Prediction of ocean import shipment lead time using machine learning methods

Saraswathi Hathikal¹ · Sung Hoon Chung² · Martin Karczewski¹

Received: 13 February 2020 / Accepted: 22 May 2020 / Published online: 23 June 2020
© Springer Nature Switzerland AG 2020

Abstract

This paper focuses on developing a predictive model for estimating the shipment lead time using machine learning methods for ocean import freight considering interests of different stakeholders such as shipper, carrier, freight forwarder, and consignee. Two different terminal criteria for calculating shipment lead time are defined considering different milestones: one with empty container return and the other with delivery confirmation at the destination. Real data obtained from an industry partner are used for implementation, and multinomial logistic regression is identified as the best classifier with the highest accuracy in each binning method, which is followed by a decision tree method. Additionally, commonly used classifiers such as multinomial logistic regression, decision tree, *K*-nearest neighbors, and support vector machine perform better than Naïve Bayes when the categorical variables are binarized, and vice versa when the categorical variables are converted into ordinal values. The proposed model has the significant potential to benefit different parties in the supply chain by providing improved visibility and predictability for shipment lead times.

Keywords Freight forwarder · Ocean freight · Shipment lead time · Machine learning

1 Introduction

Globalization is extending local supply chain across boundaries catering to the customers demands throughout the world and is changing the landscape of trade and commerce [22]. With increasing globalization, many of the products and operations are outsourced and are moved across countries. Impact of transportation on the supply chain has always been crucial and is ever expanding. In the logistics and supply chain of import and export operations, intermediate transportation entities such as freight forwarders, also known as non-vessel operating common carriers (NVOCC), play a significant role in the movement of cargo from origin to destination [1, 8, 17, 22, 25].

Freight forwarders are service providers that manage shipment logistics for individuals and/or companies to move cargo from a shipper (manufacturer or producer) to

a consignee (customer) via a carrier (company providing a transportation mode). They are logistics experts and act as a nexus between shipper, carrier, and consignee. Freight forwarders use multiple modes of transportation for the movement of freight, and for global logistics and international shipment, there are two main categories, ocean freight and air freight, based on the mode of transportation. For the US domestic shipment, the primary modes of transportation are airplanes, railroads, and trucks. From an operational point of view, the primary difference between domestic and international shipment is the additional documentation and other activities associated with international trade. This paper mainly focuses on international ocean import freight (in particular, from other countries to the US case).

✉ Sung Hoon Chung, chung@binghamton.edu | ¹Lauffer Group International, New York, USA. ²Department of Systems Science and Industrial Engineering, Binghamton University, Binghamton, USA.

1.1 Role of freight forwarders for ocean import freight

For ocean import freight, freight forwarders act as mediators between shipper, ocean carrier, and customer. Ocean carriers are companies that operate the ocean freight vessels. There are many types of cargo vessels, and the most commonly used are container ships. The transportation services offered for containerized cargo are based on the container load, namely full container load and less than container load. Freight forwarders consolidate shipments from various shippers into full container loads. Freight forwarders have a sizable relationship with ocean carriers and the substantial buying power enables them to provide services at a reasonable cost for shippers or consignees. Besides coordinating with carriers and shippers, freight forwarders perform various other activities and services such as assistance in choosing the right ocean carrier, shipment tracking to shippers and consignees, documentation, cargo insurance, trucking services and door-to-door shipment delivery for domestic shipment if a shipper or consignee requests, customs clearance, warehousing facilities, and customer service [23, 34].

1.2 Overview of ocean import freight operations

Ocean import freight operations include a series of activities conducted to move the cargo from the origin to the destination through shipper, freight forwarder, ocean carrier, and consignee. A shipper requires transportation of freight from an origin (shipper location) to a destination (customer location that can be any place in the world) at a reasonable service level and cost. The shipper holds a contract with a third-party logistics service provider (freight forwarder or NVOCC), and the freight forwarder will have an agreement with ocean carrier companies. Shippers with a large volume of freight may make a contract with ocean carriers directly without the involvement of freight forwarders [9]. The freight forwarder manages the movement of a cargo from one point to another on behalf of the freight owner and sometimes provides ground transportation before and after the ocean transport. Ocean carriers collect, store, transfer, track, load, and unload cargo and also manage their capacity assignment. A consignee receives the shipment at the destination.

An ocean import process begins when a shipper places an order with an agent stationed at the origin country by providing necessary documents such as shipment specifications, commercial invoice, and packaging

list. The agent communicates the client requirements to the ocean import representatives of international freight forwarders. Further, the pricing representatives of the freight forwarder evaluate different pricing options based on the best routing and appropriate ocean carriers available for the given time of the year to maximize the overall benefits. The contract type (spot contract or fixed contract) and service type (port-to-port, port-to-door, or port-to-ramp) dictate the pricing for different customers. In the spot contract, tariffs and rates are subject to fluctuation, and rate changes are filed with the Federal Maritime Commission (FMC) in a timely manner for any rate increase or decrease to become effective and also to hedge customers against undue cost increases. In the fixed contract or non-vessel operating service agreement (NSA), shippers commit to shipping an absolute minimum number of containers for a specified period, and freight forwarders commit to a particular rate at the specified service level. Freight forwarders are responsible for arranging shipments from the port of the origin to the port of the destination for the port-to-port service, to a warehouse or consignee facility for the port-to-door service, and to an inland facility via ocean carrier and rail/truck for the port-to-ramp service.

After the pricing is done, ocean booking representatives of the freight forwarder communicate the booking details to the consignee. Upon receiving a booking confirmation, the booking representative provides a go-ahead signal for the overseas agent to make necessary bookings with an ocean carrier and trucker (if necessary) and enters the details in the interface operating system. Subsequent to the ocean carrier booking, the agent provides a master bill of lading (MBL) to the freight forwarder, a receipt of the acknowledgement for the goods received for transportation. It contains shipper's details, consignee's details, notifying party details, goods description, and the party information responsible for paying freight charges. The freight forwarder provides a house bill of lading (HBL), a replica of MBL concerning the information included in the document, to their customer. Most often, a shipper picks up the container from the port to load the shipments at its warehouse or manufacturing facility and returns the container to the terminal before the cutoff date. An ocean carrier will load the container into the vessel that departs from the port of the origin. Besides these activities, agents and representatives ensure the Automated Manifest System and Importer Security Filing information are transmitted to the US customs within 48 h and 24 h, respectively, before the cargo leaves the origin country.

During the transit of cargo, freight forwarders prepare necessary documents for customs clearance. A unique reference number (file number) is assigned for every booking to identify the shipment in the operations office. Incoming

and outgoing documents are recorded and reviewed for each consignment, and any errors or omissions are resolved. Relevant documents are correctly labeled and are made available to the customers. Seven days before the expected time of vessel arrival, ocean import documentation representatives send arrival notice document to the related parties, and a delivery order is sent to truckers to make arrangements for cargo pickup, if the freight forwarder is handling a trucking facility as well. Once the cargo is picked up, the trucker unloads the container at the customer warehouse facility, and the customer provides a delivery confirmation on receiving the shipment. Finally, the trucker returns the empty container to the terminal.

1.3 Global container supply chain

The role of freight forwarders and ocean import freight processes can be better understood in the context of

the global container supply chain. The global container supply chain is a system that consists of containers, vessels, port facilities, railways, and trucks, all of which work together to deliver goods around the world. Beyond the scope of physical entities that constitute the system, the global container supply chain can be viewed from a different perspective that has three interdependent network layers interacting with each other [37]. The three layers are: (1) logistics layer, (2) transaction layer, and (3) regulatory layer. The logistics layer includes all the physical entities that are required for transporting the goods, the transaction layer is responsible for procurement and distribution of goods and materials by coordinating with a network of suppliers, and the regulatory or oversight layer defines the operational standards and enforces rules and law by imposing penalties, taxes, and duties. The examples of stakeholders for the transaction and logistics layers are shown in Table 1, and the interaction across layers is depicted in Fig. 1.

