NOVA IMS Information Management School

MDSAA

Master's degree Program in

Data Science and Advanced Analytics

FREIGHT AND LEAD TIME ESTIMATOR

Tonoyan Lilit

Work Project presented as partial requirement for obtaining the Master's Degree Program in Data Science and Advanced Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

FREIGHT COST AND LEAD TIME ESTIMATOR

by

Tonoyan Lilit

Work Project report presented as partial requirement for obtaining the Master's degree in advanced Analytics, with a Specialization in Business Analytics

Supervisor / Co Supervisor: Roberto Henriques

November 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 20.11.2023

ABSTRACT

The role of supply chain and logistics systems in today's global economy is vital, and accurate prediction of Freight Cost and Lead Time (FLT) is crucial for ensuring the effectiveness and efficiency of production planning and scheduling. Traditional FLT prediction methods are typically inadequate, as they rely on historical averages that fail to account for the variable nature of modern supply chains. This project introduces, from scratch, innovative predictive models through an industrial case study that leverage Machine Learning algorithms to consider key factors such as volume, weight, and destination. The new tools were created to fix past weaknesses in the company's approach to logistics, aiming to greatly improve how clients feel about the service. They provide a strong improvement by introducing prediction abilities that didn't exist before, helping the company become more agile and focused on client needs.

Keywords: Machine learning, Data Science, Predictive modelling, Supply chain management, Logistics.

INDEX

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Due to trends like e-commerce and globalisation, the logistics and transportation sector is proliferating. Logistics service providers create value by ensuring the timely delivery of goods to customers, acting as a crucial link between suppliers and end-users, and providing high-quality customer service (Wong & Karia, 2010).Companies that provide supply chain management services invest significantly in developing and implementing advanced tools and solutions within their organisation. By leveraging these tools, they anticipate an increase in their performance and the delivery of more customer-centric services. These efforts demonstrate a strong commitment to optimising their operations and providing the best possible service to their customers.

The integration of predictive systems in logistics can revolutionize how companies approach challenges such as fluctuating demand, variable shipping times, and the optimization of inventory levels. By accurately forecasting FLT, companies can pre-emptively address potential disruptions, thereby maintaining a consistent flow of goods, reducing storage costs, and increasing overall customer satisfaction. The benefits extend to enhancing strategic decision-making, allowing for more effective risk management, and fostering a proactive rather than reactive business approach.

Considering the transformative potential of predictive analytics, this study is prompted by the research question: "How can predictive analytics models for Freight and Lead Time (FLT) estimation improve supply chain management performance and customer satisfaction?" To answer this question, the research will pursue the following objectives:

1. To assess the current challenges faced by supply chain management in FLT estimation and identify the limitations of existing manual methods.
2. To develop and validate a predictive analytics model that accurately forecasts FLT using historical data and relevant variables.

The project is focused on developing a predictive analytics model for Freight and Lead Time (FLT) estimation in supply chain management. The motivation for this study stems from the increasing importance of FLT such as the enhancing supply chain operations, growing reliance on data-driven decisions, and potential benefits of predictive analytics. FLT is crucial in ensuring the efficient, timely, and cost-effective movement of goods throughout the supply chain. However, FLT estimation is often manual and prone to inaccuracy and time consumption. Hence, using predictive analytics can improve FLT estimation accuracy, increase operational efficiency, and provide valuable insights into supply chain management.

The importance of predictive analytics in supply chain management has been underscored by Brown et al. (2020), who found that AI methods can enhance lead time accuracy by 19 to 40 percent, suggesting a significant potential for cost savings and improved customer satisfaction. Similarly, Fussenegger (2022) highlighted the benefits of explainable machine learning models, which not only predict lead times with higher accuracy but also provide valuable insights into the decision-making process, making them particularly beneficial in complex manufacturing environments like the pharmaceutical industry.

The goal of creating an effective predictive model is to help Supply Chain practitioners by providing insights into the use of analytics in enhancing operations and reducing costs. The study also highlights the effectiveness of the CRISP-DM methodology by addressing real-world problems and developing effective solutions. The study comprises various chapters, including literature review, methodology, results, and discussion. The literature review chapter explores the theoretical framework of critical topics related to supply chain management, logistics, and predictive analytics for FLT estimation. The methodology chapter explains data collection and techniques of analysis. The results and discussion chapters present the findings and interpret their implications.

The primary objective of this research is to contribute to the literature on supply chain management, logistics, and predictive analytics. The results of this study offer valuable insights into the use of predictive analytics in FLT estimation and enhance the understanding of the role of FLT in supply chain operations.

This thesis aims to answer the pivotal question of how predictive analytics can transform FLT estimation to meet the demands of modern supply chains, with objectives focusing on evaluating the accuracy of AI models in FLT prediction and determining the business impacts of enhanced lead time precision, as demonstrated by the significant findings from Brown et al. (2020).

Effective logistics management is essential to the success of healthcare systems as it directly impacts patient care and outcomes. Proper management of medical supplies and equipment transport is necessary to ensure that customers receive their goods in time. This requires managing inventory levels, tracking expiration dates, ensuring proper storage conditions, and managing transportation and distribution channels. Logistics management can also play a critical role in cost control, as effective cost management can reduce waste, minimise overstocking and stockouts, and improve supply chain efficiency.

The project utilised data from a supply chain management company that ships to African countries. The study employed the CRISP-DM methodology, a structured approach to data mining projects that covers the entire process, from business understanding to deployment, on the dataset to reach its conclusion. The methodology consisted of six main phases: business understanding, data understanding, data analysis, preparation, modelling, evaluation, and deployment. Various machine learning methods and algorithms were utilised, such as Random Forest, XGBoost, Decision Tree, and Lasso CV in regressor mode. The research is significant as it provides insights into the real-world applications of machine learning in supply chain management, which can help improve operations and reduce costs.

2 LITERATURE REVIEW

This section outlines the theoretical framework of the key topics, such as supply chain, logistics, and predictive analytics and their use for FLT estimation.

The literature review starts with a general overview of Supply Chain and logistics management to get insights into the processes. The importance of logistics in the modern world economy and the contribution to the health system will follow these subchapters. Another part of the second chapter will be about the FLT and its importance for well-functioning supply chain companies. After the essential topics are covered, the study will focus on how the predictive analytics can be used in FLT estimation.

2.1 SUPPLY CHAIN

The Supply Chain is the network of business organisations, activities, and technologies involved in producing and delivering products and services from the raw materials to the final customer. It encompasses all the processes from the procurement of materials, manufacturing, distribution, delivery, and ultimately disposal of the product. The primary objective of the Supply Chain is to meet customer demand in a timely, efficient, and cost-effective way while ensuring quality and minimising the waste (Zsidisin & Ritchie, 2009). Even though the main purpose of Supply Chains is to reduce costs, identify suitable partners for product delivery, and maintain competitiveness in the market, a proper management is essential for successful operations (Monczka, 2009). Figure 1: illustrates a generic supply chain within the total supply chain network, which is a schematic overview of the typical supply chain flow, delineating the progression from suppliers through to customers. It highlights the critical stages of the supply chain including suppliers, manufacturers, distributors, retailers, and ultimately the customers, each interconnected by the movement of goods symbolized by arrows. This visual representation underscores the hierarchical and systematic approach to supply chain management that is essential for meeting consumer demand efficiently and competitively.

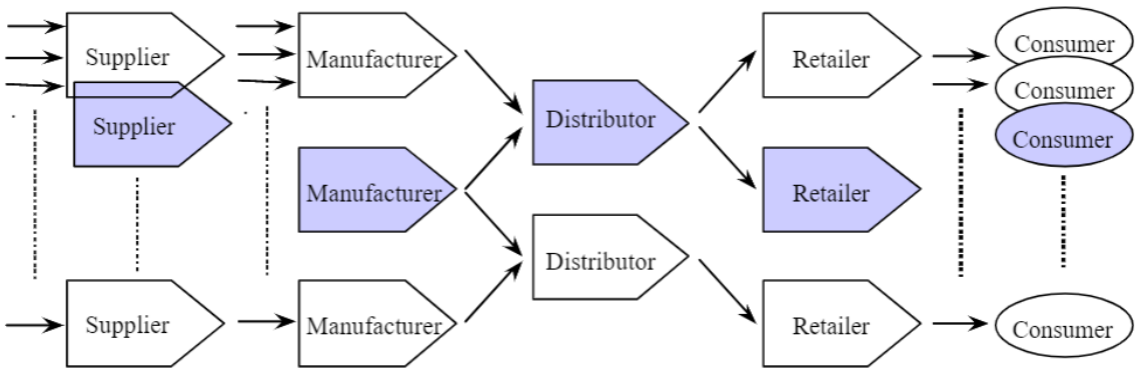

Figure 1: Schematic diagram of a supply chain network
Source: (der Vorst, 2004)

2.1.1 SUPPLY CHAIN MANAGEMENT

The goal of supply chain management is to manage relationships to maximise profits for all participants in the chain. This poses some complex problems, such as the narrow self-interest of some customers' needs to be put aside for the good of the chain. Although the term "supply chain management" is now frequently used, some might argue that it should be called "demand chain management" to assess that the market should drive the chain but not the suppliers. Supply chain management involves the design and control of all supply chain activities, including close cooperation with partners such as suppliers, intermediaries, third-party providers, and customers (Christopher, n.d.). Effective supply chain management is crucial for the success of any business, as it affects the company's profitability, competitiveness, and reputation (Zijm et al., 2019). The demands on global supply chain management in the twenty-first century are getting more and more complex (Branch, 2007).

2.2 LOGISTICS

Supply chain management aims to link and coordinate the processes of other parties involved in the pipeline, such as suppliers and customers. Logistics is a planning orientation and framework that establishes a unified plan to flow goods and information through an organisation (Christopher, n.d.). Logistics is the process of planning, implementing, and controlling the efficient flow and storage of goods, services, and related information from the point of origin to the point of consumption to meet customers' requirements. It encompasses all the activities involved in getting a product or a service from the manufacturer to the end user, including transportation, storage, handling, and distribution. (Zsidisin & Ritchie, 2009) Logistics is the art and science of management, engineering, and technical activities concerned with demand, design, supplying, and maintaining resources to support objectives, plans, and operations. (Definition by the USA Society of Logistics Engineers). Logistics is the practical aspect of supply chain management, which covers tasks like determining quantities, obtaining supplies, managing inventory, transportation, and collecting and reporting data. Supply chain management encompasses logistics activities, coordinating and collaborating with staff across levels and functions. While the supply chain includes the global supply and demand landscape, logistics tends to concentrate more specifically on program elements, for instance the health system (Logistics Management Units).

2.2.1 LOGISTICS MANAGEMENT

Logistics management is a component of Supply Chain management that focuses on the organised movement, storage, and distribution of material and goods, both incoming and outgoing, efficiently and effectively (Zijm et al., 2019).It is an integrated function, which coordinates and optimises all logistics activities with other functions including marketing, sales, manufacturing, finance, and information technology (SUPPLY CHAIN MANAGEMENT TERMS and GLOSSARY, n.d.).
The level of customer value determines the success or failure of any business, and Logistics management is almost unique in its ability to impact both the numerator and the denominator of the customer value ratio. (Equation 1)

$$Customer\ Value = \frac{Quality \times Service}{Cost \times Time}$$

Equation 1

The four main components can each be succinctly described as follows:

- *Quality*: the functionality, performance, and technical specification of the offer.
- *Service*: the availability, support, and commitment provided to the customer.
- *Cost*: The customer's transaction costs, including price and life cycle costs.
- *Time:* Time taken to respond to customer requirements, e.g., delivery lead time.

