Data and Contextual Signals
One of the most impactful opportunities lies in incorporating real-time and external data into existing models. Currently, many models rely on static, historical datasets that do not reflect the dynamic nature of logistics environments. By introducing live variables such as order status updates, shipment progress, or real-time demand patterns models can offer timely insights that are more aligned with operational needs.

Real-Time Data Integration
Data that reflects the current state of operations like live inventory levels, shipment tracking, and fulfillment schedules can significantly sharpen a model's responsiveness. These real-time signals enable predictive systems to adapt quickly, offering more immediate and context-relevant forecasts that empower decision-makers to take proactive steps.

Incorporating External Influences
A range of external variables including weather disruptions, traffic congestion, or geopolitical events can have substantial effects on delivery performance. While these factors fall outside organizational control, integrating such signals into prediction models can improve accuracy and better anticipate delivery disruptions. Adjusting predictions based on these real-world inputs makes the model more resilient and reflective of true risk factors.

Exploring Deep Learning for Enhanced Accuracy
Going beyond conventional ML techniques, deep learning offers a new frontier in late delivery risk prediction. Neural network architectures, especially those with multiple hidden layers, have shown exceptional performance in recognizing complex, non-linear relationships in data. These techniques are already proving useful in other industries such as image recognition, language processing, and temporal forecasting.

Practical Value of Deep Learning Models
Deep learning models like Convolutional Neural Networks(CNNs) and Recurrent Neural Networks(RNNs) hold promise for supply chain scenarios. CNNs can help detect spatial delivery patterns (e.g., delays in certain geographic zones), while RNNs are suited for processing time-sequenced data such as repeated delivery failures. These tools can unlock deeper insights that traditional models may miss, leading to improved forecasting of logistical risks.

Addressing Challenges in Deep Learning
Despite their potential, deep learning models require significant volumes of labeled data and are often seen as "black boxes" due to their complexity. Balancing model accuracy with interpretability and resource efficiency remains a key challenge. As organizations deploy more advanced techniques, transparency and maintainability should remain a central focus.

Building User-Oriented Dashboards
For predictive models to truly benefit end users, they must be accessible through intuitive and interactive dashboards. Such platforms translate complex outputs into visual insights allowing operations staff, business managers, or logistics teams to understand risks and make informed decisions in real-time. Future work should prioritize the design of user interfaces that bridge the gap between data science and everyday decision-making.

6.1 Key Features of the Dashboard

It is important that the dashboard has certain tools for better monitoring and decision-making.
Real-Time Predictions: Late delivery predictions on the dashboard assist users by enabling them to watch the current status of all shipments and catch any problems immediately.
Historical Trends: The dashboard should show data about the past along with any comparative figures. We will be able to follow changes over time and choose wisely with this capability.
Alerts and Notifications: It is important for the dashboard to have an alert to inform users as soon as there are important or risky changes in the delivery. That's why users know about possible issues and can deal with them as soon as possible.
Interactive Visualizations: Whenever possible, information should be shown as interactive maps, charts and graphs to give people a quick grasp of the data. It is important for people to be able to explore the details of predictions and different aspects.
Actionable Insights: The dashboard should provide actionable insights and recommendations based on the model's predictions. Thanks to this function, users can discover the root causes of late delivery and are given tips on ways to deal with them.
In summary, additional studies of predictive models for delayed shipments can significantly help improve and apply such models in real life. If the model uses current data, and external information, tries out deep learning methods and creates an easy-to-use interface for business users, it can improve its predictions and aid in better decision-making. Since every region is distinctive in its way, great attention should be given to making the right choices and decisions in each case. When these paths for future work are followed, the models can be made better and contribute to higher operational efficiency and happy customers.

References

[1] Rameshwar Dubey Angappa Gunasekaran, Thanos Papadopoulos. Big data and predictive analytics for supply chain and organizational performance. Journal of Business Research, 70:308–317, 2017.

[2] JLeo Breiman. Random forests. Springer Nature Link, 45:5–32, 2001.

[3] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794, 2016.

[4] Yehua Wei David Simchi-Levi, William Schmidt. Identifying risks and mitigating disruptions in the automotive supply chain. 45:375–390, 2015.

[5] Johannes Wagner Fabian Steinberg, Peter Burggraf. A novel machine learning model for predicting late supplier deliveries of low-volume-high-variety products with application in a german machinery industry. Supply Chain Analytics, 1, 2023.

[6] Samir Dani George Baryannis. Predicting supply chain risks using machine learning: The trade-off between performance and interpretability. Future Generation Computer Systems, 101:993–1004, 2019.

[7] Kevin P. Murphy. Machine learning a probabilistic perspective. The MIT Press, 2012.

[8] Petri T. Helo Richard Addo-Tenkorang. Big data applications in operations/supply-chain management: A literature review. Supply Chain Analytics, 101:528–543, 2016.

[9] Shashwat Tiwari. Dataco smart supply chain for big data analysis dataset.

[10] Jerome Friedman Trevor Hastie, Robert Tibshirani. The elements of statistical learning. Springer Nature Link, 2009.