Table 1 Example of stakeholders in transaction and logistics layers [37]

Layer	Example of stakeholders	Example of oversight agencies
Transaction	Walmart Target Non-Vessel Operating Common Carriers (NVOCCs)	Federal Trade Commission U.S. Customs and Border Protection World Customs Organization
Logistics	Pacific Maritime Association International Labor Organization Maersk Sealand Port of Long Beach	U.S. Department of Labor U.S. Department of Homeland Security U.S. Coast Guard World Customs Organization

Fig. 1 Interaction between different layers of container supply chain [37]

2 Problem statement and research objective

A shipment lead time can be defined as the total time for delivery of the shipment from the shipper at the origin to the consignee at the destination from the day of communication with the freight forwarder. The shipment lead time plays a critical role in supply chain management, and in most cases, it is not deterministic but stochastic as a variety of uncertain factors are involved in a supply chain [12, 18, 19]. Most customers (either shipper or consignee) are concerned about the shipment lead time even before they sign a contract with a freight forwarder, because it has a great effect on their businesses. However, calculating an exact shipment lead time is by no means an easy task mainly because of (1) uncertain factors such as weather conditions, delay in the container consolidation, ad hoc custom inspection, etc., (2) decentralized control, and (3) lack of streamlined data that can be obtained from information sharing via coordination and cooperation.

No single party is capable of controlling the entire chain of activities. Moving a cargo from an origin to a destination entails an operational cooperation and coordination across stakeholders, especially when it moves from one stakeholder to the other. While decoupling the drivers of supply chain perhaps results in increased productivity locally and temporarily, it will also cause inherent coordination issues within the supply chain, which may result in a poorly managed shipment lead time and a high variability. Therefore, information sharing across stakeholders via coordination and cooperation can have a significant impact on the supply chain performance, which has been studied in the literature [12, 19].

In this research, we focus on the freight forwarders as they are the mediators across all stakeholders in a supply chain. Shippers and consignees use freight forwarders for their 1) logistics expertise, 2) operational convenience, and 3) economical advantage. Freight forwarders focus on finding optimal balance between maximizing the customer satisfaction and offering services at a low cost. In this research, we notice that freight forwarders also have the unique capacity to access the important information obtained from different stakeholders in a supply chain, and we focus on developing effective machine learning methods to predict the shipment lead time for both customers (either shippers or consignees) and freight forwarder using a variety of information. For freight forwarders and customers, the last milestone for the shipment is the return of empty containers to the port. However, we also note that customers are mostly concerned about the actual delivery time to the consignee's warehouse facility. Accordingly, we define the

two terminal conditions to consider both parties (freight forwarders and customers) interests as follows:

1. Shipment Lead Time—Empty Container Return (SLECR): Time between the customer's first contact with overseas agent and the empty container return to the port
2. Shipment Lead Time—Delivery Confirmation (SLDC): Time between the customer's first contact with overseas agent and the shipment delivery confirmation to the consignee's facility

Based on the terminal criteria defined above, we propose the machine learning-based approaches to better predict the shipment lead time, the advantage of which from each stakeholder's perspective can be summarized as follows:

1. Freight forwarder/NVOCC:
 - (a) Improved customer service: It provides better visibility and predictability of the shipment lead time
 - (b) New opportunities/business: Informing potential customers about the more reliable shipment lead time will build credibility and significantly improve lead conversion rate

Fig. 2 Relationship between delays and capacity utilization [8]

2. Carriers: Improved asset management as summarized and shown in Fig. 2
3. Shippers and consignees: Improved inventory management, Informed decision making, Enhanced operations

In summary, the objective of this research is to provide more accurate shipment lead time that will be available to all stakeholders upfront, which will play a significant role to determine the performance of the ocean import process. This objective will be realized by way of the proposed prediction model, powered by the actual, streamlined data that encompass the entire ocean shipment process. The availability of such streamlined data motivates this research, and extensive data preparation methods such as factor identification via cause and effect analysis, subject matter expert consultation, preprocessing, and cleansing are used to extract only the useful information for the research. To the best of our knowledge, no research has addressed the ocean shipment lead time prediction model that explicitly employs the actual, streamlined data that encompass multiple stakeholders. The result of this research will make the ocean import process more streamlined, transparent, and reliable. The contribution of this paper can be summarized as follows:

1. Development of a predictive model to estimate the shipment lead time for ocean import shipment.
2. Consideration of different stakeholders' interests for ocean import shipment lead time.
3. Performance evaluation of machine learning methods by varying class label threshold value to classify the shipments into different classes.
4. Assessment of method performance with different treatment of categorical features.

3 Literature review

This section reviews the recent research, which focuses on solving prevalent problems in the field of logistics and supply chain using machine learning and artificial intelligence methods. Machine learning methods have been used extensively in the field of supply chain and logistics for a variety of topics. Recent examples include estimated arrival time prediction [32], agricultural supply chain analysis [30], international freight transportation [3], supply chain risk management [4], industry 4.0 application [13], etc.

Focusing on the shipment lead time, which is one of the most critical problems in supply chain transportation and logistics, topics such as prediction, on-time reliability, and major factor identification are extensively studied. For

example, Alcoba and Ohlund [2] propose a model for the prediction of on-time delivery of trucking activity. Authors claim that predicting on-time delivery of the shipment, one of the key metrics in the trucking industry, enables appropriate allocation of resources responsible for tracking and tracing of truckloads. On-time delivery is a binary response variable defined as truckload delivery within 1-h time window from the appointment time. The stratified sampling method is used to address the unbalanced data points for on-time and delayed delivery in the dataset. A stepwise logistic regression method is used to identify six significant variables (contract type, pickup duration, historical volume, incidents per volume, on-time pickup, delivery facility with or without appointments) that influence the probability of on-time delivery. The performance is further compared with the experimental results of the neural network and bootstrap forest model. Furthermore, Bleggi and Zhou ([5] present an experimental study that aims to identify the relationship between various independent variables and performance metrics for carriers such as on-time delivery, on-time pickup, and first tender acceptance individually and in combination. Logistic regression is used to recognize the attributes that have a higher influence on the successful performance of the carrier, and hierarchical clustering is used to group the asset- and non-asset-based carriers based on geography, industry, number of customers, and fleet size (number of truckers) to examine the underlying strategies employed by each carrier type. Furthermore, shippers are grouped to examine the portfolio of their carriers and their performance profiles. In addition, Chun [6] presents a model to capture fundamental variations in the inbound signals sources (estimated delivery date, estimated vendor delivery date, updated vendor delivery date, estimated carrier delivery date, and requested carrier delivery date) for shipment arrival based on the past error rate data for Amazon. Although the estimated delivery dates are regularly received from the vendors, most actual delivery dates are different than the estimated delivery date. As such errors or variations have a detrimental effect on the inventory management and the purchase order quantities for the future, the study focuses on mapping and developing a classification tree model to reduce the joint variation of estimated errors. The model is further used to generate a new estimated delivery date that is more accurate for the shipment arrival. Finally, the author provides potential policy changes for vendors to address the source causes of lead time variance.

