

Using Statistical and Forecasting Methods to Predict Arrival of Container Shipments

Chethan Manjunath
Krannert School of Management
Purdue University
West Lafayette, USA
cmanjuna@purdue.edu

Keerthana Nemili
Krannert School of Management
Purdue University
West Lafayette, USA
knemili@purdue.edu

Soham Patil
Krannert School of Management
Purdue University
West Lafayette, USA
patil124@purdue.edu

Sreeja Sesa
Krannert School of Management
Purdue University
West Lafayette, USA
sadhu0@purdue.edu

Gauri Vaidya
Krannert School of Management
Purdue University
West Lafayette, USA
vaidyag@purdue.edu

Matthew A. Lanham
Krannert School of Management
Purdue University
West Lafayette, USA
lanhamm@purdue.edu

Abstract— In today's global economy, organizations commonly procure goods and materials from multiple vendors across the globe through sea-going vessels. Although carriers in charge of shipments provide their own estimated time of arrivals, they are often inaccurate and provide an expected time of arrival (ETA) with larger variance. Moreover, multiple ETAs from multiple carriers degrades the decision-making of organizations causing disruptions to subsequent activities in the chain. In partnership with an award-winning container shipment tracking aggregator, we designed and tested several approaches to effectively use these noisy carrier predictions in an ensembled fashion to generate better vessel ETA predictions. Among several experiments we found that a simple . The impact of this work will not only improve the value the partner provides their clients but will help their clients plan better which will positively impacting P&L and the planet.

I. INTRODUCTION

In the modern global economy, container shipments play a critical role in the transportation of goods and the functioning of supply chains. The estimated time of arrival (ETA) of container shipments is an important factor in the effective and efficient management of the supply chain and the organizations involved in it. Accurate predictions of the ETA of container shipments can bring numerous benefits to organizations, including improved planning and scheduling, enhanced customer satisfaction, reduced costs, and a more sustainable supply chain.

Accurate predictions of the ETA can help organizations to plan and schedule their operations more effectively. This includes the management of inventory levels, production processes, and customer deliveries. With accurate ETA predictions, organizations can ensure that they have the necessary inventory to meet customer demand and avoid stockouts, minimize disruptions, and minimize the costs associated with excess inventory. Another important benefit of predicting the ETA of container shipments is the impact it has on customer satisfaction. Late deliveries can result in dissatisfaction among customers, damaging the

organization's reputation and potentially leading to lost business. Accurate ETA predictions allow organizations to provide customers with more accurate delivery dates, improving customer satisfaction and building customer trust.

The transportation of goods in container shipments also has a significant impact on the environment. The shipping industry is responsible for a significant amount of greenhouse gas emissions, air pollution, and ocean acidification. Accurate ETA predictions can play a role in reducing the environmental impact of container shipments by minimizing disruptions and improving the efficiency of the supply chain. For example, reducing the amount of time containers spend waiting at ports can help to reduce the amount of greenhouse gas emissions generated by the shipping industry. Accurate ETA predictions can also help organizations to optimize the routing of shipments, reducing the distance containers need to travel and reducing the environmental impact of transportation. In addition, by reducing the need for expedited shipping and reducing the amount of time containers spend in transit, organizations can reduce the energy consumption and carbon emissions associated with container shipments.

Shipping companies publicly advertise their vessels' estimated time of arrival (ETA) in port, and downstream supply chain activities are planned accordingly. However, delays often occur, and the ETA might differ from the vessel's actual time of arrival (ATA), for instance due to technical or weather-related issues. This impacts the entire supply chain, in many instances reducing productivity and increasing waste and inefficiencies.

Predicting the exact time, a vessel arrives in a port and starts off-loading operations poses remarkable challenges. Today, a majority of companies rely on experience and improvisation to respectively guess ATA and cope with its fluctuations. Very few providers are leveraging machine learning (ML) techniques to scientifically predict ETA and help companies create better planning for their supply chain.

This study focuses on the use of statistical and analytical techniques to understand and provide and aggregate prediction considering multiple ETAs from multiple carriers for a give vessel. Businesses need a firm commitment for supplies shipped through large vessels to

plan their downstream activities such as manufacturing, distribution, customer services etc. The complexity of ocean shipment makes it hard for carriers to give an accurate ETA. Moreover, due to multiple carrier predictions, there are different ETAs for the same vessel.