For each of these four components to maintain a competitive advantage, an ongoing program of innovation, improvement, and investment is necessary.

2.2.2 LOGISTICS FOR HEALTH SYSTEMS

Health care systems in the public sector are organised according to a vision for providing services. History shows that supply chains were developed as a secondary factor after it became clear that gaining access to commodities was essential for providing healthcare services. Staff members who are already overburdened with duties, typically lack formal logistics training, and occasionally do not have formal logistics responsibilities in their role have an impact on the efficiency of the Supply Chain. Because of this, Supply Chain failures like stockouts and product waste are frequent and expected. In-country supply chains frequently lack the necessary organisational structures to manage resources and operations effectively and an adequate number of experienced personnel devoted to routine logistics management (*Logistics Management Units: What, Why, and How of the Central Coordination of Supply Chain Management*, 2010).Logistics for health programs play a critical role in ensuring the availability of essential health supplies, such as medicines, vaccines, and medical devices, at the right place and time. It has many more important goals than just ensuring that goods are delivered. Ultimately, all public health logistics systems aim to guarantee commodity security, meaning that every person has access to high-quality essential health supplies whenever needed. Maintaining a properly functioning Supply Chain is critical for ensuring commodity security(Project, n.d.). Moreover, effective Supply Chain management helps ensure commodity security but also helps to determine the success or failure of any public health program.

An effective logistics system is essential for ensuring commodity security, reducing waste, improving the availability of essential health supplies, and, ultimately, improving health outcomes.

To maintain the flow of vital health supplies, a health logistics pipeline requires regular transportation to transfer the supplies from one warehouse to another. Supplies can be transported through various modes, such as boats, buses, or bicycles, in countries with varying geographical landscapes (Project, n.d.)

2.3 DATA-DRIVEN DECISION MAKING

Information is the driving force behind the logistics cycle, and accurate information is necessary to make better decisions. Data collection is one of the most important parts of the process and organisation. Useful data gathered across all system levels will ultimately improve customer service. The developments in IT have been very significant in the last decades and their use in daily life. Big data and analytics, which play an important role in the evolution, have opened new opportunities for

organisations (McAfee & Brynjolfsson, 2012; Shao & Lin, 2016). Big data and analytics are expected to lead to more supported decisions and better performance (Zambujal-Oliveira, 2019). Additionally, big data and analytics can be used to help organisations optimise their internal processes.
Data and process mining techniques have proven to be effective in converting raw data into valuable information, which can improve Supply Chain planning. The increased availability of computing resources and the development of optimisation methods have led to a focus on smart planning through comprehensive data analysis(Zijm et al., 2019).

2.4 AREAS OF OPTIMIZATION

Even though effective Supply Chain Management (SCM) leads to increased profits and sustainable competitive advantages for businesses, a more collaborative, proactive, and strategic approach is necessary. This can only be achieved through customer-focused strategies, resilience, efficiency, and agility (Zsidisin & Ritchie, 2009). Additionally, businesses must enhance their processes within procurement, outsourcing, information transfer and sharing practices scopes, and knowledge management. Companies must offer exceptional service and quality to capture a significant market share and improve customer satisfaction to compete effectively in the market. Price is no longer the only determining factor in purchasing decisions; consumers, who are becoming increasingly informed, prefer reasonable prices but also prioritise high-quality products and services and reasonable lead times in their purchasing decisions (Darko et al., 2018).

2.5 FREIGHT AND LEAD TIME PREDICTION IN SUPPLY CHAIN MANAGEMENT

One of the most important challenges in Supply Chain Management is Freight and Lead Time (FLT) prediction. Accurate prediction of the cost and time is fundamental for making informed decisions and ensuring that goods are available according to planned demand when and where they are needed. This requires a deep understanding of the factors that can impact freight and lead time, such as regulations, transportation modes, and market conditions. Overcoming this challenge is essential for maximising Supply Chain efficiency and improving customer satisfaction.
The accurate prediction of FLT significantly influences the quality and efficiency of production planning and scheduling (Lingitz et al., 2018). Traditional planning and control methods mostly calculate average values derived from historical data. This often results in the inefficiency of production planning and scheduling, as planners cannot consider the variability of FLT, which is affected by multiple criteria in today's complex manufacturing environment (Zsidisin & Ritchie, 2009)

Predicting a continuous target value is known as a regression task or problem. The prevalent techniques for a regression task include linear and logistic regression, which involve constructing a model to describe the association between the continuous variable (Kelleher & Mac Namee, n.d.).

2.5.1 LEAD TIME FORECASTING

Lead time is a crucial indicator of how well a Supply Chain is performing. It affects the overall cost of inventory as well as customer satisfaction (Abdel-Aziz M. Mohamed & Nermeen Coutry, 2015). The possibility to accurately predict the lead time allows for planning of stocks and detecting stockouts in

destination to the company. Also, the receiving warehouses could not guarantee the space to accommodate the medicine upon arrival and plan their spaces (Christopher, 2011). By continuously monitoring and improving lead times, organisations can increase their competitiveness and deliver products faster, resulting in greater customer satisfaction and higher levels of repeat business. If a company underestimates the lead time, it may face stockouts and supply chain disruptions that can harm customer satisfaction and the bottom line.
Conversely, overestimating the lead time can result in excess inventory and increased costs. By forecasting lead time accurately, companies can optimise their inventory levels, reduce stockouts, and improve their overall supply chain performance. (Fussenegger, 2022).

Forecasting lead time can be viewed as a regression problem, as its primary objective is to predict a continuous variable. Several studies have proved the effectiveness of regression algorithms in lead time prediction. Numerous studies utilise various models to forecast lead time in different industries and settings. For instance, one study compared the data mining approach using linear regression with other estimation methods calculated using mathematical formulas (Öztürk et al., 2006). Another research showed the effectiveness of eleven regression techniques for lead time prediction was analysed. The methodology used in the study was based on the initial five stages of the Cross Industry Standard Process for Data Mining (CRISP-DM) model. The study's findings indicate that the Random Forest (RF) model was practical and easy to comprehend(Lingitz et al., 2018)

Similarly, the study concentrated on unknown processing times in a complicated factory distribution, comparing four Machine Learning (ML) models. Ridge regression was the most effective model, resulting in a 30% reduction in the time it takes to complete a specific production process from start to finish (Yamashiro & Nonaka, 2021). Further research was about testing two approaches for estimating lead time: constructing a third model for lead time prediction and merging the outputs from separate models of actual and predicted lead time. The combination of predictions from the actual lead time and predicted random forest models improved the lead time accuracy by 38 days (Brown et al., 2020). It was tested If lead times are overestimated, orders are placed prematurely, leading to overstocking. Conversely, if lead times are underestimated, there will be backorders and reduced mission readiness. Estimations usually rely on historical data, especially for items with infrequent orders, and fail to consider other data sources that could enhance accuracy.

These models aim to accurately predict lead time, as it is a critical factor in determining inventory levels and impacts business outcomes. The use of such models has increased in recent years due to advances in technology and data analytics, allowing for more sophisticated and accurate predictions. Many researchers continue to explore and refine these models to improve their reliability and effectiveness in various applications.

2.5.2 FREIGHT COST FORECASTING

Forecasting freight rates have been an interesting topic in the shipping industry for a long time (Munim & Schramm, 2017).The significance of freight rates in the shipping industry is due to their function as a mechanism for adjusting the balance between supply and demand. This concept is emphasised in Stopford's (Lirn, 2011) work on the subject.

Freight cost calculation is crucial for suppliers and customers, as both parties need it for basic calculations and budgeting. Predicting freight costs helps to control transportation costs, which can be a significant part of a company's expenses. In addition, it is a critical factor for stakeholders in the shipping industry, including shipowners who rely on accurate forecasts to plan their operations,

lenders who assess investment risks, rating agencies involved in risk analysis, shipyards and marine equipment vendors, and ports seeking to develop new facilities, among others (Lirn, 2011). Accurate prediction of freight costs enables companies to optimise their transportation spending and ensure that they do not overspend on transportation. It also plays a crucial role in planning and forecasting, helping businesses make informed decisions about their Supply Chain operations. In addition, knowing the expected freight cost helps to ensure that goods are delivered on time and in the most cost-effective way possible.

The accuracy of freight rate forecasts is essential for stakeholders in the shipping industry, including ship owners, bankers, rating agencies, and marine equipment vendors (Randers & Goluke, 2007)However, accurate forecasting can be challenging due to the fluctuations, cycles, seasonal patterns, and randomness of the shipping industry. Researchers have attempted to model actual freight rates using approximations of freight rate functions that typically depend on the distance and weight of the cargo being transported (Swenseth & Godfrey, 1996). Given the origin and destination, such functions make it possible to calculate a straightforward market rate for a particular shipping lane. While these approximations may not provide precise forecasts, they could help achieve a high degree of accuracy in forecasting by carefully considering the factors that affect the shipping industry's volatility.

There has been growing research interest in predicting real freight rates in recent times, and scholars have started to include more lane- and market-specific variables rather than solely relying on distance and weight factors to portray the actual freight market conditions more precisely. In one study, carrier data was analysed to examine the revenues generated from various customers on different lanes through regression analysis. The researchers then compared the estimated and actual shipment rates to identify systematic differences and suggested that such differences could present opportunities for renegotiation. In this way, the carriers can potentially increase their revenue or reduce costs by renegotiating the shipment rates (USAID, 2014).

For instance, the case study conducted to identify the factors that impact the arrival time of trucks at a European Distribution Centre suggests that a "big data" approach can provide valuable insights, but it may not have as predictive solid power as anticipated. It was noted that factors beyond just data, such as human and organisational factors, could also impact arrival time. Thus, it is recommended that future predictive models should also consider these organisational factors to improve their accuracy. Overall, this study highlights the importance of taking an integrated approach to predictive modelling and considering a range of factors that could influence outcomes (van der Spoel & van Hillegersberg, 2017).

2.6 SUPERVISED, UNSUPERVISED AND SEMI-SUPERVISED LEARNING

Supervised learning is a type of machine learning that relies on labelled datasets to train algorithms, enabling accurate classification of data and predictions. In supervised learning, the model adjusts its weights based on input data that is fed into it, and this process of fitting the model to the data occurs during cross-validation (Shai Shalev-Shwartz & Shai Ben-David, 2014).
The goal of supervised learning is to create a model that can accurately predict the label of a given input feature vector (Andriy Burkov, 2019). For example, suppose the dataset contains information

about individuals and their medical history. In that case, the model could take a feature vector that describes a person's medical history as input and output a probability indicating whether that person has a certain disease, such as cancer.
The supervised learning process begins with selecting and preparing a dataset, which includes both the input feature vectors and their corresponding labels. The model is then trained on this dataset by iteratively adjusting the model weights based on the input data, with the goal of minimising the difference between the predicted labels and the actual labels in the training data (S.R.Das, 2016).
Once the model has been trained, it can be used to predict the labels of new, unseen data. This process involves inputting the feature vector of the new data into the model and obtaining a predicted label as output. The accuracy of the model's predictions can be evaluated by comparing the predicted labels to the true labels of the test data.
Overall, supervised learning is a powerful tool for creating models that can accurately classify data and make predictions. By relying on labelled datasets and iterative model training, supervised learning algorithms can be applied to a wide range of applications, from image recognition to natural language processing to medical diagnosis.

In semi-supervised learning, the dataset contains both labelled and unlabelled examples, where the labelled examples refer to data with pre-existing class labels, while unlabelled examples do not have any class labels assigned to them. Typically, the number of unlabelled examples is significantly larger than the labelled examples (S.R.Das, 2016). The main objective of the semi-supervised learning algorithm is to develop a model using the dataset. This model is expected to classify the data accurately and generalise well to new, unseen examples.