Note that lead time prediction has been studied in other fields. For example, Lingitz et al. [20] use machine learning algorithms to predict lead time in a semiconductor manufacturer. The primary purpose in [20] is to enhance production planning and scheduling by

predicting manufacturing process lead time. While there are some similarities between [20] and our research, this research significantly differs from [20] in twofold: (1) application and (2) experiment method. With respect to the application, this research aims to predict ocean import shipment lead time to enhance shipment visibility and predictability considering interests of multiple stakeholders by predicting SLECR and SLDC. It benefits different stakeholders such as freight forwarder, consignees, and shippers in a different way, which is unique to the supply chain and logistics (ocean import logistics in particular) business. Therefore, special treatments and methods, as well as deep understanding of the business nature, are required. The inclusion of such makes this research very unique and very different to that of [20]. In addition, this research treats the prediction of shipment lead time as a classification problem rather than a regression problem to account for the inherent uncertainties, because there are a lot of uncontrollable factors that affect the shipment lead time. Hence, the outcome of the experiments is a range of number of days between which the shipment lead time is expected to belong and not a definite number. Lingitz et al. [20] treat the output variable as a continuous variable and uses regression models. In this research, the same set of features are used to evaluate the performance of all algorithms, while Lingitz et al. [20] select a certain subset of numerical features for linear models and consider both numerical and categorical features for nonlinear regression models. Additionally, Lingitz et al. [20] identify the top ten most important features using random forest and gradient boosting machine algorithms. This research examines the model predictability with intra-class variation. It is observed that the accuracy of predictability declines as the range is gradually reduced, which in turn increases the number of class labels for the outcome variables. That is, higher the granularity, lower the ability to correctly predict the shipment lead time for a consignment. In summary, this research evaluates the model performance in response to categorical variable treatment: converting categorical features to ordinal values and binarization of categorical features.

In addition to the shipment lead time, other topics such as shipment tendering process, error and fraud detection, cargo loss, and documentation are studied using machine learning methods. For example, Prasad et al. [26] examine the errors in shipment data for a freight forwarder. It is one of the recent papers that apply machine learning methods in the field of logistics. Two types of errors considered for their study include system errors that are flagged within the system and operational errors that are external to the system. Descriptive analytics is implemented to analyze the relationship between the shipment attributes and occurrence of errors. Various classification models are

employed to predict the shipment data entries that have a high occurrence of errors. Clustering methods are used to categorize the data entries to identify the factors that attribute to the occurrence of erroneous data which can further be extrapolated to the future records. Predictive models such neural networks and Naïve Bayes classifiers are used to predict the probability of error in a shipment data entry. Finally, a case study of the freight forwarder is presented to explain the experimental results. Kim [15] employs machine learning methods to analyze the rejection rates in the truckload tendering processes. When shippers make a contract with shippers by way of auctions, some carriers reject shippers' request if some conditions are not met. Such conditions include volume variability, length of haul, geographical patterns, and price difference. Despite various optimization methods for carrier selection, the shippers' tenders are rejected frequently. The major factors for tender rejections are analyzed using linear regression. Wu et al. [38] propose a business analytics research framework to study the reasons for cargo loss. In this paper, the authors provide guidance about how to obtain relevant logistics data and devise mitigation strategies to address cargo loss and enhanced logistic risk management. The cargo loss severity is defined based on the financial impact, and potential attributes include forwarder type, on-board date, origin, destination, commodity type, and mode of transportation. Decision tree and logistic regression are used, and the most important factors that affect the cargo loss severity are identified as transit type, commodity type, and destination. Other application areas of decision trees and classification methods can be found, e.g., [7, 33]. Furthermore, Triepels et al. [36] use Bayesian network model to predict the commodity type listed in the shipments, whose results are further used to examine occurrence of documentation fraud by comparing it with the corresponding shipment documents.

There are a sizable amount of studies that employ various machine learning and data mining techniques to address problems central to the global supply chain in general and freight forwarders in particular. However, most studies address the problems considering an individual stakeholder (e.g., either shipper, carrier, or consignee). This research focuses on predicting the shipment lead time considering potential factors affected by various stakeholders in the ocean import logistics.

4 Methodology

This section outlines the steps implemented in this research to predict the shipment lead time for ocean import freight, which are summarized as follows: data collection and analysis, data cleansing and preparation,

data preprocessing, implementation of machine learning methods for data modeling, and interpretation of the results. The raw data are collected with the cooperation of a partner company, which is then analyzed using descriptive statistics and other methods to obtain an overview of the variables in the dataset. Irrelevant, unnecessary, and erroneous data points are removed in the data cleansing step. Several techniques, the details of which will be introduced in the subsequent section, are applied for organizing and manipulating the dataset to serve the research objective in the data preparation step, after which the data preprocessing step follows. In the implementation step, machine learning methods such as multinomial logistic regression, decision tree, *K*-nearest neighbors, and support vector machine are employed. Finally, the results are discussed in the interpretation step.

Several assumptions are made based on the domain knowledge and experts' feedback from a partner company. In particular, the experts' feedback is utilized for the entire steps of this research, especially for data processing, prediction model building, verification and validation, and interpretation phases of this research. The assumptions are made to build an appropriate prediction model and prevent any inconsistency in the predictability of the response variable, which are summarized as follows:

1. Data, obtained from the partner company's database system, are assumed to be accurate and correctly entered by the ocean import representatives.
2. In the shipping industry, there are two kinds of dates: hard dates and soft dates. Hard dates (e.g., booking date) are fixed and cannot be changed, while soft dates (e.g., shipment ready date) are rather flexible and are subject to change to accommodate customers requests. This research assumes all the container event dates in the dataset are regularly updated and, therefore, latest for both hard and soft dates.
3. The customs clearance for consignments is assumed to be performed by either a freight forwarder or an external broker within the standard or required time window.
4. Each container is considered to be a shipment.
5. The consignments with delivery confirmation dates are assumed to be port-to-door shipment.

Based on these assumption, each step for predicting the shipment lead time is illustrated in more detail in the following sections.

4.1 Data collection and analysis

In this research, all the data points are obtained from a partner company, one of the leading New York-based logistic and supply chain company. In collaboration with global partners, the company provides the end-to-end logistics solutions and diverse customized services such as freight forwarding services including ocean import, ocean export, air import, and air export, customs and brokerage, pricing and quoting, consulting, customer service, and logistics. The company uses Application System/400 (AS400), a midrange server developed by International Business Machines (IBM), to manage and facilitate shipment delivery services. Before collecting the data, significant factors that would affect the shipment lead time for ocean freight are identified. First, an exhaustive list of potential factors affecting the lead time is created using the cause and effect diagram as shown in Fig. 3. The main categories considered for brainstorming all the possible causes under each group are as follows: overseas agent, shipper, carrier, shipper, carrier, broker, location and trucker, and the attributes of these main six categories are further examined as shown in Fig. 3.

The partner company use business intelligence software to extract relevant fields from the database to generate multiple reports to obtain useful insights regarding the business performance. For ocean services, data are obtained from so-called the Key Performance Indicator report. It consists of two subreports: shipment details

Fig. 3 Cause and effect diagram for factors potentially affecting the shipment lead time

Table 2 Variables in the shipment details report

Variable	Description
Dept	Type of department: import or export
File_no	Unique identification number for each shipment
Container_no	Identification number assigned to each container
Container_Type	Container type based on capacity
HBL_no	House Bill of Lading number
Booking_no	Booking number associated with each shipment
Customer_id	Customer identification number
Shipper_id	Shipper identification number
Carrier_id	Carrier identification number
PL	Port of loading
PDIS	Port of discharge
Final_des	Final destination
PO_no	Purchase order number

Table 3 Variables in the shipment dates report

Variable	Description
Dept	Type of department: import or export
PCC	Branch office code handling the files
File_no	Unique identification number for each shipment
Container_event	Significant series of events from shipper's initial contact to empty container return to the terminal
Container_event_date	Significant series of container event dates from shipper's initial contact to container return to the terminal

and shipment dates. The shipment details report provides the detailed shipment information such as booking reference number, carrier information, container details, port of origin, port of destination, and shipper and consignee information. The shipment dates reports provide a list of available container events for every file number created within the specified date range that is transmitted either via Electronic Data Interchange (EDI) from the carrier or entered manually by the company representatives into AS400. The variables in the shipment details and shipment dates reports are summarized in Tables 2 and 3.

The container event variable shown in Table 3, beginning from the shippers' first contact with the agent at the origin until the final event of empty container return at the destination, is further broken down and listed in Table 4.