The aim of this project is to:

- Understand the container shipment, draw insights, and provide an overview of factors influencing the prediction.
- Create a model of an ensemble system that, based on current predictions made by various carriers, can anticipate the shipment arrival date that is closest to reality.
- Improve the accuracy of forecasting of the shipment arrival to increase consumer confidence.

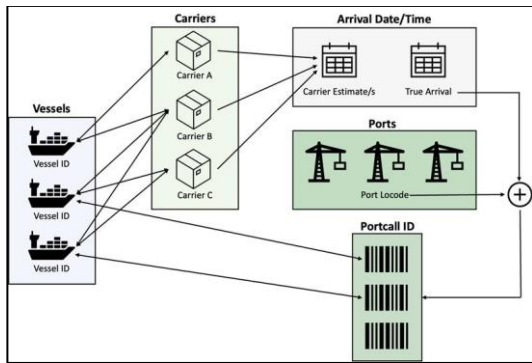


Fig 1: Data used

II. LITERATURE REVIEW

Ship container tracking is an essential aspect of modern logistics and supply chain management. The use of technology has enabled companies to always monitor the location of their shipping containers, providing greater visibility and security for their cargo. Several studies have been conducted on this topic, using methods such as logistic regression, neural networks, and hierarchical clustering to identify key factors that influence on-time delivery, and to predict lead times. In addition to this, studies have been conducted on topics such as on-time delivery prediction, major factor identification, and on-time reliability.

Predictive Modelling – Machine Learning Techniques

A study by Rong et al., (2019) proposes an Automatic Identification System (AIS) data-driven method for estimating the arrival time of sea-going vessels in port operations. AIS is a system of vessel data exchange that was made mandatory by the International Maritime Organization (IMO) in 2004 Serry (2017). The method uses Reinforcement Learning and the Metropolis-Hastings algorithm to find the optimal vessel trajectory and estimate the speed over ground, resulting in improved accuracy of arrival time predictions.

Another study conducted by Schramm & Munim (2021) presents a novel method for forecasting container shipping freight rates by incorporating practitioner sentiment, confidence, and perception measures. The study compares the performance of an ARIMA model with ARIMAX and Vector Autoregressive (VAR) models and finds that the integration of the Logistics Confidence Index in the

ARIMAX model leads to substantial improvement in forecast accuracy. The proposed methodology can be extended to other trade routes and shipping markets.

The article "Predictive Model for Estimating Shipment Lead Time in Ocean Import Freight" by Hatikal et al., (2020) focuses on finding a solution to improve visibility and predictability of shipment lead times in ocean import freight. The authors use real-world data and various machine learning techniques to develop a predictive model that can be used by different stakeholders in the supply chain. The study finds that multinomial logistic regression is the best classifier for predicting shipment lead times, with decision trees and other commonly used classifiers also performing well. The authors conclude that their proposed model has the potential to improve supply chain efficiency and reduce costs by providing improved visibility and predictability of shipment lead times.

Fleet management (FM) is a combination of data logging, satellite positioning, and data communications used to manage commercial vehicle fleets. The report by Berg Insight on Fleet Management (FM) in the Americas highlights the growth of FM solutions in North and Latin America, driven by regulatory developments and increased demand for optimization functionality. The report notes that the FM market in the Americas is expected to grow, with an estimated CAGR of 14.0% in North America and 13.1% in Latin America by 2025. The report also mentions the leading FM providers in the Americas, including Geotab, Verizon Connect, Omnitracs, and others, and notes that the Original Equipment Manufacturer (OEM) channel is expected to increase in importance in the coming years.

The article on IBM Cloud Monitoring with Sysdig highlights the challenges of limited visibility and security in hybrid multi cloud environments. The authors describe IBM Cloud Monitoring with Sysdig as a fully managed container monitoring service that provides end-to-end visibility and security for containers, reducing risk and improving performance. The authors note that IBM and Sysdig have partnered to address the need for better visibility and security in cloud-native applications, and that IBM Cloud Monitoring with Sysdig can help organizations accelerate diagnosis and resolution of performance and security incidents, better manage access to data, and troubleshoot applications and infrastructure (Preetham Kumar, 2017).