The primary idea behind using many unlabelled examples is that it provides the algorithm with additional information that can assist in creating a better model. This additional information can help in improving the model's ability to generalise well to new data. The unlabelled examples may provide valuable insights into the underlying structure of the data, which can be used to better estimate the model's parameters. Semi-supervised learning techniques attempt to leverage this additional information to create a more robust and accurate model (Andriy Burkov, n.d.).

Unsupervised learning uses algorithms to discover patterns in data sets where the data points are neither classified nor labelled. The algorithms are given the freedom to classify, label, and group the data points within the data sets without any external direction (S.R.Das, 2016).The process does not require the involvement of a supervisor, meaning that the learning algorithm is on its own to find structure in the input data without the benefit of labels. The four main types of unsupervised learning tasks are Clustering, Principal Component Analysis (PCA), Anomaly Detection, and Autoencoders. (Dridi, 2022)

- Clustering is the process of dividing a large set of data into smaller groups based on similarity.
- PCA is a technique used to reduce the dimensionality of data by transforming it into a new coordinate system that maximises the variation between the data points.
- Anomaly detection is the process of identifying data points that deviate significantly from the rest of the data set and are considered abnormal or unusual.
- Autoencoders are neural networks that are trained to reconstruct their inputs, allowing them to learn a compressed representation of the input data.

2.7 OVERFITTING AND UNDERFITTING

An incorrect inductive bias can result in two types of errors: underfitting and overfitting. Underfitting occurs when the algorithm chooses a prediction model that is too simple to capture the relationship between the descriptive features and target feature in the data. On the other hand, overfitting happens when the algorithm selects a prediction model that is too complex and fits the data too closely, making it sensitive to any noise in the data (Shaikhina et al., 2015). Overfitting is a frequent problem in supervised machine learning that cannot be entirely prevented. It occurs due to either the limitations of the training data, which may be small or contain a lot of noise, or the complexity of algorithms that require too many parameters (Ying, 2019). A simple example of the model fit is shown in Figure 2, where the model shows the relationship between apartment price and size.

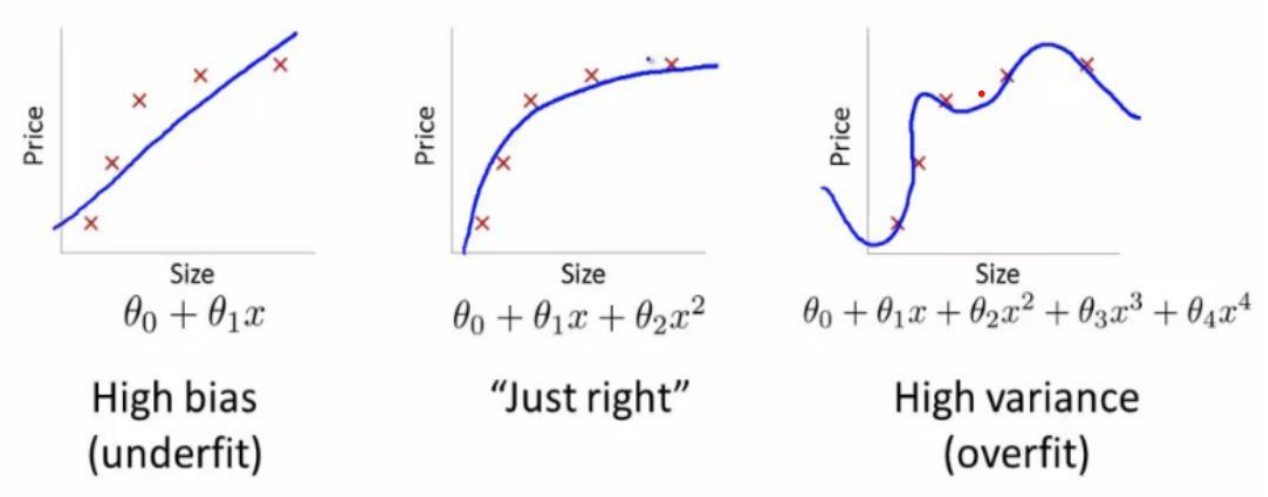

Figure 2: Relationship between apartment price and size
Source: (Vishal Maini & Samer Sabri, 2017)

Even though it is a general issue, which cannot be completely avoided, there are some common techniques used to circumvent overfitting in machine learning. It is important to experiment with different techniques and find the best approach for a specific problem. Splitting data into training and on the validation set and helps tune the model's complexity to achieve a good balance between overfitting and underfitting. Another solution example is early stopping. This technique stops the training process when the model's performance on validation set stops improving (Geron, 2019). Machine learning technique called cross-validation is a more advanced technique that uses multiple splits to get a more accurate estimate of the model's performance. It divides the data into multiple parts, trains the model on one part, and then assesses the model's performance on the other. In k-fold cross-validation, for example, the data is divided into k parts, or folds, and the model is trained k times, each time using k-1 folds for training and the remaining fold for evaluation. The average performance across all k iterations is used as the final performance metric (James, 2013)

3 METHODOLOGY

The main goal of the research is to develop predictive models through a real industrial case study for the calculations of Freight and Lead Time (FLT) with the help of Machine Learning algorithms. The data for the project is provided by a supply chain management service company that ships to African countries.

The data analysis was conducted using the CRISP-DM methodology outlined in subchapter 3.2. This approach followed the development of the model from the initial stage of Business Understanding to the final stage of Deployment. This methodology was deemed appropriate for the study as it aimed to address a real-world problem. A deeper understanding of the situation was achieved by taking a comprehensive approach to finding a solution.

3.1 PRESENTATION OF THE COMPANY

The work project at hand is a product of a well-known Supply Chain and Logistics service company. It offers customizable, high-performing, and reliable supply chain management services that enable clients to connect communities worldwide with the products they need to thrive. The company's extensive network of international logistics partners optimises delivery times and minimises disruptions, ensuring the timely delivery of life-saving commodities to those who need them most. The company aims to deliver critical health, humanitarian, and emergency response commodities to the world's most hard-to-reach places through fourth-party logistics.

3.2 PLATFORM/ENVIRONMENT SELECTION

The data analysis used Python with Visual Studio Code source-code editor. Python is widely used in data analytics and is one of the most popular programming languages. Its user-friendly syntax and rich libraries make it a versatile data manipulation, visualisation, and analysis tool. Python also has a thriving community of developers who continuously add new libraries and tools to the ecosystem, making it an excellent choice for data analytics projects of all sizes (Wes McKinney, 2013). The initial and descriptive analyses were performed using Panda’s software library written for Python.

3.3 CRISP-DM METHODOLOGY

The CRISP-DM methodology provides a structured approach to data mining projects, covering the entire process from business understanding to deployment (3 Kelleher & Mac Namee, n.d.-b).

The methodology consists of six main phases:

Define the project objectives and requirements and determine data availability and quality.

Data Understanding:

Collect and summarise the data and identify initial insights and relationships.

Data Preparation:

Cleanse and transform the data; handle missing or inconsistent values.

Modelling:

Select and apply appropriate data mining techniques; build predictive models.

Evaluation:
Evaluate the models, identify the best-performing model, and assess its strengths and weaknesses.
Deployment:
Plan for deployment and prepare a plan for model maintenance and monitoring.

This methodology offers a thorough and organised approach to data mining projects, assisting professionals in remaining concentrated, organised, and effective (Provost, 2013).

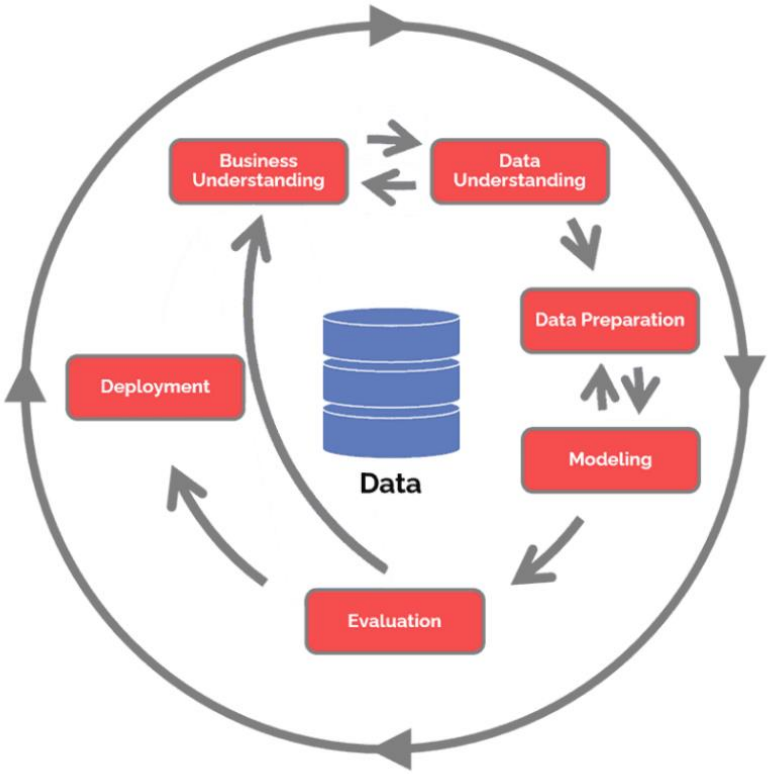

Figure 3: CRISP-DM Process
Source: (Kelleher & Mac Namee, n.d.-b)

3.4 BUSINESS UNDERSTANDING

A critical aspect of any data analytics project is identifying the specific business problem or need that the organisation is trying to solve. From this understanding, the practitioner can determine what type of information a predictive analytics model can provide to help address the problem. This definition serves as the basis for building a machine learning analytics solution (Shaikhina et al., 2015).

3.4.1 PROBLEM IDENTIFICATION AND PROJECT OBJECTIVE

A company providing Supply Chain services is facing some issues with existing forecasting tools. They are providing poor performance, resulting in inaccurate predictions and delays in shipments. This has caused disruptions in the company's operations, leading to lost revenue and unsatisfied customers. The objective of this project is to develop and implement more effective tools for FLT forecasting, using historical data to identify patterns and relationships that can be used to make more accurate predictions. The algorithms will be able to adapt and learn from new data, allowing for continuous improvement of the forecasting models over time. The goal is to improve the accuracy of predictions, which will enable the company to better manage inventory, schedule deliveries, and reduce lead times. This will result in improved customer satisfaction, reduced costs, and increased revenue.

3.5 DATA UNDERSTANDING

3.5.1 DATA COLLECTION

Once the business problems were obtained, data was collected from the Azure Synapse SQL Pool, which is a cloud-based analytics service. This cloud-based data warehousing solution enables users to store and manage large volumes of relational data for analysis and reporting., which can be accessed through a Structured Query Language. Once connected, the data was retrieved and transformed into the Parquet format, which allows for easier and faster loading of the data for further analysis. This approach to data storage and retrieval is designed to facilitate efficient and effective analysis of large amounts of data, enabling the company to make more informed decisions based on data-driven insights. After the data became accessible, an initial data exploration was performed to get primary insights into the data. It is important to understand potential data quality problems and identify steps that needs to be taken during data preparation phase.

The company collects data about the orders from customer for developing and testing purposes. Each data example consists of a set of features that describe an order: volume, weight, port of destination, transportation mode etc. The dataset provided for the project only includes orders from a single year (2021). It is important to train the model using post-COVID data, as the pandemic resulted in numerous disruptions and changes to the supply chain and logistics industry. These disruptions have the potential to significantly impact the accuracy of the model, making it difficult to predict future outcomes based solely on data from the pandemic period.
Initial datasets consist of 2,885,598 rows and 15 columns containing all the necessary information about the orders.

