Vessel Turnaround Time Prediction: A Machine Learning Approach

Abstract

Uncertainty in vessel turnaround time (VTT) is troublesome and would reduce the operational efficiency in port management, potentially causing economic losses. Despite vessels generally providing their estimated departure time (EDT), there is frequently a considerable difference between the EDT and the actual departure time (ADT) of vessels due to various factors such as unexpected port handing inefficiency. This variability complicates the coordination of efficient port operations. Our research aims to address this issue by employing an extreme gradient boosting (XGBoost) regression model to predict the VTT, using vessel arrival and departure data at the Hong Kong Port for the year 2022 and the first quarter of 2023. The proposed machine learning approach can provide more accurate predictions on VTT on average compared to the EDT data reported by vessels themselves, with a substantial reduction in both mean absolute error (MAE) and root mean square error (RMSE) of 23% (from 5.1 hours to 3.9 hours) and 24% (from 8.0 hours to 6.1 hours), respectively. These results present a significant leap forward in the predictive capabilities for the VTT and lay the foundation for further research into improving vessel scheduling efficiency, reducing port congestion and enhancing overall port performance.

Keywords: Maritime transport; Vessel turnaround time prediction; Machine learning in port management; Port efficiency improvement

1 Introduction

Maritime transport is an indispensable component of global trade and the engine of globalization. According to statistics of United Nations Conference on Trade and Development, over 90% of the world's trade is carried by sea, underpinning international commerce by providing an economical and efficient mode of transport (United Nations Conference on Trade and Development, 2021; Tian et al., 2023; Liu et al., 2023; Lin, Zheng, Chu, Mao, et al., 2023). Ports serve as pivotal nodes in maritime transportation and global supply chain, facilitating the exchange of goods across nations (Wang, Liu, et al., 2022; Lin, Zheng, Chu, Zhang, et al., 2024; Feng et al., 2022). The efficiency of these ports directly impacts the flow of international trade. A key metric that measures such efficiency is the vessel turnaround time (VTT), which represents the total time a vessel spends in a port from arrival at the anchorage or berth to departure from the port. The optimization of VTT is not merely a measure of a port's operational efficiency; it is also an indicator of the port's ability

to contribute effectively to the global supply chain. A shorter VTT implies quicker movement of goods and thus shorter schedule of vessels, reducing delays and enhancing the overall supply chain productivity. Conversely, longer VTTs can lead to bottlenecks and congestion, affecting not only the specific port's operations but also having a cascading impact on global trade networks. The turnaround of a vessel encompasses multiple processes like berthing, cargo unloading and loading and all the necessary documentation procedures and the VTT has far-reaching implications on port efficiency, cost-effectiveness and competitiveness. In daily port operation, vessels typically report their estimated departure time (EDT) data before leaving a port. However, this EDT is often significantly different from the actual departure time (ADT) data due to numerous unpredictable factors, such as unexpected port operational inefficiency and congestion. These deviations between EDT and ADT introduce uncertainties and inaccuracies to the estimation of VTT. Consequently, these inaccuracies can lead to significant vessel departure delays, increased port congestion and escalated operational expenses. This not only adversely impacts the competitiveness of shipping companies and ports, but also, by extension, hampers international trade. To reduce the inaccuracies and uncertainty of VTT prediction. The improved EDT data can subsequently aid ports in executing berth allocation more efficiently. This significantly benefits the decision-making process in port operations.

We use HKP as the case study in this research to predict VTT. Recognized as one of the world's busiest ports, HKP handled approximately 18 million 20-foot equivalent units (TEUs) of containers in 2020. Typically, vessels preparing to leave the HKP report their EDT data within 36 hours prior to their departure. Moreover, the port automatically records each vessel actual time of arrival (ATA) data upon its entry and ADT data upon its exit. Using this rich dataset, we develop an extreme gradient boosting (XGBoost) regression model to precisely predict VTT at the HKP. This prediction model considers multiple factors, including historical vessel arrival and departure data, historical vessel and port operational data, vessel generic features and tidal information. Our results highlight that the predictive model significantly minimizes the deviation error in VTT. Evaluated by the MAE, the deviation error dropped from 5.1 hours to 3.9 hours, marking a reduction of 23%. In addition, when assessed by the root mean squared error (RMSE) metric, the error was cut down by 24%, from 8.0 hours to 6.06 hours. The model also achieves an impressive R-squared value of 0.804, underscoring the effectiveness of our approach in forecasting VTT data. Moreover, the feature importance analysis indicates that the EDT data and vessel generic features are vital for VTT prediction. Furthermore, we delve into a multitude of policy insights and potential expansions of our predictive model from four main aspects: effective port management, resource optimization, commercial cost reduction and green shipping promotion. These viewpoints illuminate the wider implications of our prediction model, providing valuable directions for subsequent research. To the best of our knowledge, the present study is the first to predict VTT considering EDT data, vessel generic features, historical vessel and port operational data. The specific contributions of this paper are quadruple:

1) Propose a framework for quantitatively assessing the VTT accuracy at port;

2) Develop a highly accurate tree-based model for predicting container VTT considering multiple

innovative vessel generic and port operational features;

3) Evaluate feature importance and identify feature correlation based on the machine learning model for VTT prediction;

4) Generate policy and managerial insights of port daily operation.

The organization of this paper is as follows. In Section 2, we undertake an exhaustive literature review on the analysis and prediction of the VTT. Section 3 provides a brief introduction to the background of the HKP and describes the vessel arrival and departure dataset that utilized for our analysis. In addition, this section includes a comprehensive statistical analysis of VTT data at the HKP. Section 4 introduces and constructs an XGBoost model for predicting VTT at the HKP. This section also includes an analysis of the prediction results, discusses research limitations. In Section 5, we delve into the wider implications of our VTT prediction results, encompassing potential extensions of our model and policy insights drawn from our results. The paper concludes in Section 6 with a discussion of the primary findings and their broader impacts.

2 Literature review

VTT is a crucial port management performance measurement. A vessel's VTT contains various parts such as the duration of berthing, waiting and servicing time of the vessel and thus reflects the proficiency and efficiency of a port's overall operation. Excessive lengths and uncertainties in VTT could lead to port congestion and untimely scheduling, potentially resulting in delays in ship schedules. This, in turn, can increase vessels' operational costs and thus diminish the port's reputation, ultimately leading to considerable economic losses. Minimizing the VTT contributes to the port's capacity to handle more ships within a specific period, thereby enhancing the overall efficiency of the maritime chain. Historically, efforts to optimize VTT in port operations have predominantly focused on strategies for berth scheduling. Such a focus is well-documented in the field of maritime studies, as reflected by the substantial attention it receives in the existing literature (Kim and Moon, 2003; Xu, Chen, and Quan, 2012; Golias et al., 2009). Despite this established focus, it is worth noting that in the port industry, one of the world's oldest and most consistent sectors, operations related to berth scheduling and VTT optimization often rely more heavily on expert knowledge and the established berth allocation strategies rather than innovative and data-driven scheduling strategies (Barua, Zou, and Zhou, 2020; Brouer, Karsten, and Pisinger, 2016; Rodrigues and Agra, 2022; Politikos et al., 2023). Recent innovations in machine learning field have paved the way for developing data-driven methods in port operations management (Yan, Wang, Zhen, and Laporte, 2021; Filom, Amiri, and Razavi, 2022; Duan et al., 2022; Xu, Chen, Wu, et al., 2022; Ma et al., 2023; Liu et al., 2023). These approach show considerable potential of reducing prediction inaccuracies regarding vessel arrival and departure time from a port management perspective.

Numerous studies have addressed the issues of uncertainties in the time associated with vessels' activities in ports, aiming to predict time-related factors such as vessel arrival time, turnaround time and departure time at the port (Yan, Wang, Zhen, and Laporte, 2021; Filom, Amiri, and

Razavi, 2022; Wang and Yan, 2023). These predictions can greatly assist port operators in their decision-making for port management. For instance, by accurately predicting vessel arrival time, the following berth allocation strategy can be optimized, which in turn enhances the operational efficiency of the port (Yu et al., 2018). Compared to the extensive literature on predicting ship estimated time of arrival (ETA) data (Yu et al., 2018; Yan, Wang, Zhen, and Laporte, 2021; Filom, Amiri, and Razavi, 2022; Chu, Yan, and Wang, 2023), there are currently only a small number of studies on predicting ship EDT or VST using machine learning Filom, Amiri, and Razavi, 2022; Yan, Wang, Zhen, and Laporte, 2021. This discrepancy can be attributed to several reasons. First, the complexity inherent in VTT prediction caused by factors like the vessel's uncertain arrival time, unexpected port congestion and inefficiency in berth allocation that have an influence on VTT makes its precise predictions challenging. Secondly, due to privacy and security concerns of port data, acquiring detailed port operational data, such as real-time quay crane efficiency or berth occupancy rates, is difficult. The limited access to such crucial data impedes the ability to make accurate predictions on VTT.

Mokhtar and Shah (2006) are among the first to conduct research in the field of VTT evaluation. They employ a linear regression model to analyze the relationship between VTT and port facilities. To be more specific, the study concentrates on two ports in Klang, Malaysia, utilizing vessel call data and port operational data collected in August 2005. The results reveal a strong correlation between VTT, crane allocation and the number of containers loaded and discharged (Mokhtar and Shah, 2006). Ducruet and Merk present an overview of the VTT efficiency in world container ports in 1996, 2006, and 2011 (Ducruet and Merk, 2013). Their study indicates that Ningbo port in China is the most efficient among the largest ports during the years. Štepec et al. (2020) introduce a machine learning based VTT prediction system, utilizing 11 years of port call data at the Port of Bordeaux. This system shows excellent performance in predicting VTT. When compared to actual VTT, the system's error rates for specific types of cargo vessels dropped below 10% (Štepec et al., 2020). Li and He (2020) employ a deep neural network to predict container liner berthing time at a terminal in China using four years of container vessels berth time data. However, the advantage of the proposed model's berth time prediction results over the container berthing time reported by the vessels has not been clearly demonstrated (Li and He, 2020).

Smith (2021) analyzes factors affecting container vessels' VTT at major U.S. container ports, using vessel automatic identification system (AIS) data and port TEU volumes through statistical and comparative analysis. Results of this study show a remarkably positive correlation between the VTT and the vessel expected cargo volume per call. The research results also indicate that vessels are expected to follow the arrival and departure schedules, which make vessel schedules the primary determinants of the VTT (Smith, 2021). Abreu et al. (2023) evaluate the primary factors affecting the VTT at the port. Then they develop several classification models using data from the cargo vessel movement data in Brazilian ports in 2018 to predict the VTT. The results show that random forest (RF) demonstrate promising performance, with accuracy and F1-scores above 73%. The results also suggest that RF has the potential to be applied in real-world port management scenarios (Abreu et al., 2023). Zhai et al. (2022) utilize cargo operation data from tanker terminals and vessel AIS data to predict the VTT at Singapore tanker terminals. By employing linear regression

and decomposed distribution methods, the prediction model demonstrates remarkable accuracy at 98.81% when evaluated by vessel historical VTT (Zhai et al., 2022).

Based on a comprehensive review of prior studies, it is evident that most of the current research is focused on analyzing factors affecting VTT. The studies related to VTT prediction are predominantly based on classification analysis. However, these classification models predict VTT results for vessels in the form of time intervals, which are less precise compared to predictions provided by regression models that specify a specific time-stamp. There is only one paper that has successfully employed regression machine learning models to accurately predict VTT (Štepec et al., 2020). In that research, the authors proposed a tree-based VTT prediction system based on port call data, vessel generic features and port tidal data. The proposed system can greatly reduce error in VTT by at least 10% for various types of vessels compared to the original EDT data reported by vessels. The innovative aspects of the paper are primarily focused on the presentation of the first machine learning-based regression system for VTT prediction. Additionally, the author is the first to evaluate the impact of port call data, vessel generic features and tidal information on ships' VTT prediction. However, the authors of this article acknowledge that the model they proposed is validated by data from a relatively small port. They do not account for larger ports with more complex infrastructure and higher traffic volumes, where factors such as traffic congestion and vessel arrival patterns could significantly influence the VTT. Remarkably, even within this paper, current research related to VTT prediction has not taken into account the EDT data reported by vessels in predicting their VTT. Furthermore, there is a noticeable lack of attention paid to the influence of vessel generic features (such as vessel length and beam), vessels' historical performance at port (e.g., historical average VTT) and berths' historical operation performance (e.g., historical delays of the ships operated at the berth in terms of their predicted turnaround time) when these factors having a tangible influence on VTT. Given the existing research gaps, this study develop regression models to predict container vessels' VTT at the HKP, which is one of the busiest ports in the world. Moreover, more comprehensive data are included to construct the prediction model, including vessel generic features, historical delay data of both vessels and berths and other relevant factors. Therefore, it can be expected that our research can generate more precise predictions on the VTT in a regression way and enhance the understanding of the intricate factors affecting vessel turnaround time.

3 Introduce and analyze vessel turnaround data at the HKP

In this section, we first present an overview of the HKP background and the vessel arrival and departure data. Following this, we gather vessel arrival and departure data from the entire year of 2022 and the first three months of 2023. We then carry out data pre-processing and analysis, with the aim of quantitatively evaluating the VTT at the HKP. We analyze a total of 14,975 paired vessels arrival and departure records at the HKP in terms of the types of vessels arriving at the port, the vessels' expected VTT, the vessels' actual VTT and the delays in VTT at the HKP.