The report by Berg Insight on Fleet Management (FM) in the Americas highlights the growth of FM solutions in North and Latin America, driven by regulatory developments and increased demand for optimization functionality. The report notes that the FM market in the Americas is expected to grow, with an estimated CAGR of 14.0% in North America and 13.1% in Latin America by 2025. The report also mentions the leading FM providers in the Americas, including Geotab, Verizon Connect, Omnitracs, and others, and notes that the Original Equipment Manufacturer (OEM) channel is expected to increase in importance in the coming years.

Overall, these studies summarized above highlight the potential of machine learning techniques to improve efficiency and predictability in supply chain and logistics operations. These previous studies provide valuable context for research in the field of supply chain

management and logistics, cloud-native applications, and fleet management. The articles highlight the importance of improving visibility and predictability in these industries and the potential benefits of better monitoring and management systems.

Table 1: Literature review

Author/Source	Supply chain	Ocean-shipment	ML model
Berg Insight AB	x		
IBM, Preetam Kumar	x		
Saraswathi Hathikal, Sung Hoon, Martin Karczewk	x	x	x
Hans-Joachim Schramm , Ziaul Haque Munim	x		x
Kikun Park, Sunghyun Sim, Hyerim Bae	x	x	x
Samir Araujo and Michele Sancricca	x	x	

III. DATA

We collaborated with a technology platform company who aggregate shipment details and predictions made by multiple carriers for various organizations to conduct our research. We were provided with a part of their enormous shipment data from multiple vessels and carriers which was used for our research purpose. The data dictionary is included for illustration purposes.

Table 2: Data description

Field	Type	Description
VESSEL_ID	string	Identity representing vessels (ships)
PORTCALL_ID	string	Represents an instance of vessel arrival at a port
CARRIER_ID	string	Unique identifies of carriers
CARRIER_ETA	Date time	ETA provided by carriers
PRED_CREATED	Date time	Date on which the prediction was made
TRUE_ARRIVAL	Date time	Actual arrival time
PORT_LOCDOCE	string	Name of the ports being serviced

VESSEL_NAME	string	Name of the vessels (Ships)
CAP_TEU	int	Capacity of each vessel (Ship)

IV. METHODOLOGY

The aim of the project is to develop a model which gives a more accurate estimated time of arrival (ETA) of the vessel taking inputs from various carrier predictions. The problem-solving approach begins with basic exploratory data analysis where we arrived at some important attributes which will be guiding the model. These attributes include carriers, their predicted ETAs, the dates on which prediction was created and the actual arrival date of the vessel to calculate error. Next step is the pre-processing part where we derive some of the variables that will be used as inputs to the model. These variables are error in number of days which is basically the difference between the actual arrival date and the predicted ETA. Another derived variable is a time window. Because the accuracy of prediction depends upon how far the prediction created on date is in the timeline from actual ETA, a time window will also have impact on the model accuracy. To evaluate model performance, we finalized the accuracy measurement metric as the percentage of predictions which were within 24 hours of the actual arrival date. Next step is modelling with 70%-30% data spit into training and test data sets, to arrive at the best performing model. The best model is then validated on the test data set in the evaluate phase and once results were acceptable, the implementation plan was developed.

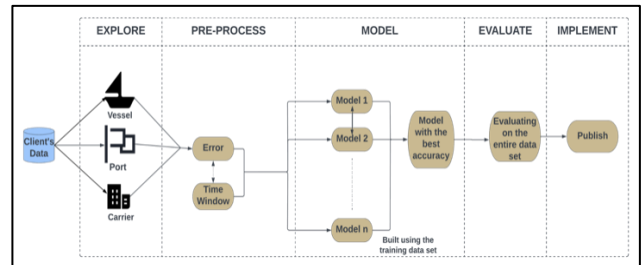


Fig 2: Methodology

V. MODELS

From the data analysis, following were the observations regarding the distribution of errors:

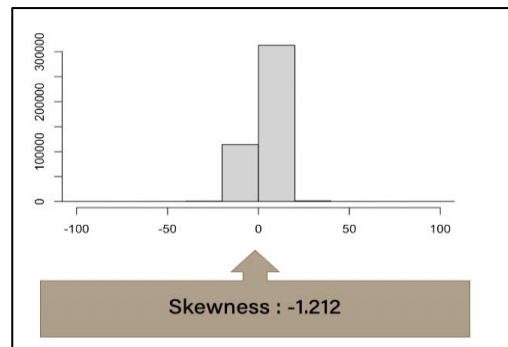


Fig 3: Distribution of errors

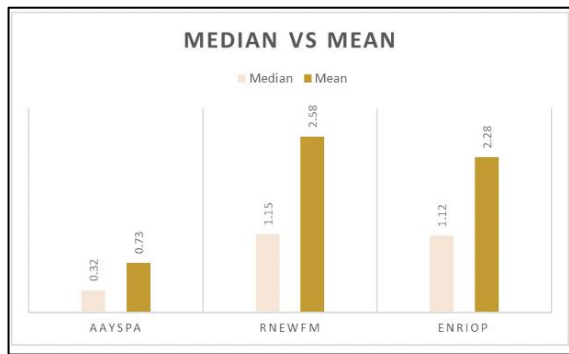


Fig 4: Distribution of errors for 3 carriers

The error distribution was highly left skewed with skewness of -1.2 therefore, median error was chosen as the input to the model rather than mean error. Another observation was the accuracy of predictions improves as the ETA approaches therefore, time window along with the carrier is considered as other input. So, the final model inputs are:

- Creation of carrier + time window pairs
- Pred_created_on = Prediction created on date
- Carrier ETA = Predictions made by carriers
- % accuracy = Percentage accuracy for each pair
- Error = Median error for each pair
- Days = Carrier ETA - Pred_created_on

Final model: The main concept underlying the final model implementation is correction of the inaccurate predictions by adding the median error to them. For illustration consider following example of a carrier with 40% accuracy, making 10 predictions. 4 out of these 10 will be correct predictions and for rest of the 6 predictions, median error will be needed to be added to each of them. Summation of both these quantities will provide the new days that need to be added to the prediction created on date to get the new vessel ETA.

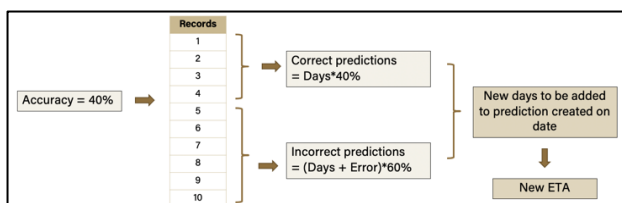


Fig 5: Final Model Calculations

Final model formula:

$$\text{New Days} = (\% \text{ accuracy} \times \text{days}) + (1 - \% \text{ accuracy}) \times (\text{Days} + \text{median error})$$

$$\text{New ETA} = \text{Pred_created_on} + \text{New Days}$$

VI. RESULTS

The proposed model gives overall accuracy improvement of $\sim 3\%$. Following figure shows simulation results on 1 port call and the black line corresponding to the proposed model is the closest to the ideal green colored line.

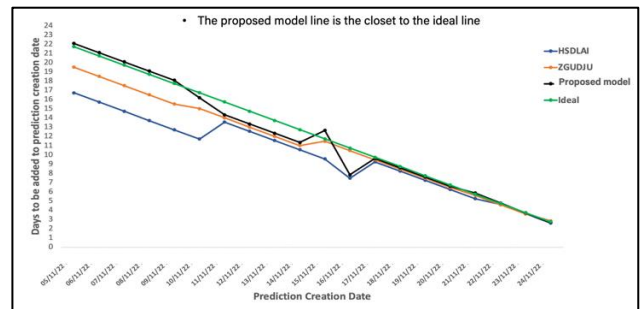


Fig 6: Model comparison

Further on evaluation of model performance across all 392 port locations, it was observed that proposed model performs better than or as good as the existing carrier predictions for 72% of the port locations. However, for 28% of the ports model accuracy is poor with average difference of 7.7% and standard deviation of 9.7% from the carrier prediction accuracies.

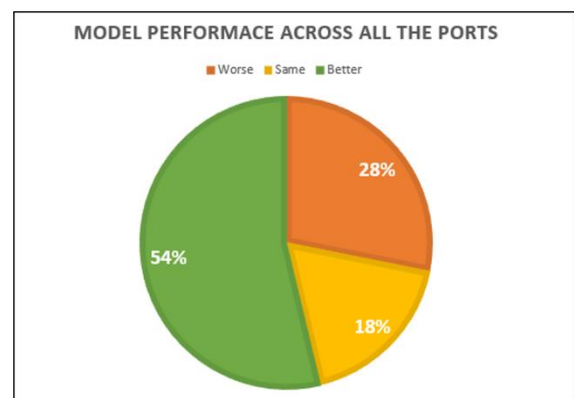


Fig 7: Model performance analysis

From the model performance analysis across time windows, it was observed that model can be utilized the best from time windows $t-15$ till $t-5$, which is a critical time for planning activities further down the supply chain.