3.5.2 VARIABLES IN THE INITIAL DATASET

The following table gives information about the variables included in the initial dataset.

Variables	Information
RECIPIENT_COUNTRY	*Type: object* *Description: Destination country*
PORT_OF_DESTINATION	*Type: object* *Description: Destination port*
ESTIMATED_DELIVERY_DATE	*Type: object* *Description: Estimated delivery date calculated by company*
ACTUAL_SHIPMENT_DATE	*Type: object* *Description: Actual day of shipment*
SHIP_DATE	*Type: object* *Description: Day of shipment calculated by company*
TRANSPORTATION_MODE	*Type: object* *Description: transportation mode.* *SEA, LAND, AIR*
TOTAL_WEIGHT	*Type: float64* *Description: Total weight of the order*
TOTAL_VOLUME	*Type: float64* *Description: Total volume of the order*
ACTUAL_QUANTITY	*Type: float64* *Description: Quantity of products*
ITEM_ID	*Type: object* *Description: Product ID*
EST_LEAD_TIME	*Type: float64* *Description: Estimated lead time of the product calculated by company. In weeks.*
APPROVAL_DATE	*Type: object* *Description: Shipment approval date*
ACTUAL_DELIVERY_DATE	*Type: object* *Description: Actual delivery date*
FREIGHT_ESTIMATE	*Type: float64* *Description: Estimated freight cost calculated by company*
TOTAL_ACTUAL_FREIGHT_COST	*Type: float64* *Description: Actual freight cost*

Table 1: Variables in the initial dataset

The initial data analysis highlighted several problems:

- Duplicated values.
  The dataset contains information on all the orders made between 09.2021 and 09.2022. Normally, order information is stored with a unique ID. However, when a customer orders multiple products at the same time, the order ID, estimated delivery date, destination country,

and other values remain the same. It is also common for orders to have the same quantity of the same product ordered, resulting in a high number of duplicated values. The duplicated values were removed from the dataset leaving 44736 rows.

- Inaccurate data format.
  To ensure accurate calculation of lead time, it is necessary to have the date data in the appropriate date format. To perform this, we used *to_datetime()* function from Pandas on columns APPROVAL_DATE and ACTUAL_DELIVERY_DATE (Wes McKinney, 2013).
- Missing information.
  A crucial component for developing a lead time prediction model was missing, namely the actual lead time data.
  ACTUAL_DELIVERY_DATE - APPROVAL_DATE will give the value for ACTUAL_LEAD_TIME.
  The EST_LEAD_TIME value in the dataset is calculated in weeks so the new variable ACTUAL_LEAD_TIME had to be weeks. Pandas to_numeric() function converts the argument to numeric type and the value was divided by 7 to get the final value in weeks(Wes McKinney, 2013).

3.5.3 DESCRIPTIVE ANALYSIS

This section presents a fundamental overview of the primary dataset in its original form without modifications other than removing duplicated values. Overall, the dataset contains 54 Recipient countries and 120 different ports of destinations.
The Figure 4 represents the distribution of categorical variable transportation mode.

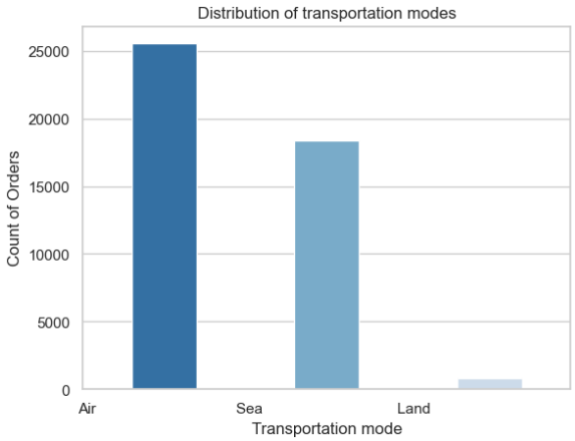

Figure 4: Distribution of transportation mode
Source: Own work

The graph shows that air transportation is the most common mode of transportation, accounting for approximately 58% of all transportation modes. Land is the second most common mode of transportation and finally, sea is the least common mode of transportation.

The boxplots illustrate the outliers in Estimated Lead Time, derived from the initial dataset where the company utilized basic Excel functions to calculate the estimated lead time based on average values from historical data, and Actual Lead Time. (Figure 5)

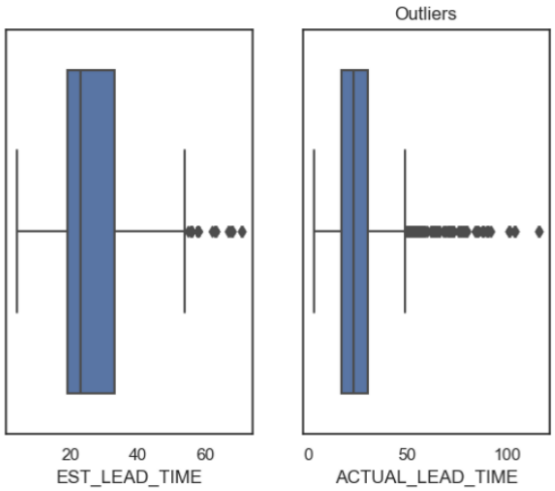

Figure 5: Distribution of Estimated and Actual Lead Time
Source: Own work

The comparison of the two boxplots suggests that there is a discrepancy between the estimated lead times and the actual lead times. Specifically, the estimated lead times tend to be lower than the actual lead times, and the actual lead times have greater variability than the estimated lead times. The fact that the median of the actual lead times is higher than the median of the estimated lead times suggests that the estimates may be too optimistic or not fully accounting for the complexity or variability of the tasks or projects. This could result in delays and may have implications for project planning and management.

The fact that the range of the actual lead times is wider than the range of the estimated lead times suggests that the actual lead times are more variable than the estimated lead times. This variability could be caused by several factors, including unexpected events, changing requirements, or variability in resources or personnel. The increased variability of the actual lead times suggests that there may be additional factors that should be accounted for in project planning and management to reduce the risk of delays. Overall, the comparison of the two boxplots highlights the importance of accurately estimating and accounting for lead times in project planning and management.

3.5.4 DATA QUALITY

Good data quality is essential for effective decision-making, accurate analysis, and the overall success of data-driven projects. Data quality issues, such as errors or inconsistencies, can have a significant impact on the accuracy and reliability of analyses and can ultimately lead to incorrect conclusions and poor decision-making. data quality is typically characterised by several key factors, including accuracy, completeness, and consistency (Batini & Scannapieco, 2006).

The data for the project was inputted manually to the system, which may explain the presence of several missing values and nulls. Upon analysis of the dataset, it became evident that there exist a significant number of duplicated rows. This is a common occurrence since each order can consist of multiple items, and each item is recorded in its own row with unique information. However, it is important to note that these duplicates have the potential to affect the accuracy and validity of the analyses. To mitigate potential analytical inaccuracies, a deduplication process was implemented. This

process involved identifying and removing or consolidating duplicate entries, ensuring the uniqueness of each data row.

To ensure the best possible results, it is imperative to give utmost priority to the quality of the data. This is because the accuracy and completeness of the data play a crucial role in any analysis. By focusing on data quality, a complete and error-free dataset can be ensured that is free from inconsistencies, errors, and missing values. This enables us to make informed decisions and draw reliable conclusions from our analyses. Conversely, poor data quality can lead to inaccurate analyses, incorrect conclusions, and costly errors. Therefore, investing time and resources into verifying and enhancing the quality of the data is essential to achieve the desired outcomes. Therefore, data preparation steps were taken, which involved removing any duplicates and addressing missing information. The focus was on creating a clean, reliable dataset ready for thorough analysis.

3.6 DATA PREPARATION

Data preparation is a crucial step in machine learning that involves transforming and cleaning raw data into a format that can be used for training the model. The goal of data preparation is to make the data consistent, accurate, and relevant so that the model can learn from it effectively. The process includes data cleansing, feature selection, data splitting etc (Jason Brownlee, 2020.).

1. **Removing Irrelevant Data and Missing Values**

During the initial descriptive analysis step a lot of missing values were detected, and to deal with them it was decided to use pandas *dropna()* method, which removes the rows that contains *NULL* values . After all the transformation the final dataset had 17833 rows and 10 columns.

2. **Categorical Variables**

There were several categorical variables Transportation Mode and Port of Destination were encoded using one-hot encoding method. One-hot encoding is a technique used in machine learning to convert categorical data into numerical data. It works by creating a new binary column for each possible category in the original column, where a value of 1 indicates that the row belongs to that category, and a value of 0 indicates that it does not (Raschka, 2015).

3. **Data Conversion**

After completing all the necessary data preparation steps, it was determined to convert the entire dataset to float data type to facilitate the modelling process. This decision was made because certain machine learning algorithms, statistical models, and data visualisation tools require input features to be in float format. Furthermore, float data type can provide greater accuracy in calculations, which can further improve the performance of the models (Fawcett & Provost, 2013). Therefore, converting the data to float data type ensures that the data is appropriately formatted for the modelling stage and enhances the potential for better model performance.

4. Data Splitting

The final phase of data preparation process is data splitting. Splitting the data into training and testing sets is a common practice in machine learning. The goal is to use a portion of the available data to train the model and then test its performance on the remaining data. This helps us assess how well the model will perform on new, unseen data (Joseph & Vakayil, 2021). The 20% of the data will be used for testing and 80% will be used for model training.

3.7 FEATURE SELECTION

Feature selection plays a crucial role in determining the most important attributes that contribute significantly to the prediction accuracy and overall performance of the model. One commonly used technique for achieving this is feature selection, which aims to identify the most important relationships or correlations between input and target features. It reduces the presence of unnecessary features, reduces learning time, and has the potential to enhance classification accuracy while preventing overfitting (Haar et al., 2019).

The F-test, which is also referred to as the ANOVA F-test, is a popular technique in supervised feature selection for machine learning. Its purpose is to assess the significance of the association between a specific feature and the target variable. By analyzing the variations among different groups or categories within the feature, the F-test determines if there exists a statistically meaningful distinction in the means of the target variable across those groups (Douglas C. Montgomery, 2015).

In the context of building a freight estimation model and predicting lead time, different F-tests were employed to assess the significance of various features for each model. By conducting separate F-tests for each model, the relevance of specific features in relation to freight estimation and lead time prediction can be determined. These F-tests help identify which features have a statistically significant impact on the respective models, allowing us to make informed decisions about feature selection and improve the accuracy and reliability of our predictions.

A. FEATURE SELECTION FOR FREIGHT ESTIMATE

In the analysis of the predictive model for freight costs, an F-test was performed to assess the significance of various features in estimating TOTAL_ACTUAL_FREIGHT_COST. This statistical test identifies which variables in the dataset have a strong association with the target variable TOTAL_ACTUAL_FREIGHT_COST. The F-test's results indicated that five features were statistically significant, suggesting they have a substantial influence on the freight cost variation. These significant features are unlikely to be impactful due to chance alone and provide meaningful contributions to the model's predictive accuracy. The importance of these features in the prediction equation is depicted in Figure 8, which serves to illustrate their relative impact on total freight cost predictionvariables are WEIGHT, VOLUME, QUANTITY, FREIGHT_ESTIMATE, PORT_OF_DESTINATION. However, due to limitations in the company's system, it was not possible to include the transportation mode as one of the features.