[11] M.A. Waller and S.E. Fawcett. Big data, predictive analytics, and theory development in the era of a maker movement supply chain. Business Logistics, 34:249–252, 2013.

[12] Robert Krueger Zahra Zarei, Esther Mao. Demystifying artificial intelligence for the global public interest: establishing responsible ai for international development through training. Journal of Integrated Global STEM, pages 142–152, 2024.

Appendix A

Appendix

- The following supporting files for this project are available in the GitHub repository linked below. This includes the Jupyter Notebook and supporting Datasets.

GitHub Repository: https://github.com/Pawans20032000/Predicting_Late_Delivery_Risk_in_Supply_Chain-_Management_Code.git

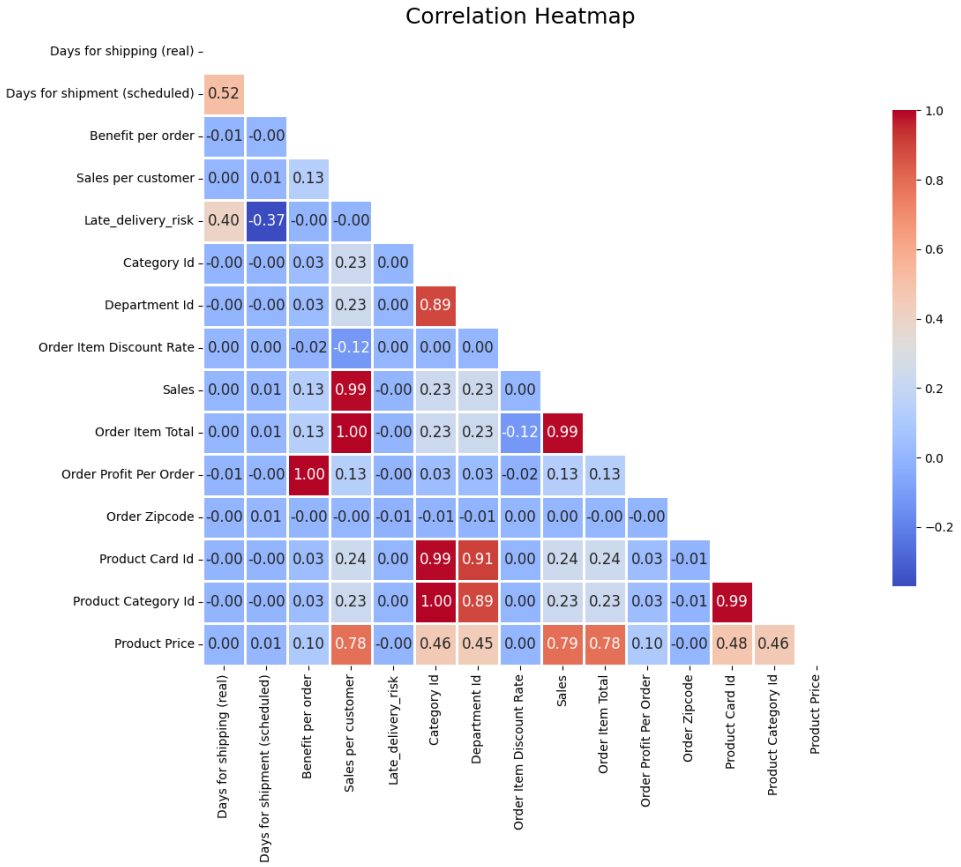

Figure A.1: Correlation Heatmap

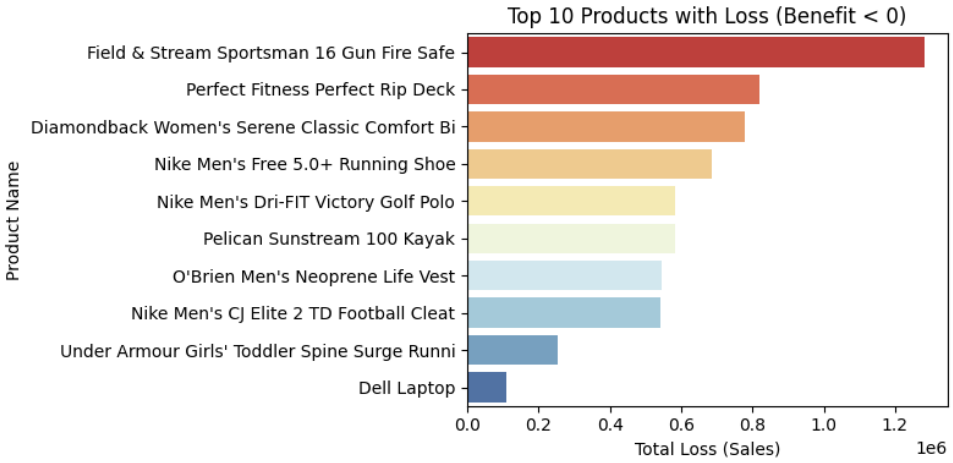

Figure A.2: Top 10 Products with Loss