3.1 Introduction of the HKP

The HKP, situated in the South China Sea, primarily caters to the needs of containerized manufactured goods. Globally recognized for its efficiency, it ranks among the top international container ports over the world (Maritime and Board, 2022). In 2021, the port managed nearly 18 million TEUs, making it the tenth largest container port worldwide (Maritime and Board, 2022). As of June 2022, HKP provided service to around 270 international container liners weekly, linking over 600 global destinations (Maritime and Board, 2022). The primary container terminals (CTs) of the HKP are operated by five operators and provide 24 berths situated in the Kwai Chung-Tsing Yi basin with precise geographical coordinates at 22.328°N, 114.122°E. An overview of the HKP is shown in Figure 1, where the top-left corner, highlighted in red, represents the Kwai Chung-Tsing Yi basin, for which we have provided specific coordinates.[1]

Figure 1: An overview of the location of the HKP

3.2 Dataset description

The Hong Kong Marine Department updates the arrival and departure information for ocean-going vessels daily on its website (Hong Kong Marine Department, 2022) that is publicly accessible. The website contains five files: vessels arrived in the last 36 hours, vessels due to arrive in the next 36 hours, in port vessels, vessels departed in the last 36 hours and vessels intend to depart the HKP in the next 36 hours. The available data, their explanations and the frequency of data updating are

[1]The figure is generated from Google Maps (https://www.google.com/maps/).

shown in Table 1 and Table 2.

Table 1: Variables and their descriptions

Variable	Description	Note
Vessel name	Name of the vessel	\
Ship type	Type of vessel	14 types in total
Trip status	Vessel trip status	Approved or pending for a vessel entering Hong Kong Waters
Agent name	Name of the vessel's agent	\
Flag	Vessel registration country	\
ETA	Vessel estimated arrival time to the HKP	Provided by the vessel operator
ATA	Vessel actual arrival time to the HKP	Recorded when a vessel arrives at the berth or anchorage area
EDT	Vessel estimated departure time from the HKP	Provided by the vessel operator
ADT	Vessel actual departure time from the HKP	Recorded when a vessel departs from the Berth in HKP
Upload time	Data upload time onto the website	Provided by the Marine Department
Next port	The next port to visit after leaving the HKP	Provided by the vessel operator
Last port	Name of a ship's last port of call before arrival	Provided by the vessel operator
IMO number	The International Maritime Organization (IMO) number of a ship	A unique vessel identifier comprising seven digits
Call sign	An alphanumeric code that uniquely identifies a vessel for radio communication	A unique vessel identifier
Berth	A vessel's current berthing location	\
Arrived location	The first location where a vessel stays after arriving at the Hong Kong waters	\

Table 2: Variables in the vessel traffic management system oeprated by the HKP

	Ships arrived in the last 36 hours	Ships due to arrive in the next 36 hours	Vessels departed in the last 36 hours	In port vessels	Intend to depart in the next 36 hours
Update frequency	Daily	Daily	Daily	20 minutes	Daily
Vessel name	✓	✓	✓	✓	✓
Ship type	✓	✓	✓	✓	
Trip status		✓			
Agent name		✓		✓	✓
Flag	✓	✓			
ETA		✓			
ATA	✓			✓	
EDT					✓
ADT			✓		
Upload time	✓	✓	✓	✓	✓
Next port			✓		✓
Last port	✓	✓			
IMO number			✓		
Call sign	✓	✓	✓	✓	✓
Berth	✓		✓	✓	✓

For the subsequent quantitative analysis and prediction of VTT, we have collected all related data from January 1, 2022 to March 31, 2023 for our analysis. The ATA, EDT and ADT datasets contain 28,126, 18,365 and 19,430 records, respectively. As indicated in the data description above, the ATA, EDT and ADT data records reported by a vessel are stored in different files. The expected VTT for a port call is the difference between the EDT and ATA for that particular voyage, while the actual VTT is the difference between the ADT and ATA. Therefore, in order to quantify the

specific duration of the VTT, further data pre-processing is required.

3.3 Data collecting and pre-processing

Since the calculation of both the actual and expected VTT involves the ATA, we utilize the ATA data as the reference during the data matching process. That is, we assign the corresponding ADT and EDT data to each ATA record. The basic steps for data pre-processing in VTT analysis are as follows:

1) Obtain ATA, EDT and ADT data:
   Hong Kong Marine Department uploads five files containing the arrival and departure information of ocean-going vessels on the government website on a daily basis (Hong Kong Marine Department, 2022). To pair a vessels' ATA with ADT data, as well as ATA with EDT for further analysis, the first step is to gather these ATA, EDT and ADT data from separate folders.

2) Unify the time format in the datasets:
   The time formats in different datasets may vary. The time formats for ATA, EDT, ADT and upload time are standardized to "Year-Month-Day Hour:Minute:Second". All records with missing time information are removed from the dataset.

3) Delete records with an ATA later than its upload time:
   As the system records and uploads a vessel's ATA after it arrives at the HKP, the upload time of the ATA data will always be later than the ATA itself. If the ATA is earlier than the upload time, it can be considered inaccurate caused by system recording errors. Utilizing such erroneous ATA data as the ground truth value could adversely affect the subsequent vessel arrival analyses and predictions. Therefore, such data are removed from the ATA dataset.

4) Drop duplicated data from the datasets:
   The port system consistently updates vessel arrival and departure data on a daily basis, ensuring coverage for the upcoming 36 hours. For each ship arrival and departure, the corresponding records of ATA, EDT and ADT are created. However, after merging the data, we noticed that a single ship's arrival might correspond to multiple reported ATA, EDT, or ADT records, each with a different upload time[2]. For these duplicate records, we only retain the first record of ATA and ADT data and delete the subsequent ones. Regarding the EDT data, when we encounter multiple EDT data records reported by the same vessel, we choose to keep all unique EDT records. These records are valuable as they reflect changes in the ship's departure schedule, providing critical information for our predictive analyses. However, if an EDT record remains consistent across multiple reports and only the upload time of EDT data changes, we eliminate the duplicates, keeping only the initial instance of the duplicated EDT. This approach is based on our assessment that the upload time of the EDT does not affect in our prediction models, and keeping duplicates would just make our training dataset bigger without helping our models.

[2]For ATA and ADT records that are unique for one port call of a ship, each record may be included in the updated files for different days. For EDT data, a ship may report different EDT during the visit to a port and all the reports are included in the files.

5) Match the ATA and ADT data of each vessel in chronological order:
During a voyage, a vessel's ADT is always later than its ATA. To match a vessel's ATA and ADT data, we first filter the relevant ATA and ADT data for each ship based on the call sign from the datasets. Next, we sort a vessel's ATA records in chronological order. For each ATA record, if the ATA precedes the ADT and the time difference between them is less than 3 days,[3] we consider them as a pair. After pairing, the ATA and ADT data are removed from the dataset for the next round of matching.

6) Match the ATA and EDT data of each vessel in chronological order:
This step is similar to the previous ADT matching step. We first use call sign to sort the ATA data and ADT data for the targeted vessel, but what we compare in the next step are the ATA data and the upload time of the EDT data. Specifically, for each ATA record, if the upload time of the EDT data plus one day is earlier than the ATA data and the time difference between the upload time of the EDT and ATA data is less than 5 days,[4] we consider them as a pair. After pairing, the ATA and ADT data are removed from the dataset for the next round of matching.

7) Delete ATA records without the corresponding ADT and EDT
Upon completing the pairing process, any ATA records without the corresponding ADT and EDT are considered erroneous and are deleted. This circumstance might occur because either the system records the vessel's ATA but fails to record the corresponding EDT or ADT, or the vessel remains docked at the port and thus does don the corresponding ADT and EDT data.

The number of ATA, EDT and ADT records before and after matching is summarized in Table 3.

Table 3: Summary of the number of data records in the data pre-processing scheme

Step	Method	ATA	EDT	ADT
1)	Data collection	21,126[5]	18,231	19,430
2)	Time format unification	20,001	17,993	19,179
3)	Abnormal ATA data deletion	18,528	17,993	19,179
4)	Duplicated data elimination	17,201	17,825	19,003
5)	Pair ADT to its corresponding ATA	16,216	17,825	16,216
6)	Pair EDT to its corresponding ATA	14,975	14,975	16,216
7)	Remove unmatched data	14,975	14,975	14,975

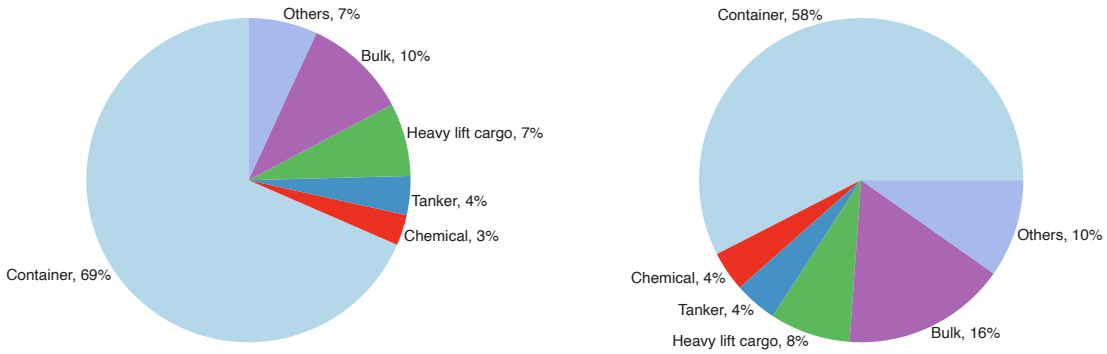
[3]We set a 3-day time difference limit to account for rare cases where the ATA is recorded upon the vessel's arrival at the port, but the port does not record the ADT when the ship departs. Without this 3-day limit, the ADT record of the ship's next voyage might be erroneously matched with the ATA record of the current voyage, which could adversely affect the subsequent data matching processes.

[4]We compare the ATA time with the upload time of the EDT data plus one day. Because EDT data is typically reported within 36 hours prior to vessel departure, there are rare cases where a vessel reports its EDT before it arrives at the port, so directly comparing ATA and EDT could result in incorrect matching for these vessels. The upload time of the EDT data is later than the EDT and we add one day to the upload time to compare with the ATA to prevent matching errors.

[5]The reason why there are significantly more ATA data records than those of ADT and EDT is that there are changes in the ATA_berth information in the ATA dataset, while the other information remains the same. Consequently, the port might record multiple ATA data records with one corresponding ADT data record during one ship visit.

3.4 Quantitative evaluation of vessel arrival data at the HKP

In this subsection, we present a quantitative evaluation of vessel arrival data at the HKP from January 2022 to March 2023. This analysis encompasses a comprehensive assessment on the types of arriving vessels, their expected VTT, their actual VTT and their delay in VTT. There are a total of 14 types of vessels that have arrived at the HKP (Maritime and Board, 2022), including container vessel, bulk vessel, heavy lift cargo vessel, oil tanker, liquefied natural gas (LNG) carrier, multi-purpose vessel, passenger vessel, reefer carrier, fishing vessel, tug, chemical tanker, nuclear fuel tanker, car carrier and other types of vessels. Because the top five ship types constitute around 93% of all of the visiting ships, we only keep the top five vessel types (i.e., container vessel, bulk vessel, heavy lift cargo vessel, oil tanker and chemical vessel) and classify the remaining nine vessel types into the "others" category. The analysis results are shown in Figure 2.



(a) Arrived vessel types analysis at the HKP in 2022 (b) Arrived vessel types analysis at the HKP in 2023

Figure 2: Analysis of the types of vessels arriving at the HKP

Figure 2 indicates that the predominant type of vessel arrivals at HKP is container vessel. This validates the fact that HKP primarily functions as a container port. Considering that the HKP is predominantly a container port and the main aspects of port operations such as berth allocation primarily involve container vessels, we therefore conduct the following quantitative analysis of VTT for all types of ships, with a specific focus on container vessels. The statistical analysis of actual VTT (in hours) are displayed in Table 4, where the actual VTT is defined as the difference between the ADT and the ATA for each vessel.

Table 4: Statistic analysis of the actual VTT at the HKP

Item	Number of records	Mean (hours)	Median (hours)	Maximum (hours)	Minimum (hours)	Standard deviation (hours)
All types of vessels in the dataset	14,975	18.20	14.63	63.61	0.20	18.59
All types of vessels in 2022	11,938	18.80	15.16	63.61	0.21	18.91
All types of vessels in 2023	3,037	16.83	13.82	56.65	0.11	17.33
Total container vessels in the dataset	9,998	17.66	13.88	49.12	1.21	14.52
Container vessels in 2022	8,178	18.29	13.90	49.12	1.39	14.96
Container vessels in 2023	1,820	15.89	12.30	43.28	1.21	13.59

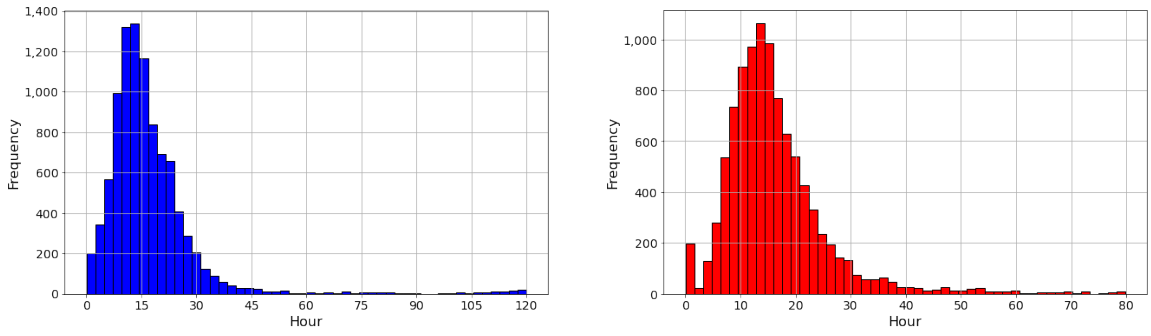
In Table 4, we find that the mean, median and variance values of actual VTT for container vessels are lower than those of all other types of vessels. This may be attributed to the fact that operations

involving container vessels at the port are primarily automated, resulting in fewer disruptions from human factors compared to other types of vessels. Additionally, we find that the maximum value of the actual VTT of all types of vessels is significantly larger than that of container vessels. The reason is that these maximum values in VTT correspond to the arrival operations of cement vessels at the HKP. Next, we evaluate the expected VTT at the HKP. The expected VTT for each vessel is defined as the difference between its EDT and its ATA. Similar to the analysis of actual VTT, we have categorized the data into two groups: all types of vessels and the container vessels. The vessel expected VTT evaluation results in hours are presented in Table 5.