4.2 Data cleansing and preparation

The two reports, shipment details and shipment dates reports, are generated by obtaining the information from the database to serve the managerial requirements that

are the primary purpose of the database. That is, the two reports are the secondary outputs and, because of this, there are some redundant data with overlapping dates. Such redundant data points need to be deleted before the organization of data can be done. The detailed ways of cleansing techniques employed in this paper are illustrated as follows:

1. In order to maintain the original dataset as much as possible, shipments with any null container event dates are removed.
2. Certain shipments not complying with the chronological order of container event are eliminated. For example, shipment data with a booking ready date occurring after the container return event at the origin port, and with the container pickup event happening before the container arrival event at the final destination are an obvious violation of the container sequence order. Such an error must have occurred due to, e.g., manual data entry. All such erroneous data entries are removed.
3. When a shipper sends a large volume of shipment, it will be distributed over multiple containers. Hence, any shipment with different container numbers with all the other fields remaining the same is eliminated.
4. It is identified that a shipper sometimes provides different purchase order (PO) numbers depending on the type of products. However, in the dataset, these are duplicate values with different PO numbers. Hence, any shipment with unique PO numbers with other fields staying identical is removed.
5. A shipper might provide different PO and HBL numbers for the same shipment depending on the type of product. If the all other details are the same but only the PO and HBL numbers are different, it means that the shipper wishes to identify the products being shipped separately. These are rare case scenarios where either the shipper or consignee has sent out special instructions to have multiple PO and HBL numbers. In this paper, any shipment with different PO and HBL numbers with all the other field remaining the same is eliminated to avoid any confusion.
6. Some shipments have a trucking conveyance from the port of arrival without involving any import services, which are known as the transload shipments. These kinds of shipments do not contain a series of container events but have only the container pickup information. Hence, all the transload shipments are eliminated.
7. As the shipment data include international transactions, there exists an inherent variability in handling the ocean shipment data mainly because of the inconsistencies in terminology, units, etc., across different countries. Such inconsistencies must be rectified.

Table 4 Container events in shipment dates report

Container Event	Description
I-Booking Create Date	Overseas agent creates booking in the AS400
Booking Create Date	The company's ocean import representative confirms booking on AS400
Booking Ready Date	Shipment will be ready and will be provided by the shipper
HBL Origin Date	Vessel is expected to depart
Booking RT Date	Container is returned to the terminal
Booking MV ETD	Vessel is expected to depart
Booking ETD	Vessel is expected to depart
HBL ETD	Vessel is expected to depart
MBL On-Board Date	Exact date container return to terminal date
MBL Return to Carrier Date	Container is returned to the terminal at the origin
File Create Date	File is created in AS400
Container LOD Date	Container is loaded into the vessel
Entry Date	File is created for custom clearance
Entry Transmit Date	Customs clearance is transmitted
Entry CR Date	Customs clearance is completed
Container CR Date	Customs clearance is completed
Container Last Free Date	Date until which storage fee is not applicable
MBL Discharge Date	Container is unloaded from the vessel
Container DIS Date	Container is unloaded from the vessel
Container AS Date	Vessel is expected to arrive
Container PU1 Date	Container is picked up from the terminal
First A/N Emailed Date	LGI communicates vessel arrival to the customer/broker
First DO Emailed Date	Delivery order is emailed to the trucker/customer for pickup
Container AD Date	Container has arrived the destination port
Container DC Date	Container delivery is confirmed
Booking ETA	Vessel is expected to arrive
HBL ETA	Vessel is expected to arrive
Container ER Date	Empty container is returned to the destination port

Otherwise, it will unnecessarily increase the dataset dimension resulting in higher computation time.

After redundant data are deleted, the two reports are combined to form a complete dataset. For each file number in the shipment details report, the corresponding container events are extracted from the shipment dates report. Monthly data obtained from the two reports are combined for every month. Furthermore, the dataset for each month is pooled together to obtain the shipment dataset for the duration of 13 months from January 2017 to January 2018. The container event dates or milestones are further arranged in a chronological order. Additionally, the variables are categorized into groups that conform with the steps that occur in a shipment process. Furthermore, after examining the data and discussing with the higher management, redundant dates are eliminated and specific relevant dates for each container event are retained. The variables that are rearranged and retained after redundant dates are removed are summarized in Table 5.

Customs clearance for the consignments is either performed by a freight forwarder or outsourced to a customs broker. For the partner company case this paper is dealing with, most customs clearance services are handled by the company. However, there are still a few cases handled by an outsourced broker for which the data for customs clearance are unavailable. Hence, it is assumed for such cases that the customs clearance is conducted within the specified or required time window.

4.3 Data preprocessing

Even after the dataset is cleaned and organized, it may not be directly used for the implementation of the prediction method. Depending on the objective of the study, the dataset still needs to be preprocessed. In this study, two terminal conditions for the shipment lead time are proposed as described earlier: SLECR and SLDC. Such terminal criteria can be calculated using the milestones in the shipment, and therefore, the variables in the dataset

Table 5 Variables retained after removing redundant date

Category	Container Event
Booking	I Booking Create Date Booking Create Date
Shipment	Booking Ready Date Booking RT Date
Loading	Container LOD Date
Departure	HBL ETD
Customs preparation	File Create Date Entry Date Entry Transmit Date First A/N Emailed Date
Customs clearance	Entry CR Date Container CR Date
Unloading	MBL Discharge Date
Arrival	HBL ETA
Delivery	Container PU1 Date First DO Emailed Date Container DC Date
Final container return	Container ER Date

need to be calculated and preprocessed based on such milestones. The total shipment lead time for both terminal criteria is calculated by taking appropriate milestones into account. SLECR and SLDC are calculated as the summation of number of days between milestones. The milestones for SLECR and SLDC are shown in Figs. 4 and 5.

Once the container has reached the final destination, a trucker will pick up the container, deliver it to the consignee's warehouse facility, and return the empty container to the port. Ground transportation may or may not be handled by a freight forwarder. If the freight forwarder is managing the ground transportation (port-to-door shipments), its documentation department will send the delivery order via email (variable name: first DO emailed date) to the trucker no less than 3 days from the container available date (HBL ETA). For such shipments, once the

Table 6 Calculated columns for SLECR

Calculated Column	Formula
tb.bc-ib	[Booking Create Date]- [I-Booking create Date]
tb.br-bc	[Booking Ready Date]- [Booking Create Date]
tb.brt-br	[Booking RT Date]- [Booking Ready Date]
tb.cl-brt	[Container LOD Date]- [Booking RT Date]
tb.hbletd-cl	[HBL ETD]- [Container LOD Date]
tb.mblldis-hbletd	[MBL Discharge Date]- [HBL ETD]
tb.hbleta-mblldis	[HBL ETA]- [MBL Discharge Date]
tb.ecr-hbleta	[Container ER Date]- [HBL ETA]

container is picked up (container PU date) and delivered to the facility, the consignee provides delivery confirmation to the freight forwarder (container DC date).

To predict the SLECR and SLDC, the entire dataset is divided into two sets, respectively: 1) SLECR dataset: the one with variables including the empty container return date (with or without the container pick-up date and delivery confirmation) and 2) SLDC dataset: the one with variables including container pickup date, delivery confirmation date, and empty container return date. The variable names and associated formula used in the dataset to calculate the time between milestones are shown in Tables 6 and 7.