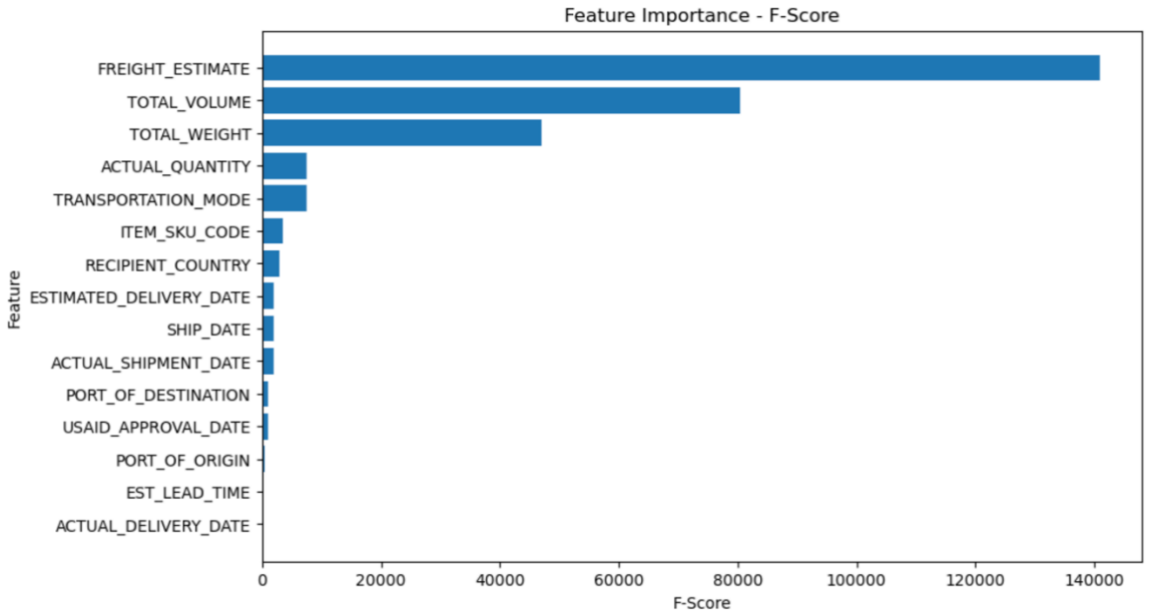

Figure 6: Feature Importance Score
Source: Own Work

Given the importance of capturing relevant information related to transportation costs and logistics, an alternative decision was made to use the port of destination as a substitute feature. While it may not directly represent the transportation mode, the port of destination is likely correlated with various factors that influence freight costs, such as distance, shipping routes, and trade regulations. By incorporating the port of destination as a feature, it is expected to provide valuable insights into the total freight cost estimation, compensating for the unavailability of the transportation mode in the analysis.

B. FEATURE SELECTION FOR LEAD TIME ESTIMATION

The feature selection process for the lead time estimation model was guided by the specific priorities and constraints outlined by the company. Due to factors such as time constraints and specific requirements, complex and time-consuming methods like F-score were not implemented. Instead, more straightforward strategy. Was used to identify the most relevant features for lead time estimation. The selected features for lead time estimation are the same as those used in the freight model, with one modification. Instead of using the freight estimate feature, it was replaced with lead time estimate. The final features for lead time model are WEIGHT, VOLUME, QUANTITY, EST_LEAD_TIME, PORT_OF_DESTINATION.

These attributes were carefully selected because they play a crucial role in improving the accuracy of lead time estimation, which is in line with the company's main goals. The weight, volume, and quantity of the shipment directly affect how long it takes for processing and transportation. Replacing the freight estimate with the estimated lead time improves the accuracy of the delivery time prediction. Additionally, including the encoding for the destination port allows the model to consider differences in lead time depending on where the shipment is headed. This thorough selection of features helps to better understand the factors influencing lead time estimation, leading to more dependable and informed predictions.

3.8 HYPERPARAMETER OPTIMIZATION

Before assessing and comparing the models, it is essential to individually optimize each model's hyperparameters to achieve the best performance. This involves experimenting with different parameter settings and measuring their performance (Bentéjac et al., 2019).
Due to the large number of parameters involved, the focus of the explanation is on the hyperparameters that were specifically optimized in the developed models. These optimized hyperparameters are the most relevant ones that can be adjusted to enhance the performance of the models.
The Random Forest model's performance can be improved by optimizing the following hyperparameters:

- max_features: The maximum number of features considered when looking for the best split in a tree.
- min_samples_split: The minimum number of samples required to split an internal node in the tree.
- min_samples_leaf: The minimum number of samples required to create a leaf/termination node in the tree.
- max_depth: The maximum depth of a tree which helps prevent overfitting.

By tuning these hyperparameters, Random Forest model's performance can be optimized and address potential issues such as overfitting and imbalanced data (RASMUS NYGREN, 2021).

Random Forest models consist of multiple decision trees. Each decision tree in the Random Forest is trained independently using a subset of the data. The final prediction in a Random Forest model is determined by aggregating the predictions of all the individual decision trees. Because of this they share some hyperparameters. For example, hyperparameters like max_depth, min_samples_split, and min_samples_leaf have similar meanings and effects in both Decision Tree and Random Forest models (Breiman, 2001)

The most relevant hyperparameters that can be optimized to improve model performance of XGBoost are: (Tao et al., 2021)

- min_split_loss: The minimum loss reduction required to make a further partition on a leaf node.
- min_child_weight: It defines the minimum sum of instance weight needed.
- max_depth: It determines the maximum depth of each tree in the XGBoost ensemble.

To maximize the efficiency of the models, a grid search and random search was conducted to evaluate their performance and determine the optimal values for the hyperparameters. It should be noted that the optimized parameters vary depending on the specific algorithm employed (Bentéjac et al., 2019)

3.8.1 OPTIMIZATION TECHNIQUES

Grid Search and Random Search are two powerful techniques for hyperparameter optimization, each with its own strengths. Both approaches were utilized to fine-tune the Random Forest and XGBoost models, as they demonstrated promising performance during initial experimentation.

For the Random Forest model, Random Search was employed to explore the hyperparameter space efficiently. Random Search is particularly effective when the search space is extensive and the computational resources are limited (James Bergstra & Yoshua Bengio, 2012).
The hyperparameters considered for Random Forest optimization included:

```
n_estimators = [int(x) for x in np.linspace(start = 50, stop = 250, num = 5)]
max_features = ['auto', 'sqrt']
max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
max_depth.append(None)
min_samples_split = [2, 5, 10]
min_samples_leaf = [1, 2, 4]
bootstrap = [True, False]
```

After Random Search was performed with 100 iterations, the best Random Forest model configuration was identified for each case.
Grid Search is a widely adopted technique for hyperparameter optimization as it exhaustively searches through a predefined set of hyperparameters to identify the combination that yields the best model performance (James et al., 2013). Grid Search for the XGBoost model was utilized, considering a range of hyperparameters. This rigorous approach enables us to uncover the optimal configuration of hyperparameters, ensuring that the XGBoost model is finely tuned for the specific challenges of freight lead time estimation.

```
objective: ['reg:squarederror'],
max_depth: [3, 5, 10],
learning_rate: [0.01, 0.05, 0.1],
gamma: [0.5, 1, 2, 5, 9],
min_child_weight: [1, 5, 10],
n_estimators: [100, 500, 1000],
random_state: [101]
```

This grid encompasses a range of values for essential hyperparameters such as 'max_depth,' 'learning_rate,' 'gamma,' 'colsample_bytree,' 'min_child_weight,' 'n_estimators,' and 'random_state.' By varying these hyperparameters, we aim to identify the combination that minimizes the negative mean squared error ('neg_mean_squared_error') during training.

3.9 MODELLING ALGORITHMS

Models are practical applications of theories that are often represented as algorithms in data science. By running these models on data, we gain insights that deepen our understanding of the world through combining theory, model, and data (S.R.Das, 2016). Various machine-learning methods were tested to achieve the most effective predictive model for this project. These algorithms include Random Forest, XGBoost, Decision Tree, and Lasso CV in regressor mode.

The ensemble approach of Random Forest is chosen for its excellence in accuracy and robustness, especially when handling diverse datasets common in FLT estimation, as evidenced by its extensive use in various predictive modeling scenarios (Sheth et al., 2022). Decision Trees are selected for their simplicity and interpretability, which are crucial for understanding and communicating the decision-making process involved in FLT predictions (Sheth et al., 2022). XGBoost is employed for its superior performance in handling structured data and its effectiveness in large-scale predictive problems, making it a mainstay in winning solutions for data-driven competitions (Didavi et al., 2021). Lastly, Lasso CV is incorporated due to its ability to perform feature selection and regularization, ensuring a parsimonious model that is both interpretable and generalizable, particularly in high-dimensional data contexts common to FLT (Didavi et al., 2021)

Proper identification and categorisation of variables can ensure the correct selection of appropriate methods for data manipulation, summary, and display. For instance, quantitative variables, such as height, age, and income, can be further categorised as either continuous or discrete, which will determine the types of statistical analysis methods used. On the other hand, qualitative variables, such as gender, marital status, and occupation, can be further classified as either nominal or ordinal, which will dictate the types of visualisations and tests used for data analysis (Jason Brownlee, n.d.). Moreover, incorrect identification or misuse of variables can result in invalid conclusions or misinterpreting results. For example, using a nominal variable, such as country of origin, as a continuous variable can lead to biased or incorrect results. Therefore, it is essential to understand the nature and characteristics of variables and to identify them accurately in the data analysis process. Proper identification of variables enables us to choose the appropriate statistical methods, make accurate predictions, and draw valid conclusions (Wes McKinney, 2013).

3.9.1 DECISION TREE REGRESSOR

Decision trees are algorithms that have a tree-like structure and are used to uncover patterns in data. They are favoured because they are simple to interpret and can be swiftly applied to problems that are not overly complicated. These models can accommodate both numerical and categorical variables and are not sensitive to outliers or missing values (Hastie et al., 2009).Binary regression trees have been introduced by Breiman (Leo Breiman, 1984). The study showed that binary regression trees could be used for regression tasks by predicting the mean value of the response variable in each leaf node, and for classification tasks by predicting the majority class in each leaf node. They demonstrated the effectiveness of binary regression trees through a series of experiments, showing that they outperformed traditional regression models in terms of prediction accuracy and interpretability.
The tree is built by splitting the data into smaller and smaller subsets based on the values of the input variables and making predictions based on the most dominant class or target value in each subset. The final prediction is made by moving through the tree from root to leaf nodes while adhering to the conditions listed at each internal node (figure 3). In decision tree regressor, the predicted value for each leaf node is typically the mean or median value of the target variable in the training instances that fall into that nod (Vishal Maini & Samer Sabri, 2017). However, they can easily overfit the training data if not pruned or combined with other models. Decision trees are highly valuable tools that can be easily affected by changes in the training data. This results in the model changing significantly even with small variations in the data. Despite this, decision trees can serve as a good foundation for ensemble methods. This is because the variance and inaccuracy can be reduced by combining them with other techniques.