Table 5: Statistic analysis of the expected VTT at the HKP

Item[6]	Total number of records	Mean (hours)	Median (hours)	Maximum (hours)	Minimum (hours)	Standard deviation (hours)
Total_vessel	14,975	17.43	13.65	119.98	−17.94	17.62
2022_vessel	11,938	17.85	14.17	119.98	−16.61	17.79
2023_vessel	3,037	15.82	11.95	156.65	−17.94	16.81
Total_container	9,998	16.34	14.63	69.95	−17.94	14.27
2022_container	8,178	14.68	14.68	69.20	−16.61	14.33
2023_container	1,820	14.48	12.08	69.95	−17.94	13.83

In Table 5, we encounter an anomaly where the minimum value of the expected VTT is less than zero. The expected VTT is calculated as EDT minus ATA, and the negative value anomaly arises from vessels that report their EDT data within 36 hours before departing from the port. However, according to our previous analysis shown in Table 4, the average actual VTT at HKP is approximately 18 hours. In some rare instances, vessels report their EDT before they arrive at the port. Nonetheless, the reported EDT data often contain significant errors and might be earlier than the ATA at the port. This discrepancy results in a negative value by deducting EDT by ATA, which, in turn, leads to a negative expected VTT value in these scenarios. These instances could potentially be outliers in the dataset, which may skew prediction outcomes. To avoid this, we will exclude these outliers during the visualization and prediction phases. Figure 3 illustrates the distribution of both the expected and the actual VTT for container vessels at the HKP.



(a) Distribution of the expected VTT of container vessels at the HKP

(b) Distribution of the actual VTT of container vessels at the HKP

Figure 3: Visualization of the expected and actual VTT of container vessels at the HKP

[6]Similar to Table 4, in Table 5, we conduct a quantitative analysis of the expected VTT for all vessels and container vessels, categorized by year.

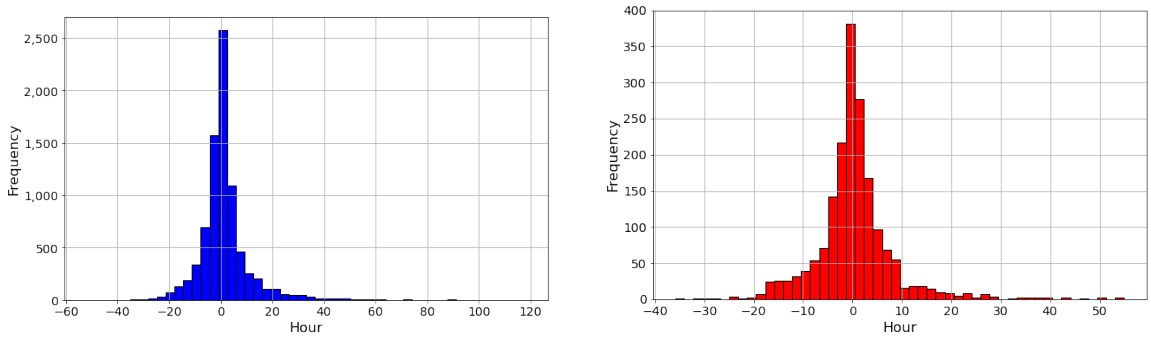
In Figure 3, we observe that the distributions of both the expected and actual VTT for container vessels at the HKP roughly follow a Gaussian distribution, which is consistent with the results in Table 4 and 5, indicating that the peak of the expected VTT distribution is approximately 14 hours and the peak of the actual VTT is around 15 hours.

Next, we give an analysis of vessel departure delay at the HKP, which is evaluated by the difference between EDT and ADT. A negative value of vessel delay shows that a vessel departs latter than expected, while a positive value suggests that the vessel departs earlier than expected. The value of 0 indicates that the vessel departed on time. There are 69 data records categorized as "on-time". The vessel departure delay analysis results (in hours) for the other two classes are summarized in Table 6.

Table 6: VTT delay analysis in hours

Item[7]	Total number of records	Minimum (hours)	Maximum (hours)	Median (hours)	Mean (hours)	Standard deviation (hours)
Early	7,637	0.02	118.25	3.45	7.44	6.02
Late	7,269	0.10	58.17	3.48	5.58	11.04
Early_container	5,261	0.16	52.02	3.27	4.93	5.40
Late_container	4,685	0.11	76.25	3.25	6.90	10.63

Finally, we visualize the annual VTT delay distribution of container vessels at the HKP, as depicted in Figure 4.



(a) Distribution of container VTT delay at the HKP in 2022

(b) Distribution of container VTT delay at the HKP in 2023

Figure 4: Visualization of the container VTT delay at the HKP

In Figures 4a and 4b, we observe that the container VTT delay data for the years 2022 and 2023 roughly follow a Gaussian distribution centered around zero. This is in line with the requirements for pre-processing the prediction set. Furthermore, we find that the VTT delay distribution for 2022 in Figure 4a aligns more closely with a Gaussian distribution compared to that of 2023 in Figure 4b. This can be attributed to the 2022 dataset encompassing data from an entire year, thus being larger, while the 2023 distribution only contains data from the first three months. In addition, both 2022

[7]In Table 6, the "early" column represents all types of vessels that their ADT is earlier than the corresponding EDT data. The "Late" column stands for all vessels that their ADT is later than the corresponding EDT data. "Early_Container" denotes container vessels that leave the port earlier than the EDT, while "Late_Container' signifies container vessels that depart from the port later than the EDT.

and 2023's VTT delays for container vessels exhibit several instances of extreme delays, which are responsible for the outlier points observed in the distribution figures.

4 Prediction of VTT at the HKP

This section is dedicated to developing a data-driven machine learning regression model to predict the VTT at the HKP. This section comprises five primary components: feature engineering, VTT prediction, the analysis of the prediction results, limitation of the research and further study.

4.1 Feature engineering

Feature engineering is the process of leveraging domain knowledge to create more effective representations that capture the essence of a problem, with the goal of enhancing a model's predictive capabilities. There are three key steps involved in our feature engineering process, namely feature selection, categorical feature encoding, and dataset splitting. "Feature selection" aims to identify the most relevant variables for prediction. "Categorical feature encoding" transforms categorical data into numerical format so that models can process them. Finally, "dataset splitting" divides the data into separate sets for training and testing, ensuring that the model can be trained and validated effectively. Each of these steps is explained in detail below.

4.1.1 Feature selection

To predict VTT, we assess and integrate five categories of features into vessel arrival dataset to establish a comprehensive training set: vessel generic features, temporal features, berth operational features, berth generic features and waterway characteristics. The actual VTT is designated as the ground truth label. The features used to construct the machine learning prediction model are detailed in Table 7. We have evaluated these features based on their demonstrated influence on VTT through factor analysis, ensuring each one contributes meaningfully to the predictive accuracy of our model. Below, we provide a rationale for including these particular features from the perspective of their influence on VTT and detail the specific parameters selected for our case study, along with the source data from which they are obtained, ensuring a comprehensive and data-driven approach to predicting VTT.

4.1.1.1 Vessel generic features

The importance of vessel generic features to VTT is a critical aspect in port management (Chu, Yan, and Wang, 2023). Generic features of a vessel, such as gross tonnage (GT) and length, directly influence how quickly it can be processed at a port. For instance, GT reflect a vessel's carrying capacity. Vessels with larger GT generally require more time for loading and unloading cargo compared to those with smaller GT. Moreover, ports may allocate specific berths based on the length and beam of the vessel, potentially influencing the VTT. Efficient evaluation of these generic features can lead to reduced waiting times, quicker cargo handling, and ultimately, a faster turnaround time. This is vital for maintaining the competitiveness and profitability of port operations as well

as the efficiency of the VTT prediction. In our study, to enrich our dataset by incorporating more ship characteristics, we integrate data from two external sources: the World Register of Ships (WRS) (WRS, 2020) and the MarineTraffic website (MarineTraffic, 2023). The WRS database, which features over 100,000 vessels, provides vessel feature information such as ship type, vessel length, IMO number and call sign. MarineTraffic, recognized as the world's leading provider of shipping tracking and maritime intelligence, offers data on vessel generic features, including length, beam and GT (MarineTraffic, 2023). By using the vessel IMO number as a common identifier, we collate vessel beam, GT and length information from these two databases and combine them with the vessel arrival dataset. It should be noted that two features, the type of the vessel and max_draft: the longest distance between the surface of water and the lowest point of vessel, can be directly extracted from the ATA data.

4.1.2 Temporal features

Temporal features like EDT and ATA significantly impact VTT. EDTs help in resource planning and allocation, ensuring efficient berth utilization. Therefore, deviations from these estimates can cause delays, affecting subsequent vessel schedules and increasing turnaround times. ATAs, indicating the vessel's actual arrival, can disrupt operations if significantly different from expected times, leading to rushed or delayed operations and inefficient use of resources. Accurate EDT and ATA are vital for effective scheduling and coordination, allowing for predictive planning and thus better management of vessel flows. Expected VTT can be calculated from ATA and ETD and serves as an essential basis for VTT evaluation and prediction. Moreover, port operational efficiency and busyness vary across different weeks, time slots, and seasons. Temporal features in ETD and ATA can serve as a basis to better assess VTT, accommodating for these variations and improving the overall efficiency and predictability of port operations. In the case study, the vessel arrival and departure information provided by the Hong Kong Marine Department website includes these time-related temporal information (e.g., ETA, ATA and their upload time) for ships in port.

4.1.2.1 Berth operational features

Berth operational features play a crucial role in determining vessel departure time. For instance, the number of vessels in the port in a real-time manner serves as an indicator of the port's ongoing operational intensity. When there are an excessive number of vessels but only a limited number of berths are available, vessels are forced to wait in holding areas like anchorages, thereby prolonging their turnaround time. On the bright side, berths with a history of low latency time indicate highly efficient loading and unloading operations at the berth. Accessing these berths can decrease the VTT, permitting more vessels to utilize the berth within a certain time frame. In contrast, berths with a track record of delays can cause vessels to spend more time in port than anticipated, subsequently leading to higher VTT. We consider three indicators in this part: the number of vessels in port, the average vessel delay and the average berth delay. The number of vessels in port refers to the number of ships in port at the ATA time. The average vessel delay refers to the average VTT delay for the particular vessel in the dataset, which is updated each time when the vessel arrives at the port. The average berth delay is divided into ATA berth delay and EDT berth delay. ATA

berth delay refers to the average VTT delay time for the berth at which the vessel arrives. The EDT berth delay refers to the average VTT delay time for the berth at which the vessel reports its EDT. The specific method for calculating the delay is defined as follows: if a vessel is associated with multiple arrival reports in the dataset, the delay data related to the first arrival is the absolute difference between the vessel ADT and the corresponding EDT. For the second arrival, the delay value is the average of the VTT delay for the first two arrivals. This pattern continues, with each subsequent delay value being calculated as the average delay of all previous arrivals.

Besides, an increasing number of ports have been establishing cooperative relationships with shipping companies. For instance, COSCO Shipping has recently acquired stakes in the Port of Rotterdam and the Port of Hamburg (Rinke and Schwartz, 2022). Through such cooperative agreements, the shipping company can enjoy privileges like privileged berthing rights and access to exclusive berths (Lun et al., 2023). However, the specific agreement and implementations of these privileges are often guarded as part of confidential port development and planning strategies. Such details are generally not available on public websites or through open-source channels. Moreover, these cooperative agreements significantly impact the evaluation and prediction of vessel arrivals at ports. To address this issue, our research has innovatively incorporated "berth operational features" into our VTT prediction model. These features, including historical vessel delay and berth ATA/EDT delay, provide valuable insights into the dynamics between shipping companies and terminal operators when evaluating VTT. For example, vessels with privileged berthing rights often bypass waiting at anchorage and proceed directly to their berths for port operations, resulting in shorter VTT delays compared to vessels without such agreements. Similarly, vessels docking at exclusive berths designated under cooperative agreements tend to experience lower VTT berth delays compared to those at regular berths. Ultimately, our prediction model, by incorporating these parameters, reflects the potential impact of cooperative agreements on vessel arrival predictions, shedding light on a crucial aspect of port operations that has been challenging to quantify due to limited data availability.

4.1.2.2 Berth generic features

Berth generic features play a significant role in vessel arrival and berth allocation in the realm of port operations. Features such as berth length and depth directly influence the processes of docking and undocking of vessels, impacting the VTT significantly. A longer berth can accommodate larger ships, which may take more time to dock and undock, and is capable of carrying more cargo, thereby increasing the average cargo handling time at the berth. Similarly, the availability of deeper berths facilitates the entry of larger vessels, which can enhance port operational speeds and necessities. These factors, in turn, contribute to the overall efficiency of port operations, impacting the final turnaround time of the vessels. Inefficiency or inappropriacy in berth allocation, arising from not adequately considering the critical influence of berth length, depth, and other features, could lead to significant operational setbacks. For example, if the berth length is not fully utilized in scheduling, it could result in the inability to accommodate larger vessels or multiple smaller ones simultaneously, leading to increased waiting times and reduced throughput. The Hong Kong Marine Department, in its specific HKP berth guide, provides a detailed description of the generic features of each berth (Hong Kong Government, 2023). In this study, we incorporate berth length, berth

length overall and berth draft into our dataset, expanding the generic information related to berth characteristics.

4.1.2.3 Waterway feature

Port waterway features encompass a range of characteristics that directly influence VTT). These features include the depth and width of navigational channels, tidal level, amongst others. The depth and width of the berth channel determine the maximum size of vessels that can safely navigate and dock at the port, known as the port's draught and beam restrictions which has been covered in the berth generic feature sub-section. Tidal level refers to the height of a river's water surface relative to a predetermined baseline and it exhibits significant seasonal fluctuations. This factor is crucial in maritime transportation, affecting both the navigational channel and the design velocities of vessels. A decrease in tidal level can narrow and reduce the depth of the port navigational channel, meaning larger vessels may have to wait for higher tides to enter or leave the berth, or they may be forced to offload some cargo to reduce draft, leading to delays and increased VTT. Smaller navigable port channels also increase the risk of accidents or collisions, necessitating slower speeds and more cautious maneuvering. This, in turn, compromises the efficiency of port operations and increases the VTT for ships, affecting the overall throughput and service quality of the port. The Hong Kong Marine Department provides updates on HKP tidal level information every 10 minutes (Marine Department of Hong Kong, 2023). In this study, we incorporate the value of tidal level (in metre) at the HKP into our dataset based on ATA time-stamp.

Table 7: Summary of features

Feature type	Detailed feature	Feature description	Source
Vessel generic features	Beam	Beam of vessel	WRS
	GT	GT of vessel	WRS
	Length	Length of the vessel	WRS
	Max_draft	The longest distance between the surface of water and the lowest point of vessel	ATA
	Vessel_type	Type of vessel	ATA
Temporal features	ATA_day	Week day of ATA	ATA
	ATA_hour	Hour shift of ATA	ATA
	EDT_day	Week day of EDT	EDT
	EDT_hour	Hour shift of EDT	EDT
	EDT_season	Season shift of EDT	EDT
	ATA_season	Season shift of ATA	ATA
	Expected VTT	Vessel EDT minus vessel ATA	EDT and ATA
Berth operational features	ATA_inport	The number of vessels in port at ATA time	In port
	ATA_berth_delay	The average historical delay of the berth at which the vessel arrives	ATA
	EDT_berth_delay	The average historical delay of the berth at which the vessel reports the EDT	EDT
	Vessel_delay	The average delay in historical VTT data of the vessel concerned	
Berth generic features	ATA_berth max draft	The longest distance between the surface of water and the lowest point of vessel when vessel arrive at the berth	Berth guide
	ATA_berth max length overall	The maximum length overall of the berth that the vessel arrives at.	Berth guide
	ATA_berth length	The length of the vessel arrival berth.	Berth guide
	EDT_berth max draft	The maximum draft of the berth where the vessel is located when reporting the EDT	Berth guide
	EDT_berth max length overall	The maximum length overall of the berth where the vessel is located when reporting the EDT	Berth guide
	EDT_berth max length	The length of the berth where the vessel is located when reporting the EDT.	Berth guide
Waterway features	Tidal level	The tidal level value corresponding to the ATA time.	Government website
Prediction target	Actual VTT	Vessel ADT minus vessel ATA	ADT and ATA

4.1.3 Categorical feature encoding

Feature encoding of categorical data is a critical component of the data engineering pipeline (Zheng and Casari, 2018). Categorical data are non-numeric and often subdivided into groups. Consider, for example, the category of "vessel type". This is a form of categorical data that includes categories such as "container vessel", "bulk vessel" and "chemical tanker vessels". These values are maintained in a string format. Such features cannot be directly processed by machine learning algorithms. In this way, categorical feature encoding is employed to convert these strings into numbers suitable for input into machine learning models. Given the nature of the categorical data involved in this

study, we primarily employ the following two feature encoding methods, one hot encoding and label encoding:

1) One hot encoding: For a feature with m categories without order between them, after One hot encoding processing, that feature is extended to m new binary features and the original feature is deleted, with each new feature corresponding to a category. The m binary features are mutually exclusive and only one of them is set to 1 considering the real feature value, with 0 given to all of the $(m-1)$ features (Zheng and Casari, 2018).

2) Label encoding: In label encoding, we assign labels according to a hierarchy. For a feature with m categories, each category is mapped to a number between 0 and $m-1$ after label encoding. The larger the assigned value, the higher the hierarchy of the category (Zheng and Casari, 2018).

Timestamps such as ATA, EDT and upload_time cannot be directly inputted into the prediction model. To tackle the issue, First, we employ One hot encoding to categorize the specific days of the week for ATA and EDT. Following this, we divide the precise time of ATA and EDT into three distinct and complimentary shifts using label encoding. These shifts are classified as: Shift 1 (from 0:00 to 8:00), Shift 2 (from 8:00 to 16:00) and Shift 3 (from 16:00 to 24:00) (Yu et al., 2018). We also used one hot encoding to label the season information corresponding to the ATA and EDT: December, January and February are labeled as winter, March, April and May as spring, June, July and August as summer, and September, October and November as autumn. The description of the categorical features is illustrated in Table 8.

Table 8: Description of categorical features

Feature name	Meaning	Encoding method
ATA_day	Week day of ATA Monday (13.9%), Tuesday (13.9%), Wednesday (13.8%), Thursday (14.3%), Friday(14.8%), Saturday (15.1%), Sunday (14.2%).	One hot encoding
EDT_day	Week day of EDT Monday (13.9%), Tuesday (13.9%), Wednesday (13.8%), Thursday (14.3%), Friday(14.8%), Saturday (15.1%), Sunday (14.2%).	One hot encoding
ATA_shift	Hour shift of ATA Shift 1 (36.1%), shift 2 (33.1%), shift 3 (30.8%). (Recall that shift 1 is 0:00∼8:00, shift 2 is 8:00∼16:00 and shift 3 is 16:00∼24:00)	Label encoding
EDT_shift	Hour shift of EDT Shift 1 (35.8%), shift 2 (34.4%), shift 3 (29.8%). (Recall that shift 1 is 0:00∼8:00, shift 2 is 8:00∼16:00 and shift 3 is 16:00∼24:00)	Label encoding
ATA_season	Season shift of ATA winter (26.1%), spring (23.1%), summer (20.8%), autumn (20.8%). (Recall that winter is December, January and February , spring is March, April and May, autumn is June, July and August)	Label encoding
EDT_season	Seaon shift of EDT winter (36.1%), spring 2 (33.1%), summer (30.8%), autumn (30.8%).	Label encoding

4.1.4 Dataset splitting

HKP is primarily a container port that can provide services to container vessels including container loading and unloading, which involves berth and crane allocation amongst other port operations. In this research, we mainly explore the prediction of VTT for container vessels at the HKP. Before splitting the dataset, we first filter out container vessels from all arriving ships based on the vessel type. Initially, there are 14,975 entries of vessel arrival data. After the filtering process, we have

10,001 records related to container vessels. Given that our dataset involves historical delay data related to vessels and berths, we cannot split the data randomly as is traditionally done. Instead, We use the 2022 ship arrival data as the training set and data from January 2023 to March 2023 as the test set. After this division, the training set consists of 8,178 data records and the prediction set consists of 1,823 data records.

4.2 VTT prediction

Machine learning (ML) is a specialized branch of artificial intelligence that centers on the use of data and algorithms to simulate the process of human learning (Zhou, 2021). An ML model autonomously extracts knowledge within the dataset and employs these latent patterns to predict future data. Typically, machine learning involves three primary components: data preparation, algorithmic representation and model optimization (Zhou, 2021; Filom, Amiri, and Razavi, 2022). Data preparation refers to the process of preparing the dataset for the training. Algorithmic representation denotes the structure and formulation of the algorithms used in the machine learning process. Lastly, model optimization involves fine-tuning the model's hyperparameters to achieve the most favorable outcome. Tabular data serves as the fundamental data format for port operations (Filom, Amiri, and Razavi, 2022), in which each row presents an observation or a sample and every column represents a feature. Tree-based models, such as XGBoost and RF, are often utilized for predictive analysis. They tend to be more adept than neural networks at extracting valuable features and information from the dataset through techniques like bagging and ensemble learning (Grinsztajn, Oyallon, and Varoquaux, 2022). The performance of these tree-based models often exceeds that of neural networks in many scenarios, especially when dealing with tabular data. In our research, our prediction dataset is a prime example of the tabular data format, thus we develop an XGBoost model to predict the VTT at the HKP. In addition, for comparison of models performance, we utilize various other prediction models including classification and regression Trees (CART), back propagation neural networks (BPNN) and RF to forecast the VTT. In the following subsections, we will introduce prediction models in details regarding model definition and results evaluation.

4.2.1 Introduction to predictive model

In this subsection, we are going to briefly introduce the XGBoost model and other predictive models employed in this study.

XGBoost is a powerful tree-based machine learning algorithm based on the boosting algorithm. The core idea of boosting algorithms is to integrate many weak classifiers or regressors to form a strong one (Chen et al., 2015). XGBoost achieves this by integrating numerous foundational CART models, creating a robust prediction model. The detailed processes to construct the CART regression model and XGBoost model are presented in Appendix A.1 and Appendix A.2.

The BPNN is a type of multi-layered, feed-forward network trained using the error backpropagation algorithm (Zhou, 2021). This algorithm calculates the gradient of the loss function during the learning process and propagates it backward with respect to each weight in the network (Zhou,

2021). This mechanism is adept at optimizing the loss function and thereby significantly enhances the model's overall performance (Zhou, 2021). The architecture of the BPNN is composed of three types of primary layers: the input layer, hidden layer(s) and the output layer. The training set is fed into the network through the input layer. Subsequently, the data flows forward through the hidden layer, culminating in the output layer where the final results are presented. Owing to its remarkable capabilities in classifying intricate patterns and mapping multi-dimensional functions, the BPNN has secured its place as one of the most widely employed neural network models. This holds true not only in academic research but also in various industry applications. The network's efficiency and versatility in handling complex datasets make it an invaluable tool for pattern recognition and predictive modeling.

The RF model is an ensemble learning method known for its robustness and versatility (Breiman, 2001). The RF model is based on the concept of creating a "forest" of decision trees and outputting a classification or regression result that is the mode or mean of the outputs of individual trees. This methodology fundamentally enhances the generalization and robustness of the model by mitigating the issue of overfitting, which is often observed in individual decision trees (Breiman, 2001). The construction of the RF model involves the generation of multiple decision trees, each grown from a bootstrapped sample of the training set. RF model has been widely applied in various fields of academia and industry. Its applications range from bioinformatics to financial forecasting, demonstrating the model's flexibility and capacity in addressing a multitude of predictive tasks (Breiman, 2001).

Linear regression (LR) is a robust statistical method used to elucidate the relationship between a dependent variable and one or more independent variables, operating under the assumption of a linear relationship between the dependent variables and the independent variable. Its primary objective is to identify the most accurate straight line that fits through the data points. Meanwhile, ridge regression, a variant of regularized LR, emerges as a potent tool in scenarios where the independent variables are highly correlated, i.e., multicollinearity. This approach extends LR by introducing a penalty term to the loss function, a technique known as regularization, which modulates the impact of less significant features, thus enhancing the stability and reliability of the model's predictions. Furthermore, ridge regression adeptly handles overfitting, a common issue in high-dimensional data, ensuring the model's robustness and its capability of generalizing across unseen datasets. By combining these methods, ridge regression ensures the model's resilience and adaptability, essential traits for effective predictive modeling.

In the field of maritime studies, tree-based models have been extensively utilized in a variety of applications. These include, but are not limited to, predicting vessel arrivals, estimating fuel consumption, and strategizing ship inspections. The diverse applications of these models have been thoroughly explored in various studies. (Chu, Yan, and Wang, 2023; Yan, Wang, and Zhen, 2023; Yu et al., 2018; Yan, Wang, Cao, et al., 2021; Yan, Wang, and Du, 2020; Luo, Yan, and Wang, 2023). In this present work, we break new ground by being the first to apply the state-of-the-art XGBoost algorithm along with other predictive models, specifically to forecast the VTT.

4.2.2 Hyper-parameter optimization

In the context of machine learning models, hyperparameters are pre-set configuration variables that govern the learning process and the performance of the models. These parameters are not learned from the training data during the training process but are manually set prior to training (Zhou, 2021). In this study, we utilize a combination of grid search and K-fold cross-validation (CV) methods to identify the optimal hyperparameters for our prediction models. Grid search is an exhaustive tuning method that systematically searches through all combinations of candidate hyperparameter values to identify the set that delivers the best performance on the validation set(s) (Zhou, 2021). For K-fold CV, we initially partition the training set into K subsets, where each subset is referred to as a "fold" (Zhou, 2021). One subset is then designated as the validation set, while the remaining $(K-1)$ subsets serve as the training set. This process is repeated K times, ensuring each subset serves as the validation set at least once. The final validation result is generated by averaging the performance across all folds when they serve as the validation set. For instance, in a scenario where we fit the model using 5-fold CV, an illustration of the hyperparameter tuning process is depicted in Figure 5.

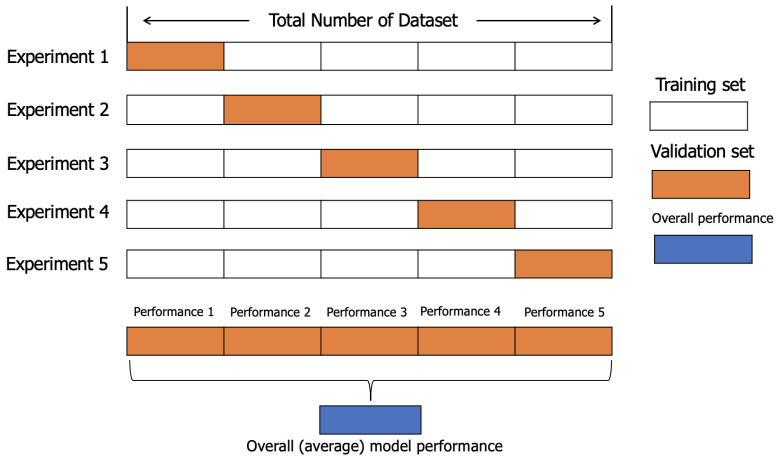


Figure 5: The illustration of five-fold CV

The XGBoost model encompasses several hyperparameters. To obtain the optimal prediction results in this study, we tune the following main hyperparameters: learning_rate, n_estimators, max_depth, sub_sample and colsample_bytree. As outlined in the XGBoost documentation (Chen et al., 2015), the definitions and default values of these selected hyperparameters are provided in Table 9.