4.4 Machine learning implementation

Recently, data mining and artificial intelligence are extensively used to discover hidden patterns and trends in the data, and help companies and organization to make better and informed decisions [16]. For the shipping industry, there are enormous opportunities to leverage the shipment data to enhance efficiency and facilitate better decision making despite particular challenges associated with the inconsistency in the data [39]. While there are many classification techniques and methods available, selecting an appropriate method for a specifically defined problem is crucial to obtain

Fig. 4 List of milestones considered for SLECR **Fig. 5** List of milestones considered for SLDC

Table 7 Calculated columns for SLDC

Calculated Column	Formula
tb.bc-ib	[Booking Create Date]- [I-Booking create Date]
tb.br-bc	[Booking Ready Date]- [Booking Create Date]
tb.brt-br	[Booking RT Date]- [Booking Ready Date]
tb.cl-brt	[Container LOD Date]- [Booking RT Date]
tb.hbletd-cl	[HBL ETD]-[Container LOD Date]
tb.mblldis-hbletd	[MBL Discharge Date]- [HBL ETD]
tb.hbleta-mblldis	[HBL ETA]- [MBL Discharge Date]
tb.cpu-hbleta	[Container PU1 Date]- [HBL ETA]
tb.dc-cpu	[Container DC Date]-[Container PU1 Date]

good results. Each method behaves differently based on certain assumptions, and hence, there is no single classifier or method that works well for all probable scenarios [28]. In this research, several machine learning methods such as multinomial logistic regression, decision tree, support vector machine, Naïve Bayes and K -nearest neighbors are used and evaluated to predict the shipment lead time using the cross validation, and a comparative evaluation of the prediction methods is conducted. Each method is briefly introduced in the subsequent sections.

4.4.1 Logistic regression

Logistic regression is one of the most widely used methods for classification. It is a generalization of linear regression where the response variable is binary and it models the probability of the outcome variable belonging to a particular class or category [14]. As it does not make any assumptions regarding the distribution of the independent variables, logistic regression can be used when the dataset consists of a mixture of categorical and continuous variables and response variable are categorical. It is recommended to use the logistic regression when the independent variables do not qualify assumptions of multivariate normality [31]. In the logistic method, the probability of occurrence $p(X)$ is modeled using a logistic function that gives outputs as 0 or 1 for all possible values of the independent variable X [35].

$$p(X) = \frac{e^{(\beta_0 + \beta_1 X)}}{1 + e^{(\beta_0 + \beta_1 X)}} \quad (1)$$

$$\frac{p(X)}{(1 - p(X))} = e^{(\beta_0 + \beta_1 X)} \quad (2)$$

$$\log \left(\frac{p(X)}{(1 - p(X))} \right) = \beta_0 + \beta_1 X \quad (3)$$

where β_0 and β_1 are the regression coefficients, which can be estimated using a maximum likelihood method. After

estimating the coefficients, the probability of an instance falling into a specific class is calculated using Eq. (1). Equations (1)–(3) are shown for one predictor variable, but they can be easily extended to include more predictor variables. In addition, while the logistic regression is in general for binary response variables, it can be extended to accommodate multi-class problems.

The main advantage of logistic regression is that it is easy to implement and computationally inexpensive, while the disadvantage is that it could result in low accuracy if the under-fitting occurs [11].

4.4.2 Support vector machine

Support vector machine (SVM) is a popular supervised machine learning method used for both linear and non-linear datasets [10]. SVM finds an optimal hyperplane that separates the instance of different classes from each other. The optimal hyperplane is searched and obtained using the support vectors (observations that are located on the margin for the respective classes). Observations are classified based on the side of the hyperplane. Unlike the maximal margin classifier, where the hyperplane exactly separates the class instances, that is extremely sensitive to changes in observations (which may cause overfitting problems), the support vector classifiers (also known as soft classifiers), allow a few instances to fall on the wrong side of the margin or hyperplane. This intentional misclassification makes the method robust and enhances the classification performance at the cost of incorrectly classifying a few training observations. The hyperplane is obtained as a solution to the following optimization problem [14] maximizing the margin width M subject to the following constraints:

$$\sum_{j=1}^p \beta_j^2 = 1 \quad (4)$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \geq M(1 - \epsilon_i) \quad (5)$$

$$\epsilon_i \geq 0; \sum_{i=1}^n \epsilon_i \leq C \quad (6)$$

where C is the nonnegative tuning parameter and ϵ_i is the slack variable for x_i , which permits an observation to fall on the wrong side of the hyperplane. The value of C reflects the measure of severity of misclassification that can be tolerated, and an appropriate selection of C helps to control the bias and variance trade-off.

SVM is an extension of support vector classifiers, which transforms the features into a higher-dimensional space using certain functions called kernels. Based on the feature

space, kernels can be linear, polynomial, or radial. SVM can be used to model nonlinear boundaries and is more robust to overfitting problems. However, one of the significant disadvantages of SVM is that training SVM models can be time-consuming.

4.4.3 Decision tree

A decision tree is simple and one of the most commonly used classification methods [21, 29]. It classifies the observations using a top-down approach by recursively implementing a divide and conquer method. Many methods are developed for training the decision tree using the greedy search since the seminal work by Quinlan [27]. Instances are segregated from the root node to many leaf nodes, which correspond to the class labels [24]. Every node in the tree tests one or the other features or attributes of the dataset and each branch from the node corresponds to any probable value of the attribute. To choose the best attribute that is beneficial for classification, a statistical measure called an information gain can be used, which measures how well a given feature or attribute can separate the training instances corresponding to the target classification. This quantity helps in selecting the best attribute at each node as the decision tree grows. Decision trees can also be converted into if–then rules for easy readability. One of the biggest advantages of decision tree method is its easy implementation and computational tractability. In addition, it can be used even when there are some missing values. However, it is prone to overfitting.

4.4.4 *K*-nearest neighbors

Given a set of training data, *K*-nearest neighbors (KNN) identifies and classifies new observations. Based on the majority vote of the similar classes, KNN assigns a class label accordingly. For a given value of *K* and test instance x_0 , it identifies *K* data points that are closer to x_0 , which is represented as N_0 . Furthermore, a conditional probability for a given class *j* is calculated as the proportion of data points in N_0 with class value *j*.

The KNN classifier is suitable for both numerical and nominal values as it makes no assumptions and it is insensitive to the outliers while providing a good accuracy. However, it may consume a significant amount of computer memory and is computationally expensive [10].

4.4.5 Resampling method: *K*-fold cross-validation

In resampling methods, samples are repeatedly drawn from the training dataset and the models are refitted on every sample selected to obtain more information (e.g., model variability) about the trained model. Since different

Table 8 List of final independent variables

Independent features
Month
PCC
Container Type
Customer ID
Shipper Code
Carrier
Port of loading
Port of loading country
Port of discharge
Port of discharge state
Port of discharge country
Final destination
Final destination state
Final destination country
tb.bc-ib
tb.br-bc

samples are trained multiple times using the same statistical learning method, it may be computationally expensive. Resampling approaches can be used for evaluating the performance. *K*-fold cross-validation is one of the resampling methods where training samples are randomly divided into *k* groups with an approximately equal number of observations in each fold. The first fold is used to validate the performance, and remaining *k* – 1 is used to train the model. This process is repeated *k* number of times where the different group is used as a validation set for each iteration.

5 Results

Given the prepared and preprocessed dataset, the machine learning methods discussed in Sect. 4.4 are implemented. The results are presented and discussed in this section. The implementation is divided into two sections: experimental setup and discussion of SLECR and SLDC results. Python 2.7 is used for the experimental setup and implementation on Windows 10 with the following system specifications: Intel Core i5-6300U CPU at 2.40 GHz, 8 GB RAM, and 64-bit operating system.

5.1 Response and input variables considered for prediction model

Identifier columns such as HBL number, file number, booking number, voyage number, purchase order number, container number, and other container event dates related to customs clearance and customs preparation are removed

from the original list of variables described in the Sect. 4.2. The final list of features considered for the shipment lead time prediction (SLECR and SLDC) is shown in Table 8.