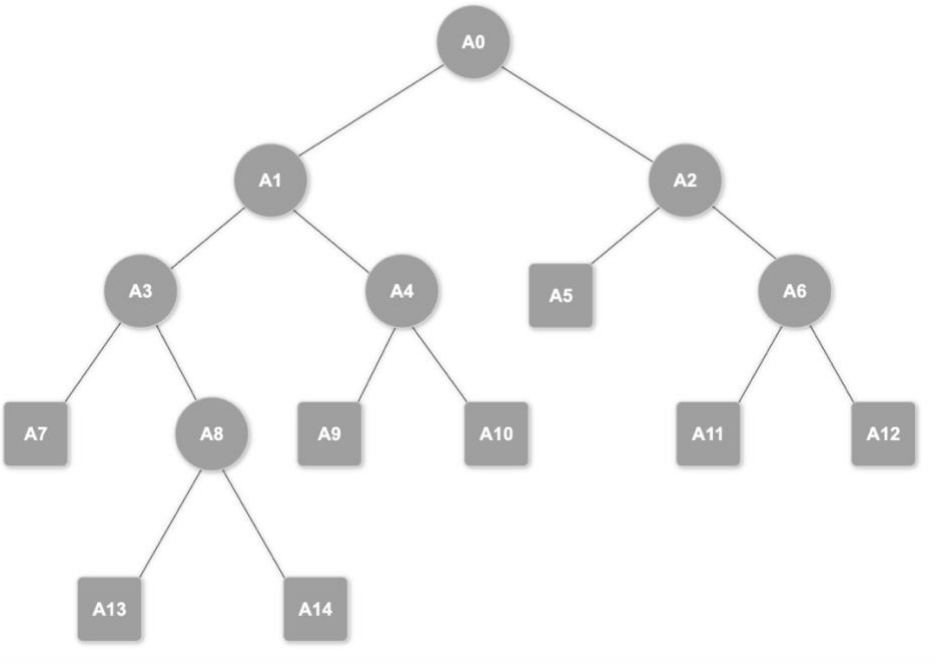

Figure 7: Decision Tree Example
Source: Own Work

Another reason decision trees are popular is because of their ability to work well even with messy data. They are relatively cheap to implement after training. Additionally, they are suitable for handling both numerical and categorical data. However, the training process of decision trees can be computationally intensive, and they have a high potential for overfitting(Vishal Maini & Samer Sabri, 2017). For instance, when building a decision tree to predict outcomes based on a set of features, having too many features and limited data can result in the tree overfitting the training data and generating too many decision rules. As a result, the accuracy of the tree's predictions for new data can be compromised (Breiman, 1984.)

3.9.2 RANDOM FOREST REGRESSOR

Random Forest is an ensemble machine learning algorithm that is used for both classification and regression tasks. A model composed of many models is called an ensemble model, and this is usually a winning strategy.
It operates by creating multiple decision trees and combining the results to produce a more robust and accurate prediction. In a random forest, each tree is built using a random sample of the data and a random subset of the features. Random Forest is one of the most widely used ensemble learning algorithms. The reason behind its efficiency is using multiple samples from the original dataset, to reduce the variance of the final model and low variance means low overfitting. (Andriy Burkov, 2019)
The main idea of the algorithm is.

- Select a random sample of data from the dataset, called the bootstrapped sample.
- For each node in the decision tree, randomly select a subset of features to consider when splitting the data.
- Select the best feature and split the data based on its values.
- Repeat the process for each split until a stopping criterion is reached (e.g., maximum depth, the minimum number of samples in a leaf node).

- Build multiple trees using different bootstrapped samples and different random feature subsets.
- Combine the predictions of the trees to make the final prediction (Breiman, 2001)

For a classification task, the final prediction is made by taking a majority vote on the predictions from the individual trees. The following formula uses class and probability to calculate the Gini of each branch on a node, indicating which branch is more likely to occur.

$$Gini = 1 - \sum_{i=1}^{c}(p_i)^2$$

Equation 2

Where:
pi - the relative frequency of the class observed in the dataset.
c - the number of classes

Where:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(fi - yi)^2$$

Equation 3

N - the number of data points
fi - is the value returned by the model.
yi - the actual value for the i point
The mean squared error (MSE) is used to solve regression problems. To determine which branch is the best choice for the forest, the formula measures how far each node is from the predicted actual value (Madison Schott, 2019)

3.9.3 EXTREME GRADIENT BOOSTING

The idea of boosting came out of the idea of whether a weak learner can be modified to become better. The goal is to repeatedly apply the weak learning method to produce a series of hypotheses, each of which is refocused on the instances where the previous one misclassified or had trouble understanding the examples (Yanofsky, 2015).
Extreme Gradient Boosting or XGBoost is a scalable ensemble technique based on gradient boosting, which has been successful in solving machine learning challenges with accuracy and effectiveness (Bentéjac et al., 2019). It works by iteratively adding new weak learners to the ensemble, each of which is trained to correct the errors of the previous weak learners. The algorithm starts by initialising the model with a simple weak learner, such as a decision tree which is called the base estimator. For the

following step the model calculates the differences between the predicted values of the base estimator and the actual values of the target variable, which are used as a target variable for the next weak learner. And the next weak learner is trained to predict the new target values instead of the original ones. After this the new weak learner is added to the ensemble, and its predictions are combined with the predictions of the previous weak learners using a weighted sum. The weights are determined by a learning rate, which controls the contribution of each weak learner to the final prediction. The adding and training steps continue until they reach stopping criteria such as a maximum number of iterations or a minimum improvement in the performance of the model.
This method has various benefits compared to other machine learning algorithms. For instance, it provides improved accuracy, can handle categorical and continuous data, and is robust to noise and outliers (Daniel Gutierrez, 2018).

3.9.4 LASSO CV

The word LASSO stands for Least Absolute Shrinkage and Selection Operator. It is a statistical formula for the regularisation of data models and feature selection. Lasso regression is a regularisation technique, which is used over regression methods for more accurate predictions. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (Ranstam J et al., 2016).
The goal of LASSO regression is to pinpoint the variables and their corresponding regression coefficients that result in a model with minimal prediction error. To accomplish this, the model parameters are constrained by 'shrinking' the regression coefficients toward zero, meaning that the sum of the absolute values of the coefficients must be less than a predetermined value (λ). This constraint limits the complexity of the model, and any variables with a coefficient of zero after shrinkage are excluded from the model. The value of λ is usually determined using an automated k-fold cross-validation method, where the dataset is divided randomly into k sub-samples of equal size. One of these sub-samples is then used for validating the model, while the remaining k-1 sub-samples are used to develop the prediction model. This process is repeated k times, and the results are combined for a range of λ values to determine the best λ for the final model. A significant advantage of this approach is that it reduces overfitting without limiting a subset of the dataset to use for internal validation (Tibshirani, 1996).

4 RESULTS

In this chapter, the results and evaluation of lead time and freight prediction models will be presented and discussed. A thorough analysis has been conducted to assess the performance of four distinct models: Random Forest (RF), XGBoost (XGB), Decision Trees (DT), and LassoCV (LCV).
The predictive accuracy of the models will be assessed using several key performance metrics:

- MAE (Mean Absolute Error) measures the average absolute difference between the predicted and actual values. It is calculated by taking the absolute difference between each predicted value and its corresponding actual value, summing these absolute differences, and then dividing by the total number of observations. A lower MAE indicates that the model's predictions are closer to the actual values, implying better accuracy (James, 2013).
- MSE (Mean Squared Error) calculates the average of the squared differences between predicted and actual values. It is obtained by squaring the error for each observation, summing these squared errors, and dividing by the total number of observations. MSE penalizes larger errors more significantly than MAE because errors are squared. It provides a measure of the model's ability to minimize outliers and is sensitive to extreme errors .(C. Bishop., 2007)
- RMSE (Root Mean Squared Error) is the square root of MSE. It is expressed in the same units as the variable being predicted (e.g., lead time). RMSE offers an interpretable measure of the average size of errors. It provides an estimate of the typical size of prediction errors in the original units of the data. Smaller RMSE values indicate better model performance in terms of prediction accuracy (Robert S. Witte & John S. Witte, 2016).
- MAPE (Mean Absolute Percentage Error) expresses prediction errors as a percentage of the actual values. It is calculated by taking the absolute percentage difference between predicted and actual values for each observation, averaging these percentage differences. It helps assess the relative accuracy of the model across different values. It is particularly useful when you want to understand how well the model performs in percentage terms. A lower MAPE suggests better accuracy (James, 2013).
- R- Squared R^2 measures the proportion of variance in the dependent variable (e.g., lead time) that is explained by the model. It is a value between 0 and 1, where 1 indicates that the model perfectly fits the data, and 0 suggests no relationship between the model and the data. It. evaluates how well the model captures the variability in the target variable. Higher R^2 values signify a better fit, indicating that a larger proportion of the variance in the dependent variable is explained by the model (Robert S. Witte & John S. Witte, 2016).

A. RSULTS: FREIGHT ESTIMATE

Comprehensive evaluation of the lead time prediction models is presented, utilizing various key performance metrics. The table below illustrates the results, employing abbreviated model names for clarity: RF (Random Forest), DT (Decision Trees), LCV (Linear Regression with Cross-Validation), XGB (XGBoost Regressor), and HT (Hyperparameter-Tuned Models).

	MAE	MSE	RMSE	MAPE %	Rscore
RF	113.907	123304.474	351.147	11390.71	0.995
XGB	0.00018	9.550	0.0003	0.018	0.999
DT	1.49587	1.53	3.91	1.49	1.0
LCV	306.67933	109917.741	331.5384	30667.933	0.983
RF_ht	2.088247	2.4930040	4.992	2.088	1.0
XGB_ht	0.0915	0.0408	0.2019	9.15	0.999

Table 2: Performance metrics results for Freight Estimate

Among the models, the XGBoost (XGB) stands out for its precision and accuracy in predicting freight. Its low Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) indicate that the predictions are very accurate and don't stray far from the real numbers. This precision is crucial in a business context where the costs of over- or under-estimating freight demand can be significant. Furthermore, the near-perfect R-squared (R^2) score demonstrates a fit to the data, suggesting that the XGB model has a strong predictive power.

While the Decision Tree (DT) model achieved a perfect R^2 score, this is typically a sign of overfitting, especially when such a score is not replicated in validation datasets. Overfitting is when a model learns the training data too well, including the noise and outliers, which means it may not perform as well when making predictions on new, unseen data.

Random Forest (RF) model also has a strong performance, with a very high R^2 value of 0.9953, indicating that it explains 99.53% of the variance in freight demand. This high degree of fit to the data suggests the RF model captures a significant portion of the variability in freight demand and is likely to make accurate predictions. However, it is important to note that while the RF's error metrics are slightly higher than those of the XGB, it still provides a balance between error minimization and risk of overfitting.

The Lasso CV (LCV) model, with an R^2 of 0.9838, shows it can explain a substantial amount of variance in freight demand. However, it does not reach the predictive performance levels of the XGB model.
The hyperparameter-optimized versions of Random Forest (RF_ht) and XGBoost (XGB_ht) aim to improve upon the baseline models.
Considering the balance between high predictive accuracy the XGB model is recommended as the most suitable and convenient for the project.

B. RSULTS: LEAD TIME

	MAE	MSE	RMSE	MAPE %	Rscore
RF	0,189682126	0,341803435	0,584639577	18,96821261	0,997179503
XGB	0,000152287	3,86209E-08	0,000196522	0,015228673	1
DT	0	0	0	0	1
LCV	1,265964124	1,815894295	1,347551222	126,5964124	0,887136335
RF_ht	0	0	0	0	1
XGB_ht	0,121029804	0,024710721	0,15719644	12,10298036	0,999922551

Table 3: Performance metrics results for Lead. Time

XGBoost (XGB) model shows precision in predicting lead times. With a Mean Absolute Error (MAE) of approximately 0.00015, the model shows an extraordinary ability to forecast lead times with negligible deviation from the actual values. This is further supported by a Mean Squared Error (MSE) that is virtually zero and an equally impressive Root Mean Squared Error (RMSE), indicating consistent accuracy across predictions.
Moreover, the Mean Absolute Percentage Error (MAPE) for XGB is just 0.015%, which signifies that the prediction errors are almost non-existent when compared to the actual lead times. Coupled with an R-squared (R^2) score that is as close to perfect as statistically possible, XGB provides an almost flawless representation of the underlying data patterns.