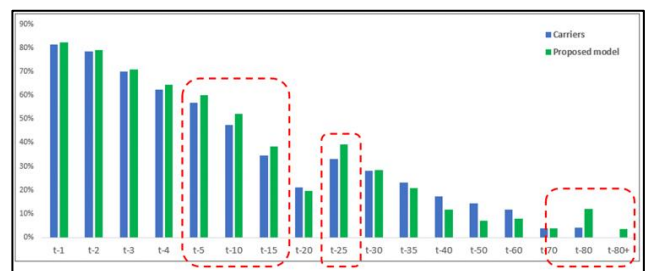


Fig 8: Model performance across time windows

VII. CONCLUSION

Implementation of model on entire data set resulted in 3% accuracy improvement. However, using this model in the time windows $t-4$ to $t-1$, does not result in significant improvement because of the proximity to the actual arrival date in the timeline, where even the existing carrier predictions are fairly accurate. The model can be best utilized from the time windows $t-15$ till $t-5$, where accuracy improvement of up-to 6% can be achieved. This can result in huge cost savings related to the activities down the value stream. To illustrate this, take example of one of the customers having monthly revenue generation of \$160 Million. An

accuracy improvement of 6% would result in \$ (160/30) * 6 million in 100 days, which converts to potential daily savings of up-to \$ 0.3 Million. This is a huge impact which would result in higher customer satisfaction and would generate a competitive edge. To conclude, implementation of this model in predicting vessel ETAs would result in

- Increased operational efficiency: Getting accurate ETAs will positively impact operational efficiency as it would enable efficient planning of the projects, reduction in idle times and more efficient resource utilization.
- Better overall prediction accuracy: A model-provided estimate is more accurate as compared to multiple carrier-provided estimates. Moreover, a single estimate is better than multiple estimates to plan downstream processes. Our model has a 3% improvement over current predictions provided by carriers.
- Cost savings: Up-to 3 times of the operational costs could be potentially saved every 100 days.
- Increased reliability: A single and reliable ETA can be provided at any given point in time, irrespective of whether the carrier has made a prediction on that day.

REFERENCES

- [1] Bäckman, M. (2021). *Trailer and Cargo Container Tracking - 9th Edition*. Berg Insight AB. https://www.researchandmarkets.com/reports/5458937/trailer-and-cargo-container-tracking-9th-edition?utm_source=CI&utm_medium=PressRelease&utm_code=n5bkbv&utm_campaign=1611553+-+Global+Trailer+and+Cargo+Container+Tracking+Market+Report+2021-2025+with+Profiles+of+100%2b+Cargo+Container+Tracking+Solution+Providers&utm_exec=chdo54prd
- [2] Kumar, P. (2017, July). *Going with the flow: exploring real-world use cases for real-time streaming analytics*. IBM Blogs - Cloud Archive. <https://www.ibm.com/blogs/cloud-archive/2017/07/going-flow-exploring-real-world-use-cases-real-time-streaming-analytics/>
- [3] Hathikal, S., Chung, S. H., & Karczewski, M. (2020). *Prediction of ocean import shipment lead time using machine learning methods*. *SN Applied Sciences*, 2(1272), 1-12. <https://link.springer.com/article/10.1007/s42452-020-2951-5>
- [4] Schramm, H.-J., & Munim, Z. H. (2021). *Container freight rate forecasting with improved accuracy by integrating soft facts from practitioners*. *Research in Transportation Business & Management*, 41, 100662. doi:10.1016/j.rtbm.2021.100662. <https://www.sciencedirect.com/science/article/pii/S2210539521000456>
- [5] Park, K., Sim, S., & Bae, H. (2021). *Vessel estimated time of arrival prediction system based on a path-finding algorithm*. *Maritime Transport Research*, 2, 100012. doi: 10.1016/j.martra.2021.100012. <https://www.sciencedirect.com/science/article/pii/S2666822X21000046>
- [6] Araujo, S., & Sancricca, M. (2021, January 13). *Using machine learning to predict vessel time of arrival with Amazon SageMaker*. AWS Machine Learning Blog. <https://aws.amazon.com/blogs/machine-learning/using-machine-learning-to-predict-vessel-time-of-arrival-with-amazon-sagemaker/>