Since there are abundant uncontrollable factors that affect the shipment lead time, this research treats the prediction of shipment lead time as a classification problem rather than a regression problem to account for the inherent uncertainties. Hence, the outcome of the experiments is a range of number of days between which the shipment lead time is expected to belong and not a singular number. For classifying the shipments, the response variable is grouped into different bins by assigning class labels for different threshold values. Since the dataset consists of both foreign-to-foreign and international shipments, median value (50 days) is considered as the initial threshold to avoid any influence of outliers present in the dataset. For SLECR, any consignment with the shipment lead time less than 50 days is labeled as 'A,' consignments with lead time greater than 50 days but less than 100 days are labeled as 'B,' and any shipments with lead time greater than 100 are labeled as 'C' for binning method 1. In binning method 2, the class interval is reduced to 25 days for each bin. In binning method 3, the shipment lead time range is further reduced to 20, and in binning method 4, the range is set

to 10 days. As the number of consignments with shipment lead time less than 10 days in the SLECR dataset is zero, the range of 0 to 10 is omitted in binning method 4 as shown in Table 9. For SLDC, a similar binning method is used. However, as the number of consignments with shipment lead time less than 16 days in the SLDC dataset is zero, the range of 0 to 16 is omitted as shown in Table 10. Based on the binning methods described above, the response variable for SLECR and SLDC is labeled as shown in Table 9 and 10, respectively.

After the cleansing and preprocessing steps, the dataset had 21,918 and 7,839 shipments with 16 independent variables for SLECR and SLDC, respectively. The classification methods introduced in Sect. 4.4 are implemented with different class labels assigned to the response variable according to the binning method discussed in the previous section. The accuracy is used as a performance metric to evaluate different methods, and the performance is compared for different categorical variable treatments by converting categorical variables into dummy variables (binarization) and ordinal values. In addition, the performance of the methods is validated with different resampling methods by dividing the dataset into training and

Table 9 Binning methods for SLECR

Binning method 1	Binning method 2	Binning method 3	Binning method 4
A - $LT < 50$	A - $LT < 25$	A - $LT < 20$	A - $10 \leq LT < 20$
B - $50 \leq LT < 100$	B - $25 \leq LT < 50$	B - $20 \leq LT < 40$	B - $20 \leq LT < 30$
C - $LT \geq 100$	C - $50 \leq LT < 75$	C - $40 \leq LT < 60$	C - $30 \leq LT < 40$
	D - $75 \leq LT < 100$	D - $60 \leq LT < 80$	D - $40 \leq LT < 50$
	E - $100 \leq LT$	E - $80 \leq LT < 100$	E - $50 \leq LT < 60$
		F - $100 \leq LT$	F - $60 \leq LT < 70$
			G - $70 \leq LT < 80$
			H - $80 \leq LT < 90$
			I - $90 \leq LT < 100$
			J - $100 \leq LT$

Table 10 Binning methods for SLDC

Binning method 1	Binning method 2	Binning method 3	Binning method 4
A - $LT < 50$	A - $LT < 25$	A - $16 \leq LT < 40$	A - $16 \leq LT < 30$
B - $50 \leq LT < 100$	B - $25 \leq LT < 50$	B - $40 \leq LT < 60$	B - $30 \leq LT < 40$
C - $LT \geq 100$	C - $50 \leq LT < 75$	C - $60 \leq LT < 80$	C - $40 \leq LT < 50$
	D - $75 \leq LT < 100$	D - $80 \leq LT < 100$	D - $50 \leq LT < 60$
	E - $100 \leq LT$	E - $100 \leq LT$	E - $60 \leq LT < 70$
			F - $70 \leq LT < 80$
			G - $80 \leq LT < 90$
			H - $90 \leq LT < 100$
			I - $100 \leq LT$

Fig. 6 Binning method 1—performance

testing dataset with shuffling with the fivefold cross-validation. The results are shown for fivefold cross-validation.

5.2 Prediction model for shipment lead time—binning method 1

Experimental results for the prediction of SLECR and SLDC using different machine learning methods, i.e., multinomial logistic regression (MLR), decision tree (DT), Naïve Bayes (NB), *K*-nearest neighbors (KNN), and support vector machine (SVM), are shown in Fig. 6. The accuracy, precision, and recall are used as performance metrics to evaluate the methods. The model is trained using cross-validation resampling method, where fourfolds are used for training and onefold is used for validation of the model.

As shown in Fig. 6, MLR, DT, and KNN outperform NB and SVM with test accuracy of 81%, 78%, and 79%, respectively, for SLECR and 86%, 85%, and 84%, respectively, for SLDC. The NB performance is lower relative to the other methods. This is mainly because NB assumes the conditional independence between features used to predict the outcome variable, which does not hold good for the dataset.

Table 11 summarizes the precision and recall metrics for all the methods. The precision metric is the ratio of true positives and the sum of true positive and false positive. It provides the proportion of shipments that are correctly classified into their corresponding class labels (shipment lead time range) out of the total shipments that are predicted positive by the model. The recall is the ratio of true positive and the sum of true positive and false negative. It provides the proportion of the shipments that are correctly predicted out of the total number of shipments that are actually positive (truly belongs to that range). For SLECR and SLDC, MLR has the highest precision of 82% and 85.6% and recall of 81% and 85.6%, respectively. It is followed by KNN, DT, SVM, and NB for SLECR and DT, KNN, SVM, and NB for SLDC.

Furthermore, experiments are conducted to evaluate the performance of the methods when the categorical features are converted into binary values and ordinal values. Figure 7 shows the results for such experiments. For SLECR and SLDC, MLR outperforms all other methods for binarization of categorical variable and DT performs best among all the other methods when categorical variables are converted into ordinal values. However, an interesting observation is that the accuracy of all the methods dropped when the categorical variables are transformed into ordinal values. This is due to undue importance attached to the variables with a higher numerical value.

5.3 Prediction model for shipment lead time—binning method 2

The same machine learning methods are implemented for SLECR and SLDC based on binning method 2. The same metrics are used to evaluate the methods. The results are shown in Figs. 8 and 9 and Table 12.

Figure 8 shows the comparative evaluation of the performance for SLECR and SLDC. Similar to binning method 1, MLR, DT, and KNN outperform NB and SVM with test accuracy of 76%, 72%, and 73%, respectively, for SLECR and 80.31%, 79.83%, and 79.02%, respectively, for SLDC. Naïve Bayes again performed poorly compared to all the other methods for the same independence assumption between predictors reason. Table 12 summarizes the precision and recall metrics for all the models for binning method 2. For SLECR and SLDC, MLR has the highest precision of 75% and 80% and recall of 76% and 80%, respectively.

Figure 9 shows the results for the transformation of categorical values using binarization and ordinal values. For SLECR and SLDC, MLR outperforms all other methods for binarization of categorical variable and DT performs best among all the other methods when categorical variables are converted into ordinal values. The accuracy of MLR, DT, KNN, SVM, and NB is reduced when the categorical features are converted into ordinal values. However, NB performs better with ordinal values compared to binarization method of treatment.

5.4 Prediction model for shipment lead time—binning method 3

The results for the prediction of SLECR and SLDC using different multi-class classifiers are shown in Figs. 10 and 11 and Table 13.

Figure 10 shows the comparative evaluation of the performance for SLECR and SLDC. Similar to binning methods 1 and 2, MLR, DT, and KNN outperform NB and SVM with test accuracy of 74%, 71%, and 72%, respectively, for

SLECR and 76.15%, 76.89%, and 75.67%, respectively, for SLDC. Table 13 outlines the precision and recall metrics for all the methods. For SLECR, MLR has the highest precision of 73% and recall of 74% respectively. It is followed by DT, KNN, NB, and SVM in the order of performance. For SLDC, DT has the highest precision of 76.60% and recall of 76.60%, respectively. It is followed by MLR, KNN, NB, and SVM in the order of performance.