While the Decision Trees (DT) and the hyperparameter-tuned Random Forest (RF_ht) models show perfect scores in all metrics, such perfection suggests a strong likelihood of overfitting to the training data. Overfitting typically results in models that do not generalize well to unseen data, which diminishes their practical utility.

The Lasso Cross-Validation (LCV) model and Ridge Cross-Validation (RCV) demonstrate moderate accuracy and fit to the data, as indicated by their higher MAE, MSE, RMSE, and MAPE values, along with lower R^2 scores compared to XGB. These models may offer benefits in terms of interpretability and could be suitable for applications where a simpler model is preferred despite a trade-off in predictive performance.

In conclusion, based on the evaluation metrics, XGBoost (XGB) stands out as the best-performing model for lead time prediction due to its superior accuracy and fit to the data. I

4.1 EVALUATION

For the purpose of evaluation and achieving more clear and understandable results, a decision was made to allocate distinct sets of data instances to each model and subsequently observe their predictions.
The visual representation depicted in the graph (Figure 8) illustrates the performance in terms of freight costs denominated in USD using Rando Forest model. Notably, three lines are discernible, each showing unique insights. The initial line corresponds to predictions made by the company before the implementation of the machine learning model. These predictions were generated using basic Excel functions, and it is obvious that the information derived from this approach was inaccurate, serving more as a placeholder than a reliable source of data.
The second line in the graph represents the actual freight costs, calculated after the delivery process. This data is considered ground truth and serves as a benchmark for evaluating the efficacy of the machine learning model. Finally, the third line shows the performance of the machine learning model itself, showing its ability to predict freight costs accurately based on the provided data.

To provide more clarity, a piece of the code related to one of the data instances is presented below. It's noteworthy that, the code is configured to assume land transport as the default mode, which of course will be modified in the future.

```
real_data = {'PORT_OF_ORIGIN': ['BOM'], 'PORT_OF_DESTINATION': ['ROB'],
'TRANSPORTATION_MODE': ['Land'], 'TOTAL_WEIGHT':[5642.0],
TOTAL_VOLUME':[30.65],'ACTUAL_QUANTITY':[9648.0],'ITEM_SKU_CODE':['100062XXC08D2DS'],'FRE
IGHT_ESTIMATE':[109936.5],'BOM':[0.0],'ROB':[1.0]}
```

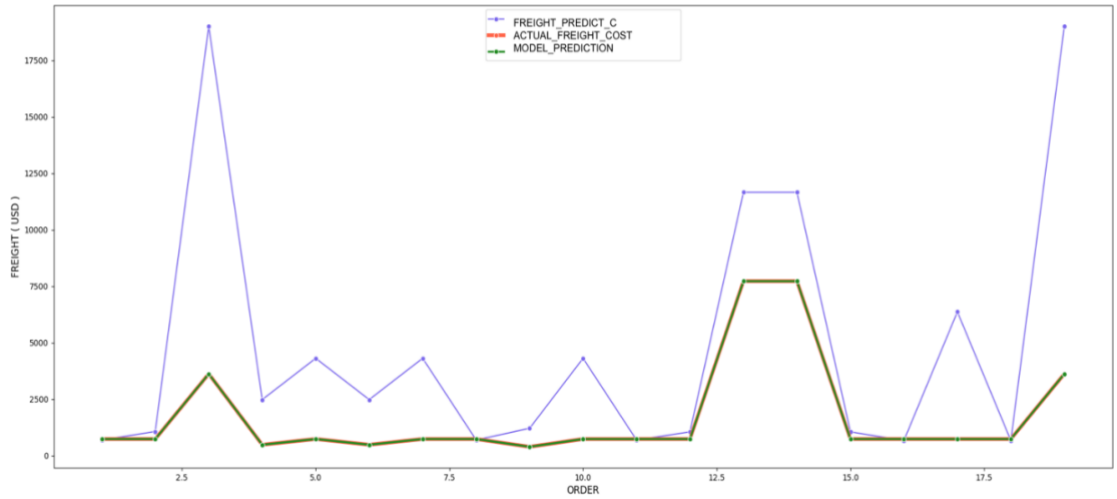

Figure 8: Model's Performance for Freight
Source: Own Work

The instance data for the lead time model presents a distinct scenario compared to the previous freight model. In this context, there is a graphical representation (Figure 9) consisting of three distinct lines, each providing valuable insights.

Blue line is representing predictions made by the company, like the freight model. This line reflects the company's forecasting efforts, which differ from the model's predictions.

Second one is the actual lead time, which is calculated by subtracting the approval date from the actual delivery date. This tangible measure of lead time serves as a benchmark and a tangible reference point for evaluating the accuracy of predictions. Lastly, the third line mirrors the performance evaluation of the lead time model, showcasing how well it aligns with the actual lead times. This aspect focuses on assessing the model's effectiveness in providing accurate lead time predictions.

In essence, the lead time model introduces a different dataset and evaluation framework, but the visual representation with its three distinct lines offers a comprehensive view of the prediction process, actual outcomes, and the model's accuracy, much like the freight model did. The code below shows the instance that was used for the model.

real_data = {'PORT_OF_ORIGIN': ['SHA'], 'PORT_OF_DESTINATION': ['HRE'], 'TRANSPORTATION_MODE': ['Land'], 'TOTAL_WEIGHT':[279.0, 'TOTAL_VOLUME':[2.29],'ACTUAL_QUANTITY':[2.0],'ITEM_SKU_CODE':['103953XUZ02X2DL'],'EST_LEAD_TIME':[46],'BOM':[0.0],'ROB':[1.0]}

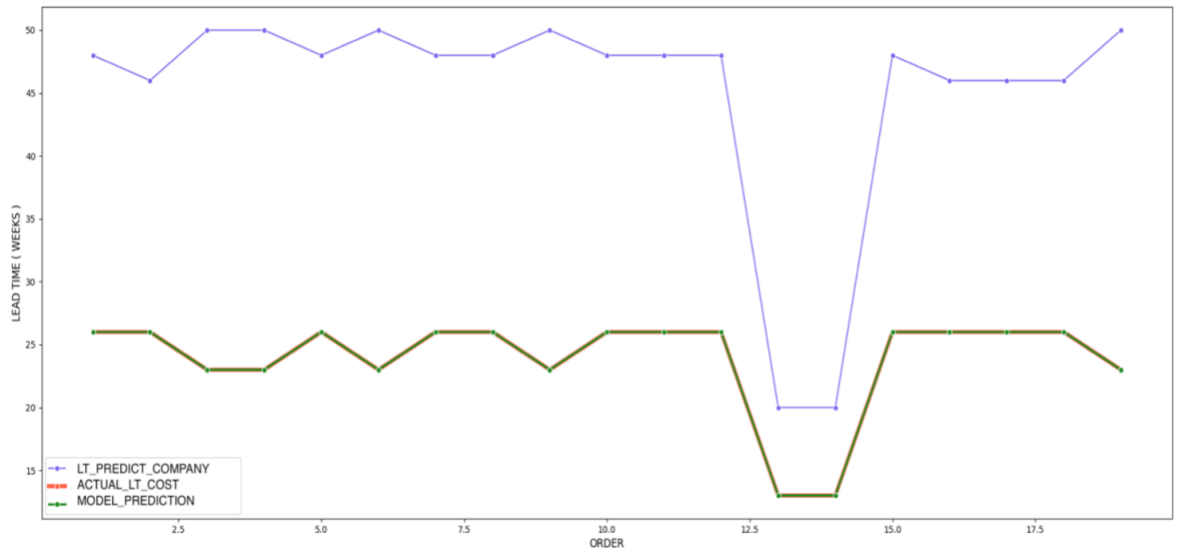

Figure 9: Model's performance for Lead Time
Source: Own Work

4.2 DEPLOYMENT

The final step of the CRISP-DM methodology is Deployment. Currently, the freight and lead time estimator models are still in the development phase. The next steps involve creating a user interface and adding more variables to improve their performance. This process requires careful consideration to ensure a smooth transition without disrupting ongoing operations.

Right now, the integration of the models is in the testing phase, where their performance is being evaluated in real-world situations. We are checking how well the models work and if they provide accurate predictions. To make the models even better, it is planned to include additional variables in the system. One example is integrating transportation mode data, which will help account for different ways of transporting freight, such as trucks, air cargo, or sea freight. This addition will make the estimations more accurate and comprehensive, which is crucial for logistics planning.

The user interface is also being developed to make it easier for people to access and use the system. The interface will be user-friendly, allowing users to input shipment details, choose transportation modes, and get estimated freight costs and lead times. By improving the user experience, we can help the company make informed decisions about logistics planning and resource allocation, ultimately optimizing their operations.

As the deployment progresses and the models prove effective, the company will be better equipped to improve their logistics processes. By making the right decisions and using the freight and lead time estimator, they can save costs and enhance their overall efficiency.

5 CONCLUSION

To wrap things up, this project has been the result of several months of hard work and in-depth research into the company's data, which was previously unknown and unanalysed. Main goal was to create two separate machine learning models, using data like 'PORT_OF_DESTINATION,' 'TRANSPORTATION_MODE,' 'TOTAL_WEIGHT,' 'TOTAL_VOLUME,' 'ACTUAL_QUANTITY,' and 'ITEM_SKU_CODE,' to predict lead times and freight costs.
Highlighting the fulfilment of the research question, "How can predictive analytics models for Freight and Lead Time (FLT) estimation improve supply chain management performance and customer satisfaction?" the study successfully developed and fine-tuned models that enhance the efficiency of supply chain management, evidenced by the more accurate budget planning enabled for clients.

This project holds great potential for the company, particularly in attracting more clients. The benefit for these clients is that they'll be able to plan their budgets more accurately, rather than relying on estimates that might not match reality. Precise lead time estimates will also be a boon for warehouse and order planning, especially since these clients typically order significant quantities of goods, necessitating careful advance planning.
In relation to the objectives, the challenge of FLT estimation was addressed, and the limitations of existing methods were identified, as laid out in the initial aims. The predictive models, particularly the XGBoost (XGB), have demonstrated the capability to overcome some of these limitations, thus answering the objectives.
Although models have performed well given the limited data available, it's important to emphasize ongoing efforts to enhance their performance. This ongoing improvement process underscores our commitment to ensuring that our models remain accurate and reliable, adapting to the changing needs of both the company and its clients.
Multiple predictive models were successfully developed and evaluated, including Random Forest (RF), XGBoost (XGB), Decision Trees (DT), LassoCV (LCV). Each of these models underwent careful training, evaluation, and in some cases, hyperparameter optimization to enhance their predictive capabilities. The evaluation of these models revealed diverse strengths and capabilities. Notably, XGBoost (XGB) emerged as the best-performing model for freight demand prediction, displaying exceptional accuracy and precision with minimal prediction errors. It is important to note that freight demand prediction, much like other predictive analytics challenges, comes with its own unique set of complexities. Some models exhibited higher accuracy, while others showcased robustness in specific areas. The choice of the most suitable model should align with the specific goals and constraints of your project.

While the results obtained from models hold promise and provide valuable insights, it is evident that further research and analysis are necessary to make a more significant impact on business operations. Access to additional data and a deeper understanding of the industry's dynamics can further enhance the precision and applicability of these predictive models The research has also met the objective of assessing current challenges in FLT estimation by applying machine learning models to real-world data, thereby confirming the hypothesis that such models can indeed improve the accuracy of freight cost and lead time predictions.

