Predicting Shipping Time with Machine Learning

By: Antoine Jonquais, Florian Krempl
Advisor: Dr. Roar Adland, Dr. Haiying Jia

Topic Areas: Machine Learning, Transportation

Summary: This research project focused on how to apply Machine Learning to make predictions regarding shipping times between South East Asia and North America, from factory to port of destination. Using a Random Forest algorithm and building four models to produce estimates at each step of the shipping process, we were able to build a functional tool that yields superior results compared to more traditional methods.

Prior to coming to MIT, Florian worked at LKW Walter managing full truck loads for Amazon and Yusen Logistics all over the EU. In his Bachelor program at the University of Economics in Vienna, he specialized in Logistics and Finance.

Before being a student at MIT, Antoine worked as an account analyst at Hasbro and a logistics project manager in France.
He holds a Master's degree in Logistics Engineering from ISEL - Université Le Havre Normandie.

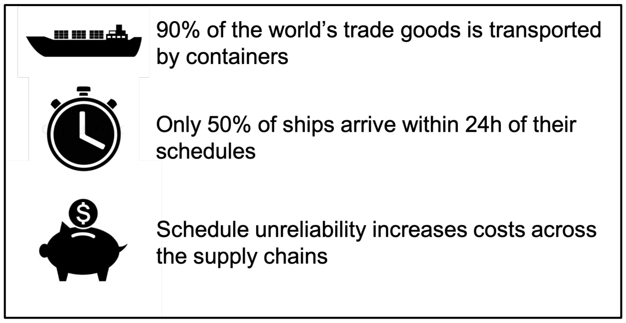
KEY INSIGHTS

1. **There is a lot of potential in the area of predictive analytics when data is available and impactful factors are identified.**
2. **Machine Learning is a good tool to predict transit times and the models will become more accurate by learning from more data.**
3. **Our approach is replicable on a larger geographic scale and the methodology can be used for any mode of transportation.**

Introduction

With the globalization of trade, maritime shipping has become a key component of international trade. As a consequence, container vessels' route scheduling reliability has become a critical point in the shipping industry as irregularities will lead to more delays further down the supply chain and will drive up the overall cost of container shipping.

The freight forwarding arm of Maersk has developed a tool, Harmony, to help the decision-making process when it comes to organizing transportation for a shipment. It uses historical data to provide the user with an average and its associated deviation of the elapsed time between transportation booking and delivery at destination. This tool can be classified as a descriptive tool as it does not make any recommendation, but simply provides statistics regarding past data.
Using Machine Learning computing, we developed a model capable of improving the predictability of shipping times. Our model delivers predictions with a 90% accuracy (measured on the mean absolute percentage errors of days) as early as the transport is booked. The model relies on historical data (for example, transit time statistics of carriers) but also on external sources of data for the most impactful factors (such as port congestion or holidays) regarding schedule reliability for container vessels.



90% of the world's trade goods is transported by containers

Only 50% of ships arrive within 24h of their schedules

Schedule unreliability increases costs across the supply chains

Figure 1: Facts about the shipping industry

We researched drivers of variability in shipping transit times and we tested several Machine Learning algorithms in order to select the one that minimizes the predictions' error margin. This algorithm, coupled with Harmony in its present state, introduces a prediction component to the output of the tool by giving the user an estimated date of unloading at the destination port for a shipment.

Methodology

The datasets we were given contained a great number of route-carrier combinations. After consulting with Maersk, we decided to focus our study to shipments loaded in South East Asia (China, Hong Kong and Vietnam) and unloaded in the United States. After defining this scope and dropping observations without an unload date, the final dataset used is composed of 1,744,278 shipment records spread across 74 distinct routes. The shipments are moved by 31 different carriers and 2,997 unique shippers.

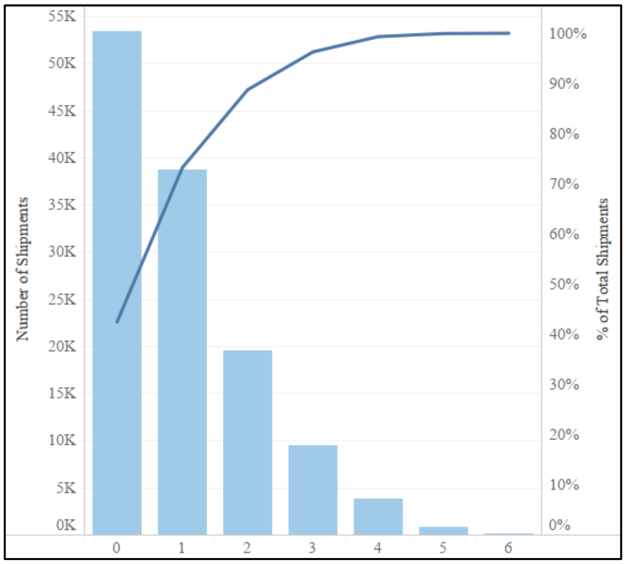


Figure 2: Distribution of time (in days) spent by containers waiting to be unloaded from their vessel in the port of Los Angeles

We then selected the relevant external factors to tie into our model. Chinese New Year and port congestion at the port of destination were the only relevant factors with data available to us. Chinese New Year helps considering the slowdown in activity around that period. Port congestion helps modelling whether a vessel will have difficulties or not to berth on time at the port of destination.
Overall, we chose 14 different features to be extracted from the historical data and to be used in the model.