Figure 11 shows the results for the transformation of categorical values using binarization and ordinal values. For SLECR and SLDC, MLR and DT outperform all other methods for binarization of categorical variable, respectively, while DT and KNN perform best when categorical variables are converted into ordinal values for SLDC and SLECR, respectively. Similar to binning methods 1 and 2, the accuracy decreases when the categorical features are converted into ordinal values for MLR, DT, and KNN. However, NB performs better when it is converted to ordinal values. For SLECR, SVM performance is not affected by the treatment of categorical features and for SLDC SVM performs slightly better when categorical features are converted to ordinal values.

Table 11 Binning method 1—model precision and recall for SLECR and SLDC

Model	SLECR		SLDC	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
MLR	82	81	85.6	85.6
DT	78	78	85.2	85.0
NB	71	68	77.8	74.8
KNN	79	79	84.4	84.4
SVM	77	76	79.6	78.6

Fig. 7 Binning method 1—performance for different categorical feature treatment

5.5 Prediction model for shipment lead time—binning method 4

Results for binning method 4 are shown in Figs. 12 and 13 and Table 14. Figure 12 shows the performance for all methods for SLECR and SLDC. Similar to the other binning methods, MLR, DT, and KNN outperform NB and SVM with the accuracy of 50%, 48%, and 48%, respectively, for SLECR, and DT, KNN, and MLR perform best with the accuracy of 59.49%, 57.64%, and 55.2%, respectively, for SLDC. Naïve Bayes again performs poorly in comparison with all the other methods for both SLECR and SLDC. As expected, the performance of all the methods for binning method 4 is inferior compared to the outcomes of other binning methods. This is because the range of each bin is too small and the number of samples in each bin is too small for the prediction tasks. Table 14 outlines the precision and recall metrics. For SLECR, MLR has the highest precision of 49% and recall of 50%. It is followed by DT, KNN, NB, and SVM in the order of precision performance. For SLDC, DT has the highest precision of 59.40% and recall of 59.40%, which is followed by KNN, MLR, NB, and SVM. For SLECR and SLDC, NB has better precision than SVM but the recall is relatively low.

Figure 13 shows the results for conversion of categorical values using binarization and ordinal values. For SLECR, DT and KNN outperform all other methods for binarization of categorical variable and when categorical variables are converted into ordinal values. For SLDC, DT performs best compared to all other methods for binarization and when categorical variables are converted into ordinal values. Similar to the other binning methods, the accuracy decreases when the categorical features are converted into ordinal values for MLR, DT,

Fig. 8 Binning method 2—performance

Table 12 Binning method 2—model precision and recall for SLECR and SLDC

Model	SLECR		SLDC	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
MLR	75	76	80.00	80.40
DT	72	72	79.80	79.80
NB	62	47	67.2	56.20
KNN	73	73	78.60	79.00
SVM	69.20	69.60	72.20	72.60

and KNN. However, NB performs better when it is converted to ordinal values. For SLECR, SVM performance is marginally reduced for binarization treatment of categorical features and for SLDC SVM performs slightly better when categorical features are converted to ordinal values.

Fig. 9 Binning method 2—performance for different categorical feature treatment

5.6 Summary of the results and discussion

This section provides the summary of results for all binning methods. For SLECR, MLR outperforms all other methods for all binning methods. For SLDC, MLR performs best for binning methods 1 and 2. However, DT provides the best results for binning methods 3 and 4. Furthermore, the effect of treating the categorical variables by transforming them into binary values and ordinal values is examined. In most cases, the accuracy is higher when the categorical values are converted into binary values than ordinal values. However, the accuracy increases for NB when categorical values are transformed into ordinal values. In addition, the performance metric gets worse as the shipment lead time range decreases as shown in Fig. 14.

The prediction model based on the classification algorithms such as MLR, DT, KNN, NB, and SVM is implemented to evaluate the accuracy of the shipment lead time (SLECR and SLDC). The response variables are grouped based on the four different binning methods with a range of 50 days, 25 days, 20 days, and 10 days. While the accuracy measures of binning methods 1, 2, and 3 are marginally different, it is also observed that the accuracy of prediction declines as the range of the bin gets smaller. That is, the higher the granularity of the range, the lower the ability to correctly predict the shipment lead time for a consignment. In particular, binning method 4 with the range of 10 days produces the accuracy of less than 50%, which is significantly lower than the outcome of other binning methods. This implies that the shipment lead time prediction range of 10 days is too small. Such poor prediction of algorithms, especially for binning method 4, can be attributed to the imbalanced proportion of the class distribution, which is the typical data characteristic in this domain. The model

Fig. 10 Binning method 3—performance

Table 13 Binning method 3—model precision and recall for SLECR and SLDC

Model	SLECR		SLDC	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
MLR	73	74	75.60	76.20
DT	71	71	76.60	76.60
NB	56	29	61.00	40.00
KNN	71	72	75.20	75.60
SVM	54	65	53.80	64.20

performance can be enhanced by use of more and adequate data for each class to learn the patterns, which is not always possible due to various circumstances. Note, however, that the accuracy metrics for SLECR and SLDC when MLR is used are reasonably good (74% for SLECR and 76% for SLDC) for binning method 3 with the range of 20 days. That is, the actual shipment lead time is within

± 10 -day time frame of the predicted shipment lead time at around 75% for both SLECR and SLDC if MLR is used. As it is identified in the real data that the maximum shipment lead time can be more than 100 days, the range of 20 days is reasonable, which is also confirmed by the industry expert. Note also that NB performance is lower compared to other algorithms. This is mainly because NB assumes conditional independence among features used to predict the outcome variable, which may not be the appropriate assumption to consider for this problem. With the given data set, SVM may not be able to generate a good decision boundary to classify the observations accurately. Retraining SVM with more data samples and additional engineered features and experimentation with more sophisticated kernel functions may positively impact the performance of learners. However, this effort can be taken up as a future enhancement to this presented research.

Furthermore, the effect of transforming categorical variables into binary values and ordinal values is examined. For SLECR, the accuracy is reduced for MLR, DT, KNN, and SVM for all binning methods when the categorical variables are converted to ordinal values. However, the accuracy of NB is increased for all binning methods except for method 1. For SLDC, the accuracy is reduced for MLR, DT, and KNN when the categorical variables are converted to ordinal values. This is due to undue importance attached to the variables with a higher numerical value. However, the accuracy measures of NB and SVM are increased for all methods except for method 1, and methods 2 and 3, respectively. It can be concluded that the commonly used classifiers such as MLR, DT, KNN, and SVM perform better when the categorical variables are treated with the binarization method and NB performs better when the categorical variables are transformed into ordinal values. The result is summarized in Table 15.

Fig. 11 Binning method 3—performance for different categorical feature treatment

Fig. 12 Binning method 4—performance

Table 14 Binning method 4—model precision and recall for SLECR and SLDC

Model	SLECR		SLDC	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
MLR	49	50	54.20	55.20
DT	48	48	59.40	59.40
NB	33	17	40.00	25.00
KNN	47	48	56.80	57.60
SVM	30.20	40.60	32.40	43.00

6 Conclusion and future work

Many logistics and supply chain companies are striving to improve visibility and predictability of shipment with the use of supply chain decision support systems and dynamic automation to create an intelligent, connected, and automated supply chain. Companies leveraging the power of data

intelligence using cutting-edge technologies such as artificial intelligence as one of the core strategies are expected to have a competitive edge, better profitability, and differentiated customer service.

This research focuses on developing a predictive model to estimate the shipment lead time starting from the shipper's first contact with the agent until the empty container is returned to the port (SLECR) and shipment delivery confirmation from a consignee at the destination (SLDC). The real industry data used in this research are obtained from the partner company, a leading logistics and supply chain company in New York, and the classification methods such as MLR, DT, KNN, NB, and SVM are used in this study. The comparative evaluation of different classifiers shows that MLR performs best in the majority of the binning methods, which is followed by DT. Note that NB, the simplest and tractable method, fails to provide reasonable accuracy for most binning methods, which can be attributed to its conditional independence between feature variables.