Table 9: Hyperparameters to be tuned in the XGBoost model

Hyperparameter	Meaning	Typical default values
learning_rate	The value that shrinks the weights associated with features.	0.1
n_estimators	The number of trees in an XGBoost model	100
max_depth	Maximum depth of a tree in the XGBoost model	3
sub_sample	he fraction of observations to be randomly sampled for the construction of each tree	1
colsample_bytree	The percentage of features that used for building each tree	0.5

The ranges and intervals of hyperparameters in the XGBoost model are detailed in Table 10.

Table 10: Range and internal of values for the specified hyperparameters in the XGboost model

Hyperparameter	Range	Interval
learning_rate	From 0.01 to 0.31	0.05
n_estimators	From 100 to 1500	200
max_depth	From 3 to 15	3
sub_sample	From 0.2 to 1.4	0.3
colsample_bytree	0.4 to 1	0.1

For the RF model, we implement it using the scikit-learn library, where we fine-tune hyperparameters such as max_depth, min_samples_leaf, min_samples_split, n_estimators and max_features. In accordance with the documentation within scikit-learn (Pedregosa et al., 2011), the definitions and default values for these chosen hyperparameters are outlined in Table 11.

Table 11: Hyperparameters to be tuned in the RF regression model

Hyperparameter	Meaning	Typical default values
max_depth	The maximum depth of each CART in the RF model	None
min_samples_leaf	The minimum number of examples that are required to be present in a leaf node	1
min_samples_split	The minimum number of examples that present in a node before it can be split further	2
n_estimators	The number of trees in a RF model	100
max_features	The number of features considered for splitting a node in each tree of a RF model	“sqrt”

The “none” value for max_depth in Table 11 signifies that there is no restriction on the depth of a tree. In other words, the tree can continue expanding until each leaf node only contains one sample, or all leaves contain samples with identical outputs.

Range and interval of values for the specified hyperparameters in the RF model are shown in Table 12. The “auto” value for max_features in Table 12 indicates that all input features are considered when identifying the best splits. “sqrt” implies that only the square root of the total number of features is considered for each node split.

Table 12: Range and interval of values for the specified hyperparameters in the RF model

Hyperparameter	Range	Interval
max_depth	From 5 to 30 or None	6
min_samples_leaf	From 1 to 10	1
min_samples_split	From 1 to 6	1
n_estimators	From 100 to 1500	100
max_features	"auto" or "sqrt"	\

For the BPNN, the primary hyperparameters that we tune include hidden_layer_sizes, activation function, batch_size, max_iter, and learning_rate (Pedregosa et al., 2011). The definitions and default values of these selected hyperparameters for the BPNN can be found in Table 13.

Table 13: Hyperparameters to be tuned in the BPNN

Hyperparameter	Meaning	Typical default values
Hidden_layer_sizes	The number of neurons in the hidden layer of the neural network	100
Activation	Activation function for the neural network	"relu"
Learning_rate	The initial learning rate that controls the step-size in updating the weights	0.001
Max_iter	The maximum number of iterations permitted during the training process	200

The default value of "relu" activation function in Table 13 refers to the rectified linear unit function, represented as $f(x) = \max(0, x)$. The "relu" function is commonly used as the activation function in neural networks due to its computational efficiency. In this study, we implement a BPNN consisting of three layers: an input layer, a hidden layer and an output layer. Table 14 lists the search ranges and intervals for the hyperparameters used in the BPNN.

Table 14: Range and internal of values for the specified hyperparameters in the BPNN

Hyperparameter	Range	Interval
Hidden_layer_sizes	From 10 to 200	10
Activation	"logistic", "relu" and "tanh"	\
Learning_rate	From 0.0001 to 0.01	0.0001
Max_iter	From 50 to 200	10

The value of "logistic" for the activation function in Table 14 refers to the logistic sigmoid function, represented as $f(x) = 1/(1 + e^{-x})$. Similarly, "tanh" refers to the hyperbolic tangent function, which is written as $f(x) = \tanh(x)$. In the experimental section, we employ a grid-search methodology along with a five-fold CV approach to obtain the optimal hyperparameters for the model. The resulting optimal hyperparameters for the XGBoost model, the RF model and the BPNN are elaborately presented in Tables 15, 16 and 17.

Table 15: Adopted hyperparameters for the XGBoost model

Hyperparameter	learning_rate	max_depth	n_estimators	sub_sample	colsample_bytree
Selected value	0.01	6	1000	0.5	1.0

Table 16: Adopted hyperparameters for the RF model

Hyperparameter	max_depth	min_samples_leaf	min_samples_split	n_estimators	max_features
Selected value	'None'	10	7	660	'auto'

Table 17: Adopted hyperparameters for the BPNN

Hyperparameter	hidden_layer_sizes	activation	learning_rate	max_iter
Selected value	100	"relu"	0.0002	200

4.2.3 Model assessment metrics

For offline evaluation of VTT prediction, the model's output is considered as the predicted value, while the difference between the ADT and ATA is treated as the ground truth of the actual VTT. To provide a comprehensive assessment of the model's performance, three commonly adopted metrics are utilized: RMSE, the mean absolute error (MAE) and R-squared (R^2) (Veenstra and Harmelink, 2021). R-squared, also known as the coefficient of determination, is a statistical measure used in regression analysis. It represents the proportion of the variance of a dependent variable that can be explained by an independent variable or variables. This value can range from 0 to 1. Given a total number of n ships, y_i is the ground truth VTT, $\bar{y}$ is the mean value of ground truth VTT and $\hat{y}_i$ is the predicted VTT for ship $i, i = 1, \ldots, n$, the definitions of the metrics are as follows.

RMSE:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}. \tag{1}$$

MAE:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| . \tag{2}$$

R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{3}$$

The three metrics are used in the following sections to comprehensively assess the prediction performance of VTT at the HKP.

4.2.4 Analysis of VTT prediction results

Utilizing the previously identified optimal hyperparameters, we employ XGBoost, RF, BPNN, LR and ridge regression on the dataset to evaluate their predictive performance in estimating VTT. Subsequently, we compare the predicted VTT by these models against the EDT records provided in the original test set, using MAE, RMSE and R-squared (R^2) as metrics for comparison. Additionally, to assess the impact of EDT and its related features on the prediction of VTT, we perform an analysis where EDT-related features, specifically "EDT_berth" and "expected VTT", are excluded from the dataset. Furthermore, to demonstrate the superiority of our model and the importance of selecting

novel parameters, we also conducted a comparative analysis with the model and parameters used in the paper by Štepec et al. (2020). In this comparison, we applied their parameters and model to our dataset. This means that features like vessel maximum draft, berth generic features, and EDT-related features are excluded from the dataset for training. Following this modification, we then proceed to retrain the RF and XGBoost models, using the revised dataset to conduct further prediction tasks. The prediction results are presented in Table 18.

Table 18: VTT prediction results by different machine learning models

Model	RMSE	MAE	MSE	R^2
Test_set	8.0	5.12	64.51	None
Decision_tree	6.64	4.32	44.09	0.761
RF	6.15	4.04	37.81	0.795
XGBoost	**6.06**	**3.94**	**36.80**	**0.804**
BPNN	8.21	5.33	67.41	0.02
Ridge_regression	7.67	4.99	58.83	0.465
LR	7.71	5.11	59.44	0.441
RF_no_edt	7.89	5.08	57.61	0.119
XGBoost_no_edt	7.71	4.83	51.99	0.174
RF_S	7.90	5.08	62.41	0.108
XGBoost_S	7.75	4.88	60.06	0.165

In Table 18, the row labeled "Test_set" refers to the error between the original EDT that is reported by individual ships and the real ADT in the test set. We use this data as the basis for comparing the results of our model predictions. "RF_no_edt" indicates the results predicted by the RF model, which does not take into account any EDT related information. Similarly, "XGBoost_no_edt" refers to the results predicted by the XGBoost model, again without considering any EDT related features. "RF_S" and "XGBoost_S" indicate the prediction results obtained by applying the RF and XGBoost models, along with the features proposed by Štepec et al. (2020). to our dataset. The results presented in Table 18 reveal that, except for the BPNN, all the other seven VTT prediction models show reasonably more accurate results on the test set, when compared with the initially reported EDT data. In addition, the BPNN performance lags behind, even underperforming in comparison to the original test data. The primary reason for this discrepancy is that BPNN with one hidden layer, as opposed to tree-based models, do not yield as effective results when applied to tabular data (Grinsztajn, Oyallon, and Varoquaux, 2022). Regarding the models that have shown predictive improvements, tree-based models work well in the prediction tasks and XGBoost achieves the best results across all four metrics. The delay error, as measured by the MAE evaluation metric, decreases from 5.12 hours in the original EDT test set to 3.94 hours as predicted by the XGBoost model, marking a decrease of 23%. The RMSE also sees a decrease from 8.0 hours to 6.06 hours with a reduction of 24.3%. The MSE drops from 64.51 h^2 to 36.80 h^2, indicating a substantial decrease of 43%. Furthermore, XGBoost also achieves the best performance in the R^2 evaluation, yielding a

score of 0.804.

Table 18 also illustrates that if the RF model or XGBoost model do not consider EDT related features during prediction, the effectiveness of the model prediction would significantly diminish. The VTT predictive results from XGBoost and RF without considering EDT related features are only marginally better than the original EDT test set data, with the XGBoost model reducing the delay error by a mere 0.11 hours for RMSE and 0.04 hour for MAE error. The results shown in the RF_no_edt and XGBoost_no_edt rows in Table 18 emphasize the significant role that EDT and its related features play in real-world port operation and predicting the actual VTT. Although there may be error in the reported EDT, this feature is still the most crucial basis for VTT evaluation and prediction when ATA is known. As for the comparative experiments with the approach of Štepec et al. (2020), the results show that the RF_S and XGBoost_S models, which presumably apply the features from Štepec et al. (2020), demonstrate a reduced performance across all metrics compared to our proposed prediction models. Specifically, RF_S and XGBoost_S have higher RMSE values (7.90 and 7.75, respectively) compared to our RF and XGBoost models (6.15 and 6.06, respectively), indicating that our models outperform in terms of RMSE, with the lowest RMSE observed in our XGBoost model. For MAE, which similarly benefits from lower values, RF_S and XGBoost_S again show inferior performance compared to our RF and XGBoost, which have the lowest MAE values among all models (4.04 and 3.94, respectively). The situation is similar for MSE and R^2 metrics. Overall, the experimental results corroborate the effectiveness of our approach. The VTT prediction results can be further analyzed to generate strategic and managerial implications for policymakers and port practitioners.

In addition to evaluating model performance, we also identify the top ten most important features for VTT prediction from the constructed XGBoost model. The feature importance score can be automatically computed using a built-in function, XGBoost.feature_importances_, from the XGBoost Python library (Chen et al., 2015). Feature importance is presented here in the "total_gain" method, where the sum of all feature importance values equals 1 and it reflects the contribution of each feature towards improving the model's accuracy. The higher the score, the more significant the feature is in terms of its importance to the model. Feature importance of the top 10 most important features in the XGBoost model constructed is shown in Table 19.

Table 19: The top 10 most importance features and their importance scores of the constructed XGBoost model

Rank	Feature	Importance score
1	Expected VTT	0.5651
2	Length	0.1329
3	Beam	0.1004
4	GT	0.0812
5	Max_draft	0.0542
6	ATA_inport	0.0358
7	EDT_berth_dealy	0.0310
8	Vessel_delay	0.0219
9	ATA_shift_2[8]	0.0181
10	ATA_berth_delay	0.0108

The feature importance analysis presented in Table 19 indicates that several key factors primarily determine the performance of VTT. These factors include the expected VTT (calculated as the vessel's EDT minus its ATA), vessel generic features (such as length, beam and GT), the number of vessels in port at the ATA time and the average historical turnaround time delay data for both the berth and the vessel. Specifically, the expected VTT is the most critical determinant among the features, as it is reported by the vessel itself and the port uses it as a reference for berth scheduling planning. The vessel's length and beam, which rank second and third, respectively, directly influence the allocation of berth (Yu et al., 2018; Chu, Yan, and Wang, 2023). Additionally, the vessel's GT and the maximum draft during navigation, ranking fourth and fifth, as they reflect the carrying capacity of the container vessels, and thus affect the turnaround time. As for the number of vessels in port, it reflects the busyness of the port at the time when the ship arrives at the port. When there are too many vessels in port, the port becomes busy and may be unable to effectively schedule berth allocations for the incoming vessel, which in turn can impact the actual turnaround time of the vessels. Besides, the historical delay data of the berth where the vessel's EDT data is reported, as well as the vessel's own historical delay data, have significant impacts on the vessel final turnaround time. Low historical VTT delay data of the berth reflects the efficiency of that berth, making the ADT of the vessel closer to the EDT. This, in turn, leads to more accurate results in predictive modeling. Besides, vessels with a history of less delay in VTT tend to provide more trustworthy EDT information and reliable EDT information ultimately enhances the accuracy of model predictions.