To test and validate the performance of the different algorithms, we split the data in a training set and a test or validation set. The training set consist of the first two years of data and was used to train the models and select the best hyperparameters. The test set is the data from the most recent year in our dataset and was used to validate the performance of our models. All algorithms were trained and tested on the same training and test set so that we could compare their performance.

We used Python programming language to write the code for our models. To create, train and use the final models for prediction, there are three relevant scripts in the final program. The first script is used to clean the data, remove missing values and transform all columns to the right format. The second script automatically reads in all the relevant data, trains (on new data or adding data to an existing model) and saves the final model. The third script is used for prediction in a production environment. The output delivers an estimated date of unloading at the port of destination with a 90% confidence interval with the earliest and latest unloading dates.

Results and Analysis

After comparing the results yielded by three different algorithms (Random Forest, Neural Network and Linear Regression), we decided to build the final prediction model with a Random Forest algorithm.
We selected the set of Random Forest models as they result in the highest accuracy (ranging from 2 to 3.75 days on the mean absolute error, depending on the model segments), compared to the baseline and Neural Network (see figure 3).

Not only does the Random Forest model performs better on predicting the target variable, it is also easier to train and implement in a production system. The time required to train the Random Forest model is less than for the Neural Network. Furthermore, the model is also not a black box like the Neural Network. It is thus easier to explain to stakeholders what a Random Forest does compared to a Neural Network. Neural Networks perform very well on very complex nonlinear problems. In this project the Random Forest performs better, probably because there is not enough complexity in the data for the Neural Network to really shine.

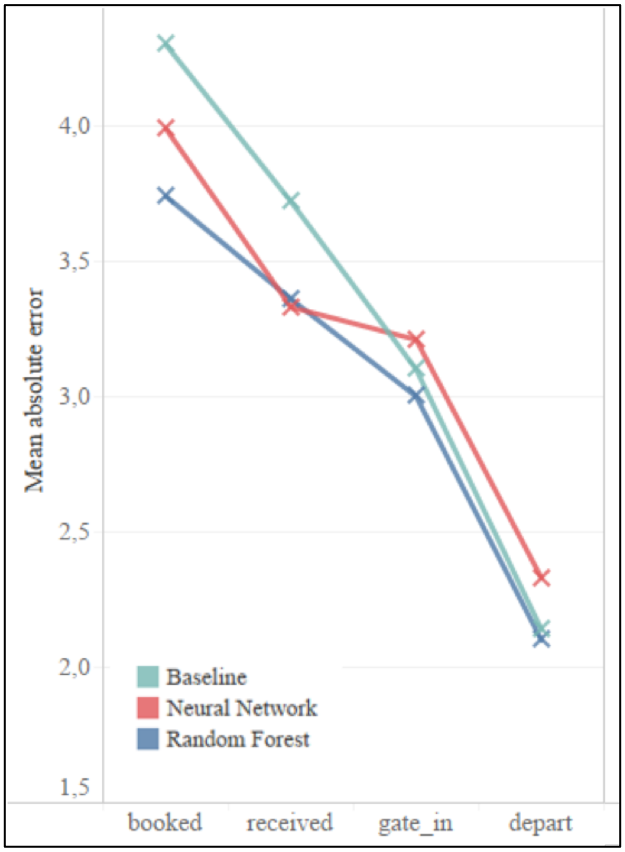


Figure 3: Performance of different algorithms tested for the models

Conclusions

Shipping goods across the world will always involve variability in some ways. This is not a deterministic environment. What we sought to accomplish with this project was not to eliminate this variability but rather mitigate its effects on the supply chains it impacts through better prediction of time spent in transit.

Given the encouraging results we have obtained through our model, we can say that using Machine Learning to predict an ETA for a shipment is a valid use of this computing discipline. The accuracy of the models get better as the cargo gets closer to the port of destination. However, we also found our models only perform better than more traditional approaches when variability is high. Our model *depart* demonstrates a performance similar to a simple historical transit-time average of the carrier. The model *depart* predicts the last leg of the trip with the fewest days left on the trip and the least variability.

In this study, we limited our scope to the South East Asia to North America routes but the same models developed for this purpose could be used for any routes, given that the models are trained and supplied with the appropriate data. As a consequence, the relevant factors we have identified as impactful for transit times, such as port congestion or time of the year in these models, are relevant for any part of the world.

Furthermore, another insight of our study is that our approach could be applied to any shipping industry. We could envision a similar project being conducted for the trucking industry for example. As long as the necessary data is available and the impactful factors can be identified, the method can be adapted for any shipping route in the world.

In the future, it would be worth investigating the possibility of including weather patterns in the model as a way to improve its accuracy, if access to reliable data can be guaranteed. Another way to improve the model would be to know the position of the container on the ship. In interviews with industry professionals, we found out that there are premium spots for containers on the vessels. The position of the container determines when it gets loaded and unloaded from the ship. We are confident that this would increase the accuracy of the model even more.

With industry 4.0 approaching rapidly, data collection and its use in predictive analytics is a must for every company that wants to be an industry leader.

Our model can be used as a prototype when developing a complete predictive analytics tool that will cover the whole journey of a shipment.