In summary, the prediction model proposed which can be implemented by classification methods such as MLR, DT, KNN, and SVM can significantly improve the visibility and predictability of the ocean import container shipment lead time. Such improved visibility and predictability is expected to bring more business opportunities and has great potential to provide a differentiated customer experience. It can also benefit consignees to make informed decisions and plan their inventory management according to hedge against the lost sales due to the delayed cargo arrival. In addition, shippers can use this prediction results to backtrack the date on which they need to contact the manufactures to ensure their freight is ready before a vessel departure date that ensures the arrival of shipment at the destination on the agreed date.

Fig. 13 Binning method 4—performance for different categorical feature treatment

Fig. 14 Performance for different binning methods

This research also provides potential for future study. The port-to-door shipments take a few days more than port-to-port shipments. However, models in this research are developed without differentiating such consignments due to the unavailability of shipment type details. Future accessibility of relevant information can be used to develop separate models for more precise and accurate prediction. In addition, additional features such as container custom clearance dates, trucker details (for port-to-door shipments), commodity type, customs and brokerage details, shipment types data, detention (charges applied

while the container is in customer care beyond the defined container return date), and demurrage charges (penalty applied when the container is in the steamship line custody beyond the specified container free date) can be used to enhance the predictability of the model. The inclusion of aforementioned relevant data in the implementation is expected to enhance the accuracy and performance of the prediction model. The natural extension of this paper is the development of the prediction model for other types of services such as ocean export, air import, and air export.

Table 15 Summary of results for SLECR and SLDC

Measure	SLECR	SLDC
Result	MLR > DT > KNN > SVM > NB	MLR > DT > KNN > SVM > NB

Acknowledgements This study was partially funded by Laufer Group International (Grant No. 1144442). The authors would like to thank for the financial support.

Compliance with ethical standards

Conflict of interest The authors declare no competing interests to declare.

References

- Agarwal R, Ergun Ö (2008) Ship scheduling and network design for cargo routing in liner shipping. *Transp Sci* 42(2):175–196
- Alcoba RD, Ohlund KW (2017) Predicting on-time delivery in the trucking industry. Ph.D. thesis, Massachusetts Institute of Technology, Supply Chain Management Program
- Barua L, Zou B, Zhou Y (2020) Machine learning for international freight transportation management: a comprehensive review. *Res Transp Bus Manag* 100453
- Baryannis G, Dani S, Validi S, Antoniou G (2019) Decision support systems and artificial intelligence in supply chain risk management. In: *Revisiting supply chain risk*. Springer, pp 53–71
- Bleggi CC, Zhou F (2017) A study of freight performance and carrier strategy. Ph.D. thesis, Massachusetts Institute of Technology
- Chun M (2014) Improving inbound visibility through shipment arrival modeling. Ph.D. thesis, Massachusetts Institute of Technology
- Criminisi A, Robertson D, Pauly O, Glocker B, Konukoglu E, Shotton J, Mateus D, Möller AM, Nekolla S, Navab N (2013) Anatomy detection and localization in 3d medical images. In: *Decision forests for computer vision and medical image analysis*. Springer, pp 193–209
- Fransoo JC, Lee C-Y (2008) Ocean container transport: an underestimated and critical link in global supply chain performance. Technical report, working paper, Eindhoven University of Technology
- Fransoo JC, Lee C-Y (2013) The critical role of ocean container transport in global supply chain performance. *Prod Oper Manag* 22(2):253–268
- Han J, Pei J, Kamber M (2011) *Data mining: concepts and techniques*. Elsevier, Amsterdam
- Harrington P (2012) *Machine learning in action*, vol 5. Manning Greenwich, CT
- Heydari J (2014) Lead time variation control using reliable shipment equipment: an incentive scheme for supply chain coordination. *Transp Res Part E: Logist Transp Rev* 63:44–58
- Hofmann E, Sternberg H, Chen H, Pflaum A, Prockl G (2019) Supply chain management and industry 4.0: conducting research in the digital age. *Int J Phys Distrib Logist Manag* 49(10):945–955
- James G, Witten D, Hastie T, Tibshirani R (2013) *An introduction to statistical learning*, vol 112. Springer, Berlin
- Kim YJ (2013) Analysis of truckload prices and rejection rates. Ph.D. thesis, Massachusetts Institute of Technology
- Larose DT, Larose CD (2014) *Discovering knowledge in data: an introduction to data mining*. Wiley, Hoboken
- Lee C-Y, Song D-P (2017) Ocean container transport in global supply chains: overview and research opportunities. *Transp Res Part B: Methodol* 95:442–474
- Lee HL, Billington C (1992) Managing supply chain inventory: pitfalls and opportunities. *Sloan Manag Rev* 33(3):65–73
- Lee HL, So KC, Tang CS (2000) The value of information sharing in a two-level supply chain. *Manage Sci* 46(5):626–643
- Lingitz L, Gallina V, Ansari F, Gyulai D, Pfeiffer A, Monostori L (2018) Lead time prediction using machine learning algorithms: a case study by a semiconductor manufacturer. *Procedia CIRP* 72:1051–1056
- Loh W-Y (2011) Classification and regression trees. *Wiley Interdiscip Rev Data Min Knowl Discov* 1(1):14–23
- Mangan J, Lalwani C, Lalwani CL (2016) *Global logistics and supply chain management*. Wiley, Hoboken
- Michael B, Stroh A (2006) *Practical guide to transportation and logistics*, 3rd edn. Logistics Network Inc.
- Mitchell TM et al. (1997) *Machine learning*. wcb. McGraw-Hill, Inc., New York, NY
- Pope DJ, Thomchick EA (1985) US foreign freight forwarders and NVOCCs. *Transp J* 24(3):26–36
- Prasad R, Malaj G et al (2017) Analysis of inefficiencies in shipment data handling. Ph.D. thesis, Massachusetts Institute of Technology
- Quinlan JR (1986) Induction of decision trees. *Mach Learn* 1(1):81–106
- Raschka S (2015) *Python machine learning*. Packt Publishing Ltd, Birmingham
- Seni G, Elder JF (2010) Ensemble methods in data mining: improving accuracy through combining predictions. *Synth Lect Data Min Knowl Discov* 2(1):1–126
- Sharma R, Kamble SS, Gunasekaran A, Kumar V, Kumar A (2020). A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. *Comput Oper Res* 119:104926
- Sharma S, Sharma S (1995) *Applied multivariate techniques*, 1 edn. Wiley
- Straube F, Friedrich H, Ludwig A (2019) Realization of ETA predictions for intermodal logistics networks using artificial intelligence. In: *Advances in production, logistics and traffic: proceedings of the 4th interdisciplinary conference on production logistics and traffic 2019*. Springer, p 155
- Tavallali P, Yazdi M, Khosravi MR (2017) An efficient training procedure for viola-jones face detector. In: *2017 International conference on computational science and computational intelligence (CSCI)*. IEEE, pp 828–831
- Tongzon JL (2009) Port choice and freight forwarders. *Transp Res Part E: Logist Transp Rev* 45(1):186–195
- Trevor Hastie JF, Tibshirani Robert (2009) *The elements of statistical learning*. Springer, Berlin
- Triepels R, Daniels H, Feelders A (2018) Data-driven fraud detection in international shipping. *Expert Syst Appl* 99:193–202
- Willis HH, Ortiz DS (2004) Evaluating the security of the global containerized supply chain. Rand Corporation Santa Monica CA, Technical report
- Wu P-J, Chen M-C, Tsau C-K (2017) The data-driven analytics for investigating cargo loss in logistics systems. *Int J Phys Distrib Logist Manag* 47(1):68–83
- Zaman I, Pazouki K, Norman R, Younessi S, Coleman S (2017) Challenges and opportunities of big data analytics for upcoming regulations and future transformation of the shipping industry. *Procedia Eng*. 194:537–544

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.