Additionally, based on the results of the feature importance analysis, we select the ten most important parameters from Table 19 and actual VTT for feature correction analysis and visualization. The results of the feature correlation matrix visualization are shown in Figure 6.

[8] ATA_shift_2 is the feature representing the hour shift of ATA from 8:00 to 16:00.

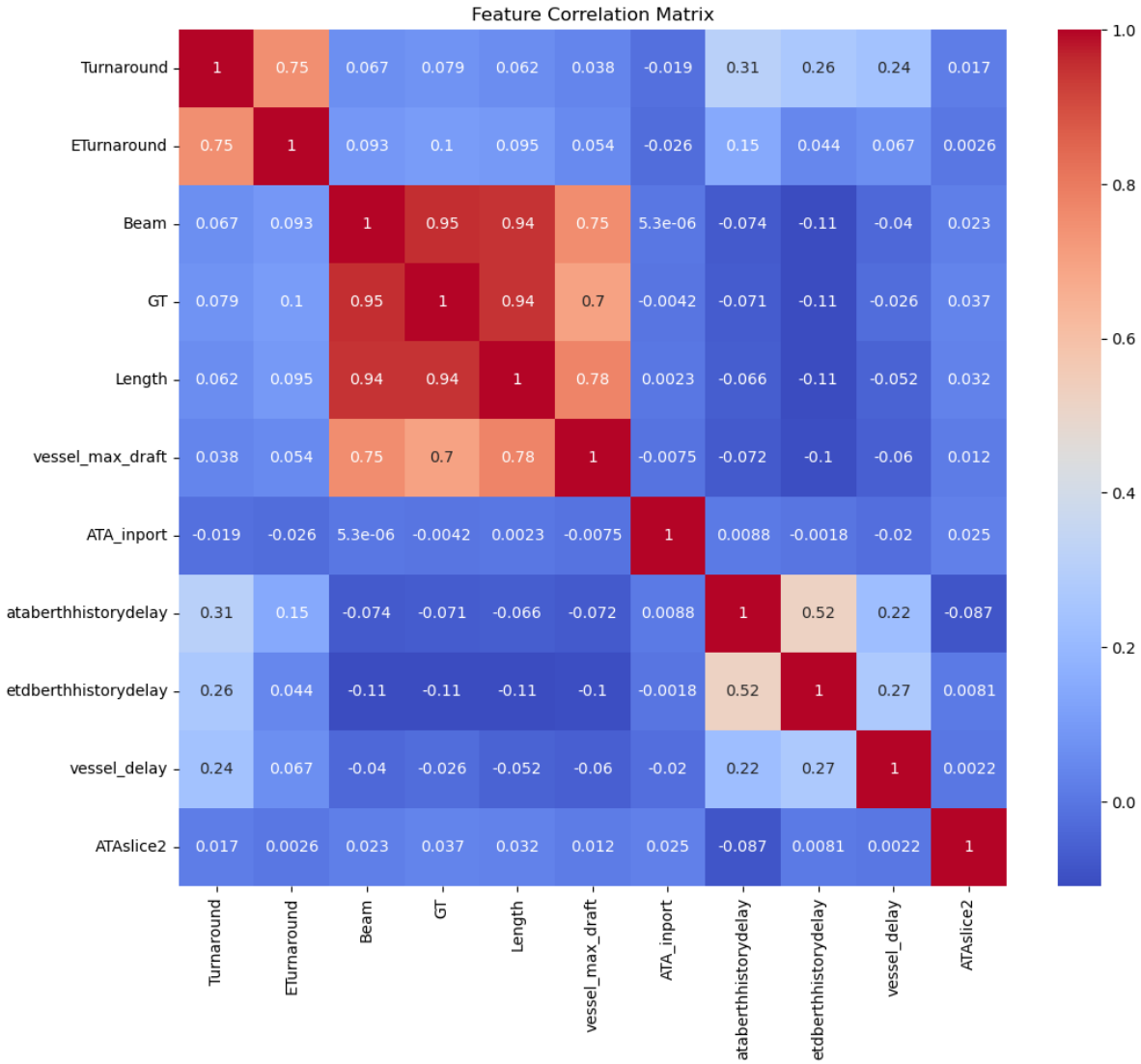


Figure 6: Feature correlation matrix visualization results

In correlation matrix analysis, each cell in the matrix represents the correlation coefficient between the variables on the corresponding row and column. A correlation coefficient can range from -1 to 1, where: 1 indicates a perfect positive correlation (as one variable increases, the other also increases). -1 indicates a perfect negative correlation (as one variable increases, the other decreases). In this part of the analysis, we focus on the actual VTT, represented as "Turnaround," and its related parameters. The correlation coefficient of 0.75 between "Turnaround" and the expected VTT ("ETurnaround") suggests a substantial positive relationship, indicating that longer expected VTT is often associated with longer actual VTT. The analysis reveals that the expected VTT as a significant indicator of the actual VTT which is similar to the findings from the feature importance analysis. Besides, the vessel's physical attributes such as beam, GT, length, and maximum draft

show a week linear relationships with "Turnaround" feature. Port operational features including berth and vessel historical delay records, also exhibit very weak correlations with "Turnaround." This indicates that these elements potentially affect the VTT.

5 Model extension and policy insights

In the following section, we will introduce the potential extensions from the VTT prediction models proposed and the policy insights that can be generated from the VTT prediction results. To systematically present our findings, we have consolidated all the relevant details into Table 20. This table provides a comprehensive overview of the extensions and implications of the VTT prediction model. It highlights potential improvements and strategic insights that are intended to guide policy decisions in the field of port operation.

Table 20: Summary of extensions and policy insights from the VTT evaluation and prediction results

General perspective	Specific scenarios
Generate rational management insights	1) Achieve more accurate VTT prediction in a quantitative way 2) Manage and allocate port resources from a strategic perspective 3) Serves as a reference for vessel scheduling
Enhance effective port operation	1) Mitigate port congestion 2) Improve port productivity 3) Refine port coordination
Reduce overall expenses	1) Port authorities: increase vessel handling revenue and reduce port operational cost 2) Cargo owner: cut costs related to unexpected vessel schedule changes and expediting goods delivery to the market 3) Vessel operators: reduce fuel consumption and cargo delivery delay
Promote green shipping practices	1) Conserve carbon energy in port operation 2) Minimize vessel emissions 3) Bolster the reputation of shipping companies and port authorities in sustainable development

5.1 Generate rational management insights

This research provides invaluable insights into the quantitative evaluation and prediction of the VTT at the HKP. VTT is vital in port operations; port operators rely on it to strategize port resources management in advance. Enhancing VTT accuracy can significantly improve the efficiency of port management. Compared to the original EDT data reported by vessel operators on their way to the port or in port, the proposed method exhibits superior effectiveness in evaluating and estimating

the VTT at the HKP. Our prediction model can reduce the VTT delay error by approximately 20% in terms of MAE and RMSE when compared to the reported VTT by vessel operators. The analysis results also indicate that the reported VTT and generic vessel features play a crucial role in VTT prediction at the HKP.

In practical port operations and management, accurate prediction of VTT is essential for efficient port resource management. They serve as a foundation for intelligently optimizing port operations. For example, berth allocation have typicality relied on expert knowledge and established allocation strategies. Port staff would arrange the allocated berths based on the EDT data reported by vessels. However, with the introduction of VTT prediction models, berth scheduling can now be done with greater intelligence and precision. The VTT models allow for more accurate forecasting of VTT at the port, enabling the port staff to make more informed decisions and dynamically adjust berth scheduling in real-time. From a quantitative perspective, in the HKP we discussed, if the average VTT is reduced by 24% through our predictive model, the port could theoretically handle 24% more ships within a given planning horizon or achieve a 24% reduction in port operation time over a fixed time period. By harnessing the insights provided by VTT prediction, port administrators can significantly reduce idle berth time, increase efficiency, and adopt innovative, data-driven scheduling strategies that are previously unattainable. This enables effective management of critical port resources, including pilots, cranes, tugboats, workforce, amongst operational facilities. Such foresight ensures that the allocation of these indispensable resources is maximized, leading to enhanced flow and efficiency of operations within the port.

Besides, VTT prediction can serve as a valuable reference for formulating vessel scheduling management. For shipping companies, vessel scheduling is a complex system, with each node symbolizing a port call. Understanding the VTT at every port enables these companies to devise an accurate liner vessel voyage plan. Moreover, when shipping companies possess a clear grasp of anticipated VTTs, they benefit from operational flexibility. This knowledge empowers them to make well-informed decisions. For ports, an accurate VTT prediction is not just a luxury, but a necessity. With a clear understanding of VTT, ports can make more rational scheduling decisions for vessel arrivals. It also allows ports to better plan and coordinate the arrival time of vessels, helping to achieve "just in time" vessel arrival. Consequently, this reduces berth congestion and increases utilization, translating into significant cost savings for both the port and shipping companies.

5.2 Enhance port operational efficiency

Accurate VTT predictions are integral to enhancing the effectiveness of modern port operations. A clear understanding of a vessel's departure time enables port operators to allocate time slots both rationally and efficiently for incoming and outgoing vessels. By adopting this data-driven strategy, the time vessels spend waiting at anchorage points is minimized, leading to reduced congestion and bolstered planning. Furthermore, managing traffic within port waters becomes more streamlined, ensuring navigable and safe waterways. This is particularly vital for bustling ports, where congestion can lead to significant delays and potential hazards. In essence, VTT predictions are foundational in achieving optimized efficiency and responsiveness in port operations.

The ripple effect of efficient VTT prediction extends well beyond mere port congestion mitigation. By determining and minimizing the duration a vessel spends at the port, port authorities can accommodate more vessels within the same time window, and as a result, the productivity within port operations is significantly amplified. For instance, using this study as an example, if our predictive model reduces the average VTT by 24%, it creates a win-win situation for both the port and the ship operators. The decrease in port idle time can be leveraged to enhance the port's productivity and also to lower the operational costs of the vessels and port authorities. Theoretically, within a specific time frame, this could lead to an increase in the port's productivity efficiency by up to 24%. Furthermore, accurate VTT prediction significantly increases economic sustainability of port operations by lowering operational costs. Currently, the cost of handling a container at HKP is 2,000 HKD (Hong Kong Trade Development Council, 2022). By utilizing EDT predictions, we can potentially increase the port's productivity by an ideal 24%. This means that, for the same time and expenditure, we can handle 1.24 times the number of containers compared to before. As a result, the cost to handle a single container could be reduced to approximately 1,600 HKD. This significant cost reduction can greatly enhance the port's competitiveness. For shipowners, reducing vessel port idle time through EDT predictions means a reduction in operational costs. For example, at the HKP, when a ship remains at anchor in Hong Kong waters beyond the required departure time, it is charged at a rate of 0.02 HKD per 100 tons (Hong Kong Trade Development Council, 2022). If a vessel weighs 50,000 tons, our prediction model can reduce the waiting time by approximately 24%, around 2 hours. This reduction potentially saves 1,000 HKD in penalties per visit.

Besides, the benefits of precise VTT predictions extend even beyond the waterfront. Ports are intricate hubs of activity, intertwined with a multitude of stakeholders. By leveraging accurate VTT predictions, the stakeholders can significantly enhance their coordination efforts. For instance, since ports often collaborate with trucking and rail services, anticipating vessel delays enables these entities to modify their schedules, thereby minimizing unnecessary waiting periods and reducing operational costs.

5.3 Reduce overall commercial expenses

Effective VTT prediction can lead to the mitigation of costly delays and substantial cost savings at ports, benefiting port authorities, vessel operators, and cargo owners alike.

For port authorities, precise VTT predictions facilitate optimal berth allocation, reducing instances of unoccupied berths and enhancing capacity utilization. This optimization can not only accommodate more vessels within a given period but also can lead to increased port operation revenue. Furthermore, effective VTT management reduces port congestion, thereby minimizing vessel delays and decreasing the time berths remain unoccupied. This, in turn, lowers port operational costs.

For cargo owners, reliable VTT predictions enable more efficient logistics planning. By providing precise estimates of when a ship is ready to depart from a port of call, these predictions cut costs related to unexpected vessel schedule changes or delays. This efficiency expedites goods delivery to the market, leading to improved customer satisfaction and potentially higher sales, and thus increase

profitability.

For vessel operators, efficient VTT predictions lead to reduced vessels' idling time at ports, which in turn translates to decreased fuel consumption — a major component of a vessel's operating costs. A quicker turnaround time enables vessels to make more trips within a specific period, thereby enhancing their revenue-earning potential. Furthermore, minimizing delays can be an attractive proposition for more business. In a competitive market, the ability to predict and manage delays can be a key differentiator.

In summary, effective VTT prediction is more than a technical advantage; it's a strategic asset that enhances various aspects of port operations. It contributes to a more responsive and economically sustainable port ecosystem, creating a win-win scenario for all stakeholders involved.

5.4 Promote green shipping practices

The maritime industry, while indispensable for global trade, has historically been a significant contributor to environmental pollution. With increasing awareness of climate change and environmental degradation, there has been a push towards more sustainable operations within this sector. Incorporating VTT predictions into green shipping initiatives provides a holistic approach to sustainability, balancing operational efficiency with environmental responsibility.

For port authorities, accurately predicting the VTT can considerably reduce the idling duration of port berths, which in turn results in port operation energy conservation and promotes the ethos of green shipping. On the vessel front, idle vessels, while waiting at ports, can contribute significantly to greenhouse gas emissions, due to the necessity to keep their engines running. By accurately predicting the start of a vessel's next journey, shipping companies can save on this wasted fuel and optimize their routes and speeds to increase fuel efficiency. Reducing the idle time at ports doesn't just cut down on fuel consumption, but also results in a significant reduction in emissions. This precision and optimization in operations, driven by accurate VTT forecasting, is instrumental in enabling vessels to meet their green shipping sustainability goals and is a meaningful stride towards the broader goal of reducing the shipping industry's carbon footprint.

Besides, adopting green shipping practices and technologies can bolster the reputation of shipping companies and port authorities, making them more attractive to clients, investors, governments and the broader public. Companies demonstrating proactive measures to reduce their environmental impact are likely to garner favorable views and thus attract more business. For ports, being recognized as an environmentally responsible organization can significantly enhance their reputation, rendering them more appealing to customers and investors who place high value on sustainability. Furthermore, ports that prioritize green shipping could potentially draw business from shipping companies with similar commitments to environmental sustainability. Such companies might prefer to partner with ports whose values align with their own and can support their green technologies. In practice, numerous ports are making strides in enhancing green port and port sustainability. For example, the Maritime & Port Authority of Singapore (MPA) has unveiled a blueprint targeting carbon neutrality from seven aspects, including port terminals, harbour craft, and research development (Maritime Singapore, 2023). The plan also includes dedicated funds to develop capabilities

in port energy management and modeling together with the universities and companies (Maritime Singapore, 2023).

6 Conclusion

This study makes a preliminary attempt to predict VTT at the HKP using data-driven models. An XGBoost regression model has been developed to enhance the precision of VTT predictions. When tested with 2023 container vessel VTT data records, the XGBoost model demonstrates a significant improvement in VTT prediction at the HKP compared to the originally reported EDT data by the vessel operator. The model is able to reduce the MAE error from 5.1 hours to 3.9 hours on the test set, signifying a 23% reduction in error when compared with the original reported EDT data from the vessel operator. Similarly, a 24% reduction in error is noted when evaluating the model performance using RMSE metric, decreasing the error from 8.0 hours to 6.1 hours. Moreover, the proposed XGBoost model exhibits an impressive R-squared value at 0.804, indicating a strong degree of correlation. Furthermore, feature importance analysis is conducted based on the XGBoost model. The findings suggest that the reported VTT by vessels themselves, is the principal factor influencing the model's performance.

Based on our prediction results, our study underscores the potential of advanced data-driven models, especially XGBoost, for achieving predictive accuracy superior to that of traditional models and originally reported data in VTT prediction. Besides, our proposed XGBoost model is designed with flexibility to accommodate the prediction of VTT for various ship types. This can be achieved by incorporating ship type as an additional feature and then use one-hot encoding method for feature preprocessing. These parameters, once integrated into our XGBoost model, would enable it to differentiate among various vessel types and train to predict turnaround times for these different types of vessels. Moreover, we offer valuable insights for model extension and policy-making which presents crucial opportunities for enhancing operational efficiency and resilience at ports. We anticipate that our research will lay the groundwork for future studies in this area, heralding new opportunities for the integration of advanced machine learning models into port operations and shipping management.

However, it is crucial to remember that while our VTT prediction results are promising, there are still several limitations in our study. The primary limitation of this research pertains to data availability. In the HKP, vessel arrival and departure data are readily accessible online - a privilege not afforded by every port. Consequently, our predictive models and research findings may primarily apply to ports with data attributes similar to those of HKP, which is a study limitation. However, it's critical to note that this constraint mainly impacts researchers and might be less significant for port management teams. The port staff typically has access to the vessel traffic management system internally used by the port that provides crucial vessel time features, such as the EDT. Therefore, port staff can still employ our predictive model to estimate VTT.

The second limitation of this study is related to the robustness of the model. The VTT prediction results, presented in Tables 18 and 19, highlight the importance of EDT related features in VTT prediction. If vessels do not report their EDT prior to departure, the efficacy of the model proposed in this study could be compromised. Furthermore, generic vessel features are vital for VTT prediction.

If these generic features are not available when a vessel arrives at the port, the VTT prediction for that vessel might be inadequate.

The third limitation involves the nature of our dataset, which does not include real-time data from vessels and port operations. For instance, the VTT dataset lacks real-time AIS data for ships, and thus the latitude and longitude of the vessel voyage and the speed of the vessel. Similarly, real-time port operation data (such as berth occupancy status, quay crane operational efficiency, etc.) is absent from the dataset. These real-time datasets are not publicly available due to security and privacy considerations.

For further research, many research questions remain open, particularly in terms of prediction and optimization. For instance, we have considered factors such as vessel time-related elements, historical vessel and berth VTT performance, and generic vessel features. However, in this current study, we have not fully incorporated weather conditions, such as daily rainfall and wind speed, which directly affect port operational efficiency. The study only accounts for port tidal information, neglecting other weather elements. Adverse weather conditions, including high winds, storms, and heavy rainfall, can lead to temporary port closures or restrictions. These disruptions impact the availability of berths and other port services, subsequently influencing the VTT. For future research, it would be beneficial to consider a broader range of weather factors in our dataset. To integrate these additional weather features effectively, we should first encode them into a model-compatible format. Once formatted, these weather features can be synchronized with the existing VTT training dataset. This integration will empower our model to more precisely assess the influence of diverse weather conditions on vessel VTT predictions. By doing so, we can substantially refine the precision of our predictive models, offering a more robust tool for maritime navigation and safety planning.

Additionally, if we can obtain real-time ship AIS voyage data and port operational data, we can integrate the real-time vessel geographic information and speed from AIS and port operation data into the VTT dataset. This could enable a more comprehensive evaluation and prediction of VTT, potentially leading to increased prediction accuracy.with regard to the VTT prediction task, the proposed XGBoost model is the most suitable for ports with similar EDT data. The predictive performance of the XGBoost model may be sub-optimal for ports that do not provide EDT data, a limitation that is further illustrated by the prediction results in Table 18. As for optimization, we suggest future research to propose an optimization model for planning daily port operations such as berth allocation and quay crane schedule. This could potentially improve the efficiency of daily port operations based on the VTT prediction results. And we could develop a smart prediction-then-optimization method to prescribe more efficient port operating decisions. Specifically, our first step would involve predicting more accurate VTT. These predictions can then be used as key parameters to optimize berth scheduling, thereby improving port operational efficiency.

In essence, accurate VTT prediction is a crucial tool for efficient resource management within port operations. It enables optimal berth allocation and effective utilization of port resources. This, in turn, increases port productivity and cost-effectiveness. This study represents a significant stride towards achieving streamlined and efficient port operations, enhancing the potential for the creation of increasingly digitized and smart ports of the future.

References

Abreu, Levi R., Ingrid S.F. Maciel, Joab S. Alves, Lucas C. Braga, and Heráclito L.J. Pontes (2023). "A decision tree model for the prediction of the stay time of ships in Brazilian ports". In: *Engineering Applications of Artificial Intelligence* 117, 105634.

Barua, Limon, Bo Zou, and Yan Zhou (2020). "Machine learning for international freight transportation management: a comprehensive review". In: *Research in Transportation Business & Management* 34, 100453.

Breiman, Leo (2001). "Random forests". In: *Machine Learning* 45, 5–32.

Brouer, Berit Dangaard, Christian Vad Karsten, and David Pisinger (2016). "Big data optimization in maritime logistics". In: *Big Data Optimization: Recent Developments and Challenges*, 319–344.

Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. (2015). "XGBoost: extreme gradient boosting". In: *R package version 0.4-2* 1.4, 1–4.

Chu, Zhong, Ran Yan, and Shuaian Wang (2023). "Evaluation and prediction of punctuality of vessel arrival at port: a case study of Hong Kong". In: *Maritime Policy & Management*, 1–29.

Duan, Hongda, Fei Ma, Lixin Miao, and Canrong Zhang (2022). "A semi-supervised deep learning approach for vessel trajectory classification based on AIS data". In: *Ocean & Coastal Management* 218, 106015.

Ducruet, César and Olaf Merk (2013). "Examining container vessel turnaround times across the world". In: *Port Technology International* 59.

Feng, Yuanjun, Dong-Ping Song, Dong Li, and Ying Xie (2022). "Service fairness and value of customer information for the stochastic container relocation problem under flexible service policy". In: *Transportation Research Part E: Logistics and Transportation Review* 167, 102921.

Filom, Siyavash, Amir M Amiri, and Saiedeh Razavi (2022). "Applications of machine learning methods in port operations–A systematic literature review". In: *Transportation Research Part E: Logistics and Transportation Review* 161, 102722.

Golias, Mihalis M, Georgios K Saharidis, Maria Boile, Sotirios Theofanis, and Marianthi G Ierapetritou (2009). "The berth allocation problem: Optimizing vessel arrival time". In: *Maritime Economics & Logistics* 11, 358–377.

Grinsztajn, Leo, Edouard Oyallon, and Gael Varoquaux (2022). "Why do tree-based models still outperform deep learning on typical tabular data?" In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. New Orleans.

Hong Kong Government (2023). *The Complete Berthing Guidelines for Port of Hong Kong*. https://www.mardep.gov.hk/en/pub_services/pdf/berthguide.pdf. Accessed: 2023-11-29.

Hong Kong Marine Department (2022). *Vessel Traffic Management System Report*. = https://data.gov.hk/en-data/dataset/hk-md-mardep-vessel-traffic-management-system-report. Accessed: 2023-08-24.

Hong Kong Trade Development Council (2022). *Overview of Hong Kong Port Transportation*. https://info.hktdc.com/shippers/vol26_1/vol26_1_chi11_01.htm. Accessed: 2023-11-29.

Kim, Kap Hwan and Kyung Chan Moon (2003). "Berth scheduling by simulated annealing". In: *Transportation Research Part B: Methodological* 37.6, 541–560.

Lewis, Roger J (2000). "An introduction to classification and regression tree (CART) analysis". In: *Annual meeting of the society for academic emergency medicine in San Francisco, California*. Vol. 14. Citeseer.

Li, Bin and Yuqing He (2020). "Container terminal liner berthing time prediction with computational logistics and deep learning". In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2417–2424.

Lin, Bowen, Mao Zheng, Xiumin Chu, Wengang Mao, Daiyong Zhang, and Mingyang Zhang (2023). "An overview of scholarly literature on navigation hazards in Arctic shipping routes". In: *Environmental Science and Pollution Research*, 1–17.

Lin, Bowen, Mao Zheng, Xiumin Chu, Mingyang Zhang, Wengang Mao, and Da Wu (2024). "A novel method for the evaluation of ship berthing risk using AIS data". In: *Ocean Engineering* 293, 116595.

Liu, Zhao, Hairuo Gao, Mingyang Zhang, Ran Yan, and Jingxian Liu (2023). "A data mining method to extract traffic network for maritime transport management". In: *Ocean & Coastal Management* 239, 106622.

Lun, YH Venus, Kee-hung Lai, TC Edwin Cheng, and Dong Yang (2023). "Business Strategy in Shipping". In: *Shipping and Logistics Management*. Springer, 69–83.

Luo, Xi, Ran Yan, and Shuaian Wang (2023). "Comparison of deterministic and ensemble weather forecasts on ship sailing speed optimization". In: *Transportation Research Part D: Transport and Environment* 121, 103801.

Ma, Dongliang, Jine Wei, Ye Li, Fang Zhao, Xi Chen, Yuchao Hu, Shanshan Yu, Tianhao He, Ruihe Jin, Zhaozhao Li, et al. (2023). "MLDet: Towards efficient and accurate deep learning method for Marine Litter Detection". In: *Ocean & Coastal Management* 243, 106765.

Marine Department of Hong Kong (2023). *Latest tidal information*. https://www.mardep.gov.hk/en/pub_services/fees.html. Accessed: 2023-11-29.

MarineTraffic (2023). *Marine Traffic*. URL: https://www.marinetraffic.com/ (visited on 05/20/2022).

Maritime, Hong Kong and Port Board (2022). *Port of Hong Kong Introduction*. URL: https://www.hkmpb.gov.hk/en/port.html.

Maritime Singapore (2023). *Maritime Singapore Decarbonisation Blueprint: Working Towards 2050*. = https://www.mpa.gov.sg/docs/mpalibraries/mpa-documents-files/sustainability-office/mpa-decarb-blueprint-2050a.pdf. Accessed: 2023-08-24.

Mokhtar, Kasypi and Muhammad Zaly Shah (2006). "A regression model for vessel turnaround time". In: *Tokyo Academic, Industry & Cultural Integration Tour*, 1–15.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. (2011). "Scikit-learn: Machine learning in Python". In: *The Journal of Machine Learning Research* 12, 2825–2830.

Politikos, Dimitris V, Argyro Adamopoulou, George Petasis, and Francois Galgani (2023). "Using artificial intelligence to support marine macrolitter research: A content analysis and an online database". In: *Ocean & Coastal Management* 233, 106466.

Rinke, Andreas and Jan Schwartz (Oct. 2022). *German go-ahead for China's Cosco stake in Hamburg port unleashes protest*. https://www.reuters.com/markets/deals/german-cabinet-approves-investment-by-chinas-cosco-hamburg-port-terminal-sources-2022-10-26/. Accessed: 2023-11-29.

Rodrigues, Filipe and Agostinho Agra (2022). "Berth allocation and quay crane assignment/scheduling problem under uncertainty: a survey". In: *European Journal of Operational Research* 303.2, 501–524.

Smith, Daniel (2021). "Big data insights into container vessel dwell times". In: *Transportation Research Record* 2675.10, 1222–1235.

Štepec, Dejan, Tomaž Martinčič, Fabrice Klein, Daniel Vladušič, and Joao Pita Costa (2020). "Machine learning based system for vessel turnaround time prediction". In: *2020 21st IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 258–263.

Tian, Xuecheng, Ran Yan, Shuaian Wang, and Gilbert Laporte (2023). "Prescriptive analytics for a maritime routing problem". In: *Ocean & Coastal Management* 242, 106695.

United Nations Conference on Trade and Development (2021). *The Review of Maritime Transport 2020*. URL: https://unctad.org/system/files/official-document/rmt2020_en.pdf (visited on 06/21/2023).

Veenstra, Albert and Rogier Harmelink (2021). "On the quality of ship arrival predictions". In: *Maritime Economics & Logistics* 23.4, 655–673.

Wang, Shuaian and Ran Yan (2023). "Fundamental challenge and solution methods in prescriptive analytics for freight transportation". In: *Transportation Research Part E: Logistics and Transportation Review* 169, 102966.

Wang, Xinyu, Zhao Liu, Ran Yan, Helong Wang, and Mingyang Zhang (2022). "Quantitative analysis of the impact of COVID-19 on ship visiting behaviors to ports-A framework and a case study". In: *Ocean & coastal management* 230, 106377.

WRS (2020). *World Shipping Register*. URL: https://world-ships.com/ (visited on 05/20/2023).

Xu, Xueqian, Xinqiang Chen, Bing Wu, Zichuang Wang, and Jinbiao Zhen (2022). "Exploiting high-fidelity kinematic information from port surveillance videos via a YOLO-based framework". In: *Ocean & Coastal Management* 222, 106117.

Xu, Ya, Qiushuang Chen, and Xiongwen Quan (2012). "Robust berth scheduling with uncertain vessel delay and handling time". In: *Annals of Operations Research* 192, 123–140.

Yan, Ran, Shuaian Wang, Jiannong Cao, and Defeng Sun (2021). "Shipping domain knowledge informed prediction and optimization in port state control". In: *Transportation Research Part B: Methodological* 149, 52–78.

Yan, Ran, Shuaian Wang, and Yuquan Du (2020). "Development of a two-stage ship fuel consumption prediction and reduction model for a dry bulk ship". In: *Transportation Research Part E: Logistics and Transportation Review* 138, 101930.

Yan, Ran, Shuaian Wang, and Lu Zhen (2023). "An extended smart "predict, and optimize"(SPO) framework based on similar sets for ship inspection planning". In: *Transportation Research Part E: Logistics and Transportation Review* 173, 103109.

Yan, Ran, Shuaian Wang, Lu Zhen, and Gilbert Laporte (2021). "Emerging approaches applied to maritime transport research: Past and future". In: *Communications in Transportation Research* 1, 100011.

Yu, Jingjing, Guolei Tang, Xiangqun Song, Xuhui Yu, Yue Qi, Da Li, and Yong Zhang (2018). "Ship arrival prediction and its value on daily container terminal operation". In: *Ocean Engineering* 157, 73–86.

Zhai, Deqing, Xiuju Fu, Xiao Feng Yin, Haiyan Xu, and Wanbing Zhang (2022). "Predicting berth stay for tanker terminals: a systematic and dynamic approach". In: *arXiv preprint arXiv:2204.04085*.

Zheng, A. and A. Casari (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly. ISBN: 9781491953242. URL: https://books.google.co.jp/books?id=Ho0UvgAACAAJ.

Zhou, Zhi-Hua (2021). *Machine Learning*. Singapore: Springer Nature.

A Theoretical introduction of prediciton model

A.1 Detailed construction of an CART regression model

CART, standing for Classification and Regression Tree, is a methodology used for constructing both classification and regression trees. This algorithm is capable of generating models for both types of trees. In our project, we are focusing on regression problems, and the CART model serves as a foundational element for the subsequent XGBoost model. This subsection will delve into the specific composition of the CART regression model, exploring how it functions and its application in our regression-focused project. Suppose our dataset D has n samples and each sample has m features: $D = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}, \mathbf{x}_i \in R^m, y_i \in R$, where $\mathbf{x}_i$ is the vector of features for sample i and y_i is the label or the corresponding real target value.

The process of splitting the tree begins at the root node. Initially, a feature m_i is chosen along with its corresponding value s_i. This pair, represented as (m_i,s_i), serves as a potential splitting point. Utilizing this point, the entire dataset D is divided into two distinct sub-regions, R_1 and R_2 and these two areas are expressed as: $R_1 = \{y_i \mid x_{i,m_i} \leq s_i\}, R_2 = \{y_i \mid x_{i,m_i} > s_i\}$. In the CART model, the average target values for all samples in regions R_1 and R_2 are determined and assigned as the predicted targets for the samples in their respective regions. The effectiveness of each candidate splitting point is evaluated by calculating the sum of the MSE for both R_1 and R_2. The algorithm iterates over all features and their corresponding values to generate various candidate splitting points. The aim is to identify the splitting point that results in the lowest combined MSE for the two regions. This optimal pair, which minimizes the sum of MSE, is then chosen as the final point to split the current node. These steps are repeated for each node until a pre-defined stopping condition for halting the growth of the tree is met. This systematic approach ensures the development of an efficient and accurate model (Lewis, 2000).

To convey the process of constructing a CART regression tree more precisely, we can present it in mathematical terms as follows:

1. Beginning at the root node, a feature-value pair (m_1, s_1) is chosen to divide the dataset D into

two distinct regions, R_1 and R_2.

2. Calculate the mean targets of samples in the two sub-regions R_1 and R_2, which are denoted by C_1 and C_2, respectively:

$$c_1 = \frac{1}{n_1} \sum_{x_i \in R_1} y_i, c_2 = \frac{1}{n_2} \sum_{x_i \in R_2} y_i. \tag{4}$$

3. Iterate through all features d_i and their corresponding values s_i to identify the split pair that minimizes the loss function in Equation (5). The optimal split pair of the current node is denoted as (m^*, s^*).

$$(m^*, s^*) = \min_{m_i, s_i} \left\{ \min_{x_i \in R_1(m_i, s_i)} (y_i - c_1)^2 + \min_{x_i \in R_2(m_i, s_i)} (y_i - c_2)^2 \right\}. \tag{5}$$

4. Use the optimal splitting pair (m^*, s^*), the samples are divided into two new areas. These two new areas are defined as follows:

$$R_1(m^*, s^*) = \{y_i \mid \boldsymbol{x}_{i,m_i} \leq s^*\}, R_2(m^*, s^*) = \{y_i \mid \boldsymbol{x}_{i,m_i} > s^*\}. \tag{6}$$

5. Carry out steps 1 and 2 on each node until any one of the pre-defined tree growth conditions is met, and no further splitting of nodes is allowed. The nodes in the final layer become leaf nodes. As a result, the entire training set is divided into K regions, $R_1, \ldots\ldots R_k$, where k also represents the number of leaf nodes. The model generated from the iterations can be written as:

$$f(\boldsymbol{x}) = \sum_{i=1}^{K} c_i \mathbb{I}(x \in R_i), \tag{7}$$

where $k = 1, \ldots, K$, and $\mathbb{I}$ is an indicator function with the following form:

$$\mathbb{I} = \begin{cases} 1 & \text{if } (\boldsymbol{x} \in R_k) \\ 0 & \text{if } (\boldsymbol{x} \notin R_k). \end{cases} \tag{8}$$

A.2 Detailed construction of an XGBoost model

In regression task, XGBoost comprises of K basic CART regression models and functions cumulatively to predict outcomes. We can represent the model's output as:

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{t=1}^{K} f_t(\mathbf{x}_i), f_t \in F, \tag{9}$$

where $\hat{y}_i$ is the predicted value by the XGBoost, f_t is the t-th basic CART tree, F is the set of functions for all K CART trees. And the training loss can be represented as:

$$L = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right). \tag{10}$$

A practical choose for the loss function is the mean squared error (MSE), where:

$$\sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2. \tag{11}$$

The accuracy of a prediction model is determined by both its bias and variance (Zhou, 2021). The bias of the model is represented by the loss function, L, while Ω, is utilized to penalize the complexity of the model and evaluate the variance of the output. In this way, the object function (Obj) of the XGBoost is:

$$Obj = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{t=1}^{K} \Omega\left(f_t\right). \tag{12}$$

The first term in Eq. (12) is the total training loss of n samples in the training set The second term in Eq. (12) represents the sum of the complexities of K trees. The complexity for each individual tree is expressed as follows:

$$\Omega\left(f_t\right) = \gamma T_t + \frac{1}{2}\lambda \sum_{j=1}^{T_t} w_j^2, \tag{13}$$

where γ and λ are pre-set hyperparameters, T_t is the number of leaves for t-th tree and w_j is the weight in the j leave. To minimize the object function in Eq. (12), we cannot directly implement the gradient descent method as is traditionally utilized in boosting model. Instead, we formulate and train the model in an additive approach. Suppose the model after t, $t = 1, \ldots, K$ iterations, the XGBoost model currently has t trees and the prediction value for the i-th sample by the current t trees is:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k\left(\mathbf{x}_i\right) = \hat{y}_i^{(t-1)} + f_t\left(\mathbf{x}_i\right). \tag{14}$$

In this way, we can rewrite our object function in Eq. (12) as:

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{t} \Omega(f_k) \\ &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{t-1} + f_t\left(\mathbf{x}_i\right)\right) + \sum_{k=1}^{t} \Omega\left(f_k\right) \\ &= \sum_{i=1}^{n} \left(y_i - \left(\hat{y}_i^{(t-1)} + f_t\left(\mathbf{x}_i\right)\right)\right)^2 + \Omega\left(f_t\right) + \sum_{k=1}^{t-1} \Omega\left(f_k\right). \end{aligned} \tag{15}$$

Recall that the second order Taylor expansion:

$$f(x+\Delta x) \simeq f(x)+f'(x)\Delta x+\frac{1}{2}f''(x)\Delta x^2. \tag{16}$$

Following the rule, by viewing $\hat{y}_i^{t-1}$ as x and $f_t(x_i)$ as Δx, Eq. (15) can be rewritten as:

$$Obj^{(t)} \simeq \sum_{i=1}^{n}\left[L\left(y_i,\hat{y}_i^{t-1}\right)+g_i f_t(\mathbf{x}_i)+\frac{1}{2}h_i f_t^2(\mathbf{x}_i)\right]+\Omega(f_t)+\sum_{k=1}^{t-1}\Omega(f_k), \tag{17}$$

where g_i and h_i are the first and second order gradients of the Eq. (15): $g_i=\partial_{\hat{y}^{(t-1)}}L\left(y_i,\hat{y}^{(t-1)}\right)=2\hat{y}_i^{(t-1)}-2y_i$, $h_i=\partial^2_{\hat{y}^{(t-1)}}l\left(y_i,\hat{y}_i^{(t-1)}\right)=2$. When training the t-th iteration, as $\hat{y}_i^{(t-1)}$ has already been determined, the first term in Eq. (17): $l\left(y_i,\hat{y}_i^{t-1}\right)$, which is the training loss of the $(t-1)$-th iteration, is a constant. And the last term in Eq. (17): $\sum_{k=1}^{t-1}\Omega(f_k)$, which represents the total penalty complexity of previous $t-1$-th iterations, is also a constant. In this way, we can rewrite the approximated object function in Eq. (17) as:

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^{n}\left[g_i f_t(\mathbf{x}_i)+\frac{1}{2}h_i f_t(\mathbf{x}_i)^2\right]+\Omega(f_t) \\ &= \sum_{i=1}^{n}\left[g_i f_t(\mathbf{x}_i)+\frac{1}{2}h_i f_t(\mathbf{x}_i)^2\right]+\gamma T_t+\frac{1}{2}\lambda\sum_{j=1}^{T_t}w_j^2. \end{aligned} \tag{18}$$

Suppose the sample set in leaf j is defined as:

$$I_j=\{i \mid q(\mathbf{x}_i)=j\}, \tag{19}$$

where $q(\mathbf{x}_i)$ is the given fixed tree structure, we can rewrite Eq. (18) as:

$$Obj^{(t)}=\sum_{j=1}^{T_t}\left[\left(\sum_{i\in I_j}g_i\right)w_j+\frac{1}{2}\left(\sum_{i\in I_j}h_i+\lambda\right)w_j^2\right]+\gamma T_t. \tag{20}$$

As the tree structure $q(\mathbf{x}_i)$ is fixed, $\sum_{i\in I_j}g_i$, $\sum_{i\in l_j}h_i$ and T_t are also fixed. To obtain the optimal w_j^* for leaf j, we can set the first derivative of the objective function to be 0, and the optimal value of w_j^* can be derived as follows:

$$w_j^*=-\frac{\sum_{i\in I_j}g_i}{\sum_{i\in I_j}h_i+\lambda}. \tag{21}$$

By substituting Eq. (21) into the objective function Eq. (20), we obtain the following optimal value of the objective function:

$$Obj^{(t)*}=-\frac{1}{2}\sum_{j=1}^{T_t}\frac{\left(\sum_{i\in I_j}g_i\right)^2}{\sum_{i\in I_j}h_i+\lambda}+\gamma T_t. \tag{22}$$

During the derivation of the optimal value for the object function, we make the assumption that the tree structure $q(\mathbf{x}_i)$ is predetermined. Therefore, our next step involves determining the tree

structure of the XGBoost model under this optimal value. Practically, the XGBoost algorithm implements a greedy method that starts from a single node and progressively adds branches to the tree to form its structure. Suppose I_L and I_R are the samples of the left and right nodes of the tree after splitting and $I = I_L \cup I_R$ is the set of nodes before splitting, then we can write the object function before and after the splitting as:
before splitting:

$$Obj_{L+R}^{(t)} = -\frac{1}{2}\frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda} + \lambda, \tag{23}$$

after splitting:

$$Obj_{L}^{(t)} + obj_{R}^{(t)} = -\frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda}\right] + 2\lambda. \tag{24}$$

Then, the gain of the splitting is expressed as:

$$\begin{aligned} \text{gain} &= obj_{L+R}^{(t)} - \left(obj_{L}^{(t)} + obj_{R}^{(t)}\right) \\ &= \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma. \end{aligned} \tag{25}$$

When splitting the node, we consider all candidates that make the splitting gain in Eq. (25) larger than 0 and select the value of features that corresponding to the largest value of gain in Eq. (25) to split the node.