In conclusion, research has demonstrated that predictive analytics approaches can be effectively applied to the complex task of freight demand prediction. These models have the potential to aid in

optimizing business processes, resource allocation, and risk management in the context of package delivery. However, their performance can be further enhanced with more extensive data sources and industry-wide compliance practices. These models can serve as valuable reference tools and foundations for future data analyses within organization. The potential impact of this project on the company's operations is significant. It has the capacity to attract a broader client base by offering them a substantial advantage: the ability to plan their budgets with precision. Unlike relying on estimations that often deviate from actual outcomes, clients can now have access to accurate lead time predictions. This is not just about fiscal planning; it extends to the core of warehouse management, overall delivery optimization, and efficient order planning. Given the substantial volume of goods typically ordered by these clients, meticulous advance planning is essential to streamline operations and maximize efficiency.

While machine learning models have demonstrated solid performance, considering the limitations of a small dataset, it is essential to emphasize an enduring commitment to continuous improvement. The pursuit of enhancing model accuracy and reliability is an ongoing effort, ensuring adaptability to the evolving needs of the organization and its valued clients. This commitment underscores a determination to provide optimal service and solutions, positively influencing the company's growth and success.

This work symbolizes a commitment to addressing real-world challenges in the field of freight and lead time prediction, setting the stage for ongoing research and innovation in this vital field.

6 LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

One of the foremost recommendations for enhancing the model's performance centers around the augmentation of the training dataset. It is important to underscore that the quality and quantity of data play a crucial role in model performance. As the dataset expands, issues within the current models are increasingly uncovered. In essence, the adage "the more, the merrier" applies to data augmentation efforts. By enriching the dataset, doors are opened to uncover and address hidden issues in the existing models.
Additionally, in terms of the company's interface with the model, there exists considerable room for improvement. Currently, the company operates its own portal with a user interface. To use the full potential of the model, it is highly recommended that the user interface undergoes a substantial upgrade. This improvement should focus on offering a more complete display of the model's capabilities, ensuring that users can extract maximum value from the system.

However, it's essential to understand certain limitations that the company has. A notable constraint is the limited data within the company's possession. Gathering a large amount of data can take a long time, possibly extending over several years. While artificial data extension techniques and the possibility of procuring data from external sources are viable, these options raise important questions that need careful thought. The specifics of how to effectively expand the dataset, whether through artificial means or external data acquisition, represent issues that will require more extensive exploration in the future. Dealing with these issues is crucial to improve the model, which is currently held back by data limitations.

7 BIBLIOGRAPHY

Abdel-Aziz M. Mohamed, & Nermeen Coutry. (2015). *Analysis of Lead Time Delays in Supply Chain: A Case Study*.

Andriy Burkov, B. (2019). *The Hundred-Page Machine Learning*.

Batini, C., & Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques*. https://doi.org/10.1007/3-540-33173-5

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). *A Comparative Analysis of XGBoost*.

Branch, A. E. (2007). *Global Supply Chain Management and International Logistics*. Routledge.

Breiman, L. (2001). *Random Forests* (Vol. 45).

Brown, S. D., Khan, H., Salley, R. S., & Zhu, W. (2020). *Lead Time Estimation Using Artificial Intelligence*.

C. Bishop. (2007). Pattern Recognition and Machine Learning . *Springer, 1 Edition*.

Christopher, M. (n.d.). *LOGISTICS & SUPPLY CHAIN MANAGEMENT*. www.pearson-books.comwww.pearson-books.com

Christopher, M. (2011). *LOGISTICS & SUPPLY CHAIN MANAGEMENT*. www.pearson-books.comwww.pearson-books.com

Classification_and_Regression_Trees. (n.d.).

Daniel Gutierrez. (2018). *Gradient Boosting and XGBoost*.

Darko, S., Terkper, V., Novixoxo, J., & Lucy, A. (2018). *ASSESSING THE EFFECT OF LEAD TIME MANAGEMENT ON CUSTOMER SATISFACTION*.

der Vorst, J. (2004). Supply Chain Management: theory and practices. *Bridging Theory and Practice*.

Didavi, A. B. K., Agbokpanzo, R. G., & Agbomahena, M. (2021). Comparative study of Decision Tree, Random Forest and XGBoost performance in forecasting the power output of a photovoltaic system. *2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART)*, 1–5. https://doi.org/10.1109/BioSMART54244.2021.9677566

Douglas C. Montgomery. (2015). *Introduction to Linear Regression Analysis*.

Dridi, S. (2022). *Unsupervised Learning - A Systematic Literature Review*. https://doi.org/10.31219/osf.io/kpqr6

Fawcett, T., & Provost, F. (2013). *Data Science for Business*.

Fussenegger, P. (2022). *Explainable Machine Learning for Lead Time Prediction A Case Study on Explainability Methods and Benefits in the Pharmaceutical Industry*. www.kth.se

Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.). O'Reilly Media, Inc.

Haar, L., Anding, K., Trambitckii, K., & Notni, G. (2019). Comparison between supervised and unsupervised feature selection methods. *ICPRAM 2019 - Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, 582–589. https://doi.org/10.5220/0007385305820589

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*.

James Bergstra, & Yoshua Bengio. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* , 282–302.

James, G. and W. D. and H. T. and T. R. (2013). *An Introduction to Statistical Learning*.

Jason Brownlee. (n.d.). *Data Preparation for Machine Learning*.

Joseph, V. R., & Vakayil, A. (2021). SPlit: An Optimal Method for Data Splitting. *Technometrics*, *64*, 1–23. https://doi.org/10.1080/00401706.2021.1921037

Kelleher, J. D., & Mac Namee, B. (n.d.-a). *FUNDAMENTALS OF MACHINE LEARNING FOR PREDICTIVE DATA ANALYTICS*.

Kelleher, J. D., & Mac Namee, B. (n.d.-b). *FUNDAMENTALS OF MACHINE LEARNING FOR PREDICTIVE DATA ANALYTICS*.

Leo Breiman, J. F. C. J. S. R. A. O. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Lingitz, L., Gallina, V., Ansari, F., Gyulai, D., Pfeiffer, A., Sihn, W., & Monostori, L. (2018). Lead time prediction using machine learning algorithms: A case study by a semiconductor manufacturer. *Procedia CIRP*, *72*, 1051–1056. https://doi.org/https://doi.org/10.1016/j.procir.2018.03.148

Lirn, T.-C. (2011). Martin Stopford, Maritime Economics Routledge. Taylor & Francis Group (2009), 815. *The Asian Journal of Shipping and Logistics*, *27*, 355–361. https://doi.org/10.1016/S2092-5212(11)80016-7

Logistics Management Units: What, Why, and How of the Central Coordination of Supply Chain Management. (2010).

Madison Schott. (2019). *Random Forest Algorithm for Machine Learning*. Capital One Tech .

Monczka, R. M. (2009). *Purchasing and supply chain management.* South-Western.

Munim, Z., & Schramm, H.-J. (2017). Forecasting container shipping freight rates for the Far East – Northern Europe trade lane. *Maritime Economics & Logistics*, *19*, 106–125. https://doi.org/10.1057/s41278-016-0051-7

Öztürk, A., Kayalıgil, S., & Özdemirel, N. E. (2006). Manufacturing lead time estimation using data mining. *European Journal of Operational Research*, *173*(2), 683–700. https://doi.org/10.1016/j.ejor.2005.03.015

Project, D. (n.d.). *The Logistics Handbook: A Practical Guide for Supply Chain Management of Health Commodities, 2011 (corrected Nov 3, 2014)*.

Provost, F. and F. T. (2013). *Data Science for Business.*

Randers, J., & Goluke, U. (2007). Forescasting turning points in shipping freight rates: Lessons from 30 years of pratical effort. *System Dynamics Review*, *23*, 253–284. https://doi.org/10.1002/sdr.376

Ranstam J, Cook JA, & Br J Surg. (2016). *Statistical models: an overview*.

Raschka, S. (2015). *Python Machine Learning*. Packt Publishing.

RASMUS NYGREN. (2021). *Evaluation of hyperparameter optimization methods for Random Forest classifiers*.

Robert S. Witte, & John S. Witte. (2016). *Statistics*.

Shai Shalev-Shwartz, & Shai Ben-David. (2014). *Understanding Machine Learning: From Theory to Algorithms*. http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., & Khovanova, N. (2015). Machine Learning for Predictive Modelling based on Small Data in Biomedical Engineering. *IFAC-PapersOnLine*, *48*, 469–474. https://doi.org/10.1016/j.ifacol.2015.10.185

Sheth, V., Tripathi, U., & Sharma, A. (2022). A Comparative Analysis of Machine Learning Algorithms for Classification Purpose. *Procedia Computer Science*, *215*, 422–431. https://doi.org/https://doi.org/10.1016/j.procs.2022.12.044

S.R.Das. (2016). *Data Science: Theories, Models, Algorithms and Analytics*.

SUPPLY CHAIN MANAGEMENT TERMS and GLOSSARY. (2013).

Swenseth, S. R., & Godfrey, M. R. (1996). ESTIMATING FREIGHT RATES FOR LOGISTICS DECISIONS. *Journal of Business Logistics*. https://api.semanticscholar.org/CorpusID:166811393

Tao, H., Habib, M., Aljarah, I., Faris, H., Afan, H. A., & Yaseen, Z. M. (2021). An intelligent evolutionary extreme gradient boosting algorithm development for modeling scour depths under submerged weir. *Information Sciences*, *570*, 172–184. https://doi.org/10.1016/J.INS.2021.04.063

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. http://www.jstor.org/stable/2346178

USAID. (2014). *The Logistics Handbook: A Practical Guide for Supply Chain Management of Health Commodities, 2011 (corrected Nov 3, 2014)*.

van der Spoel, C. A., & van Hillegersberg, J. (2017). Predictive analytics for truck arrival time estimation: a field study at a European distribution centre. *International Journal of Production Research*, *55*(17), 5062–5078. https://doi.org/10.1080/00207543.2015.1064183

Vishal Maini, & Samer Sabri. (2017). *Machine Learning for Humans*.

Wes McKinney. (2013). *Python for Data Analysis*.

Wong, C. Y., & Karia, N. (2010). Explaining the Competitive Advantage of Logistics Service Providers. *International Journal of Production Economics*, *128*, 51–67. https://doi.org/10.1016/j.ijpe.2009.08.026

Yamashiro, H., & Nonaka, H. (2021). Estimation of processing time using machine learning and real factory data for optimization of parallel machine scheduling problem. *Operations Research Perspectives*, *8*, 100196. https://doi.org/https://doi.org/10.1016/j.orp.2021.100196

Yanofsky, N. (2015). Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World. *Common Knowledge*, *21*, 340. https://doi.org/10.1215/0961754X-2872666

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, *1168*, 22022. https://doi.org/10.1088/1742-6596/1168/2/022022

Zambujal-Oliveira, J. (2019). *2019/3 Operations Management and Research and Decision Sciences Book Series*.

Zijm, H., Klumpp, M., Heragu, S., & Regattieri, A. (2019). Operations, Logistics and Supply Chain Management: Definitions and Objectives. In *Lecture Notes in Logistics* (pp. 27–42). Springer Science and Business Media B.V. https://doi.org/10.1007/978-3-319-92447-2_3

Zsidisin, G., & Ritchie, B. (2009). *Supply Chain Risk: A Handbook of Assessment, Management, and Performance*. https://doi.org/10.1007/978-0-387-79934-6

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa