Predicting Shipping Time with Machine Learning

by

Antoine Charles Jean Jonquais

Diplôme d'Ingénieur, Logistique, Université le Havre Normandie – Institut Supérieur d'Etudes Logistiques, 2012

and

Florian Krempl

Bachelor of Science in Business, Economics and Social Sciences, Wirtschaftsuniversität Wien, 2016

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2019

Signature of Author: _____

Department of Supply Chain Management
May 10, 2019

Signature of Author: _____

Department of Supply Chain Management
May 10, 2019

Certified by: _____

Dr. Haiying Jia
Senior Researcher
Capstone Advisor

Certified by: _____

Dr. Roar Adland
Professor
Capstone Co-Advisor

Accepted by: _____

Dr. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

Predicting Shipping Time with Machine Learning

by

Antoine Charles Jean Jonquais

and

Florian Krempl

Submitted to the Program in Supply Chain Management
on May 10, 2019 in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

With the globalization of trade, transit time reliability has become a critical point in the shipping industry as irregularities will lead to more delays further down the supply chain. Our sponsoring company, A.P. Møller – Mærsk A/S (Maersk) provides freight forwarding services to its clients, offering them a complete set of supply chain solutions for shipping their goods across the world. Currently Maersk uses an in-house tool, Harmony, which provides descriptive analytics for shipment times and their variations based on historical distributions. However, Maersk is facing commercial pressure from its customers for a better estimation of its shipment transit time reliability, which has become a key measurement of its operational performance. The goal of our project was to determine whether Machine Learning and predictive analytics can improve the estimated time of arrival for a shipment. Using Machine Learning computing, we developed a model capable of predicting shipping times by training the algorithms on historical shipment data, and incorporating external sources of data related to the most impactful factors regarding schedule reliability (e.g. holiday seasons and port congestion levels). We found that Machine Learning in this instance might be a partial answer to this problem, as it performs better on long lead time than on short lead time when comparing to more classical approaches. Our model has a mean absolute error (MAE) of 3.74 days when making a prediction at the time of booking transportation whereas our baseline model (which only considers historical average transit times on a shipping lane) predicts with a 4.3 days MAE at the same time. When making a prediction at the time the vessel leaves the port of origin, the two models actually perform similarly, with a MAE of 2.1 days for both.

Capstone Advisor: Dr. Haiying Jia
Title: Senior Researcher

Capstone Co-Advisor: Dr. Roar Adland
Title: Professor

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. 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 (Chung and Chiang, 2011).

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.

Maersk's Supply Chain Development department is determined to use the latest progress in supply chain technology to enhance this tool's capabilities. It wishes to integrate Machine Learning capabilities and apply them to improve prediction of cargo arrival time at destination by combining historical and contextual data.

Using Machine Learning computing, we developed a model capable of improving the predictability of shipping times. Our model is 15% more accurate than a naïve mean model and 10% more accurate than a linear regression model for long lead times (measured on the mean absolute error in days). 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.

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.

2. LITERATURE REVIEW

This chapter reviews the literature we consulted to help us answer the question posed by our research project: how can we use Machine Learning to improve prediction of cargo arrival time at destination port by combining historical transit variability and data pertaining to external relevant factors?

We researched the relevant literature on the maritime shipping industry in order to find out what factors are generally accepted as main drivers of variability in cargo transit time, how to handle big data, the associated data mining and where to find publicly available data in a transportation context, and finally, get a better view on the Machine Learning methods that exist and their different uses (see Figure 1).
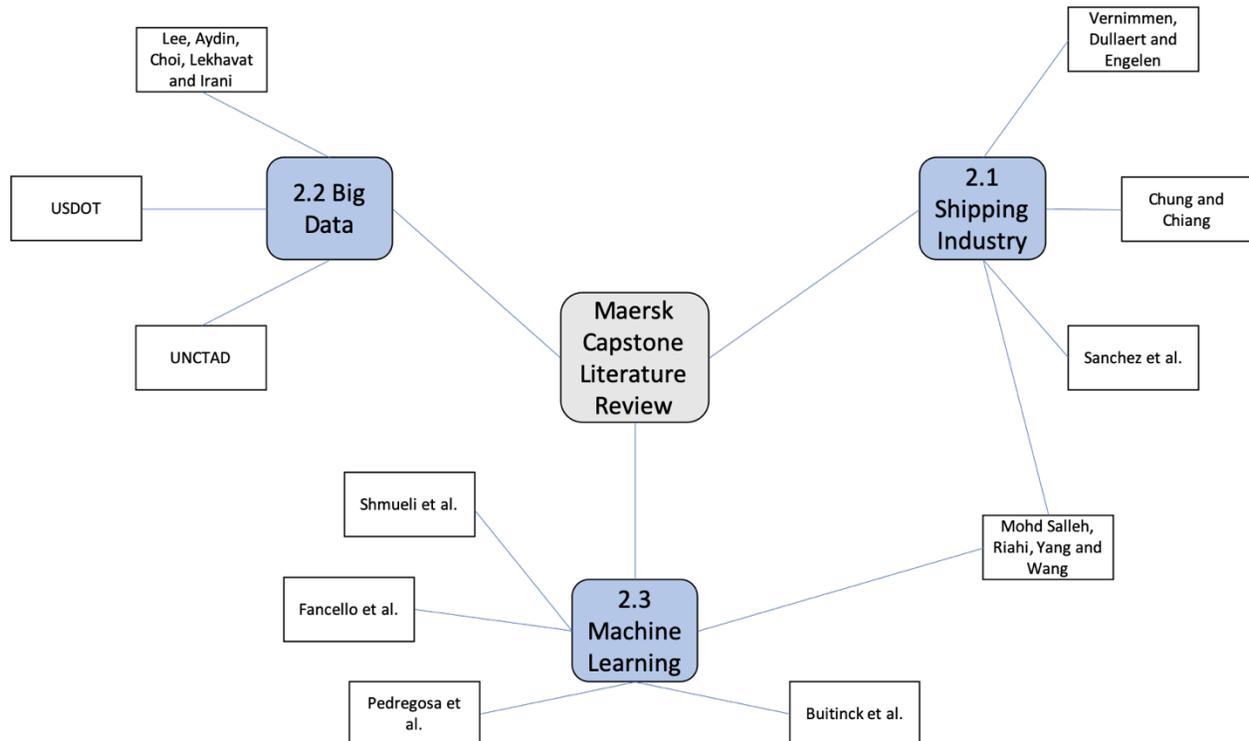
Figure 1: Map of literature sources

2.1. Maritime Shipping Industry

Our study has been primed mainly by four papers, each talking about different subjects in relation to the maritime shipping industry. Chung and Chiang (2011) investigated the consequences on hinterland supply chains of schedule reliability for container liners. It showed that if these liners are not on time, this will have an operational effect on the downstream supply chain and will increase the overall cost of shipping for the shipper. This publication helped us to assert the importance of schedule reliability in container shipping.

Sanchez, Hoffman, Micco, Pizzolitto, Sgut and Wilmsmeier (2003) explored various port performance indicators (PPI), such as terminal turnaround or loading and unloading of a container or productivity of the terminal. These PPI were used to build a model to evaluate the cost of maritime transport. They used the Principal Component Analysis method to build their model off of these PPI. It was found that all the PPI relating to port congestion are part of the resulting first component, which explains over 40% of the total variance of the model. We used this paper to validate our hypothesis about port congestion being a relevant factor in schedule reliability.

Vernimmen, Dullaert, and Engelen (2007) addressed the importance of on-time cargo shipping for an entire supply chain. It also looked at potential factors driving variability in transit time for container liners and separated them in two groups: the factors that are within the control of the shipping company (for example, not planning buffer time to account for contingencies in their weekly schedule) and those out of the company's hands (such as weather, strikes etc.). This publication was a valuable resource to select factors impacting ship schedule reliability.

Mohd Salleh, Riahi, Yang and Wang (2007) shed light on the shipping industry and the factors that may drive variability in transit time. They categorized the critical factors identified as influential regarding arrival punctuality in two overarching groups: port conditions (such as tidal window and weather conditions at port, inland corridors…) and vessel conditions (like speed, crews' reliability etc.). It also detailed the model they used to predict a container's arrival, using a Fuzzy Rule-Based Bayesian Network (FRBBN), which we will talk about more in details in section 2.3.

2.2. Big Data

We found some datasets coming from big data in the context of port and container ship operations. However, the main point we have learned from our search is that the data publicly available is scarce and when it is available, it is not necessarily up to date. This is a theme that is well documented in the Review of Maritime Transport 2016, UNCTAD (2016).

The main challenge in this study is the big data part. As described in the UNCTAD publication, port data (even though ports are often operated by state organizations) is scarce because it is seen as a competitive advantage not to disclose them. Moreover, shipping lines data are also hard to obtain since they are operated by private companies that do not disclose their operational performance indicators.

Nonetheless, it is still possible to collect and process data for at least one of the factors at play in schedule reliability for container ships: weather. Lee, Aydin, Choi, Lekhavat and Irani (2018) talked about how to collect data from the European Spatial Agency's Earth Observation satellites network Copernicus and how to apply found weather patterns to the operations of the maritime shipping industry.

2.3. Machine Learning Methods

Mohd Salleh et al. (2007) described how the use of a FRBBN is useful in predicting a containership's arrival at destination port. Their model was more accurate than the Neural Network model proposed by Fancello, Pani, Pisano, Serra, Zudas and Fadda (2011). However, the limitation in the FRBBN model is that it is more suitable for a small-scale use case so that it is difficult to replicate for new cases and data sets. Machine Learning models, on the other hand, are powerful in making the computer algorithms to "learn" from existing data, and the models are subsequently validated and tested with old and new data. Despite being less accurate, the Neural Network model described in Fancello et al (2011) is more suitable for a business application. We will also explore different Machine Learning algorithms such as Linear Regression and Random Forest.

The basis for the theoretical background and the application of Machine Learning in a business context was described in Shmueli, Bruce, Yahav, Patel, & Lichtendahl, (2018). They described in details how Linear Regression, Neural Networks and Random Forest work and how they are applied in practice on different datasets. We were also able to utilize the vast knowledge in the python open-source community for algorithm development, for instance Scikit-learn, Buitinck et al. (2013) and Pedregosa et al. (2011) and the Keras library (Cholet, 2015).

3. DATA AND METHODOLOGY

In order to answer the research question about how to use Machine Learning to improve prediction of cargo availability at the destination port by combining historical transit variability and data pertaining to external relevant factors, we perform the following procedures in the analysis. Note that in this report we use the shipping routes between South-East Asia and North America, which represent the most used routes for the company, as a base to build a model that will be a proof of concept for the company's future use on a larger perimeter, which may include all of their routes.

3.1. Terminology and definitions

First, it is important to establish data science naming conventions that will be used throughout this report. In this project, we used supervised Machine Learning algorithms. To train these algorithms, we need as an input various samples that consist of features and a target variable. A sample is one row in the dataset and consists of all the information available about one shipment. A feature is the name for one specific type of information about this row. Since a row represents a shipment, a feature could be, for example, the average time a carrier needs for a specific route. The target variable is defined as the variable the model tries to predict. For the model *depart*, this is the time period in days from when the container ship left from the port of origin to the date when the container is unloaded at the port of destination.

We divide the shipment process into four segments which match our four different models: *booked*, *received*, *gate-in* and *depart*. The model *booked* starts at booking date. The model *received* starts at the date when the shipment is received by the shipper. The model *gate-in* starts at the date when the shipment arrives at the port of origin. Finally, the model *depart* starts the prediction at the time the containership leaves the port of origin. The four models yield four outputs regarding the time of unloading for the shipment in the port of arrival, each becoming more precise as the journey goes along.

3.2. Analyzing the historical data and selecting relevant external factors

Our partner company Maersk provided us with the following: 10 files, in the comma separated values (CSV) format covering a total of 2,804,989 shipments that occurred between October 10, 2015 and November 3, 2018. Each CSV file is composed of 44 columns corresponding to different information about the shipments, with each line of data representing a shipment. For our project, the interesting part of the dataset consists of 18 timestamps indicating when the shipment reached a milestone during the transport.

The important milestones for this project are, in chronological order:

- The booking date, which is the date when the transport from origin to final destination is booked (called "book date" in the dataset);

- The receive date, which is the date when the cargo is received by the shipper at the port of origin (called "Actual Receipt Date" in the dataset);

- The gate-in date, which is the date when the cargo enters the yard where it will wait to be loaded on the vessel (called "Gate In Origin-Actual" in the dataset);

- The vessel departure date, which is the date when the cargo leaves the port of origin (called "Vessel Actual Departure" in the dataset)

- The vessel arrival date, which is the date when the cargo arrives at the port of destination (called "Vessel Actual Arrival-Actual" in the dataset)

- The unload date, which is the date when the cargo is unloaded from the vessel ("Container Unload From Vessel-Actual")

In addition, the dataset provides the name of the carrier, shipper, customer and port of origin and destination for each shipment. The full list of columns with the corresponding description can be found in Appendix A.

Lastly, as every record of the dataset represents one shipment, it is important to note that we cannot use every single observation for modelling as multiple shipments get consolidated into one container and then multiple containers get loaded onto one vessel. In order to avoid duplicates, we changed the level of granularity at which we exploit the dataset based on the model that is run. For example, using all of the

shipment data for the model that makes a prediction at the booking date (using the model *booked*) is necessary, whereas only data consolidated to the container level is relevant for the model that makes a prediction at the vessel departure date (model *depart*).

The whole dataset contains information about 131 unique routes between nine different countries. The ports on these routes are located in the United States, Europe and Asia. There are 51 unique carriers and 3280 unique shippers using these routes. It is important to note that all routes do not have the same number of data points. Consequently, the accuracy of the models' predictions on routes with fewer observations could be worse than others with many observations.

Considering all the carriers on one route, we observe that the transit time can significantly vary between shipments. Figure 2 depicts the observed transit times in days between the ports of Ningbo, China and Long Beach, California. According to the dataset, the minimum transit time a carrier needs for this route is 15 days while the longest transit time is more than double with a median of 21 days. 80% of all shipments will arrive in Long Beach after 22 days at sea.
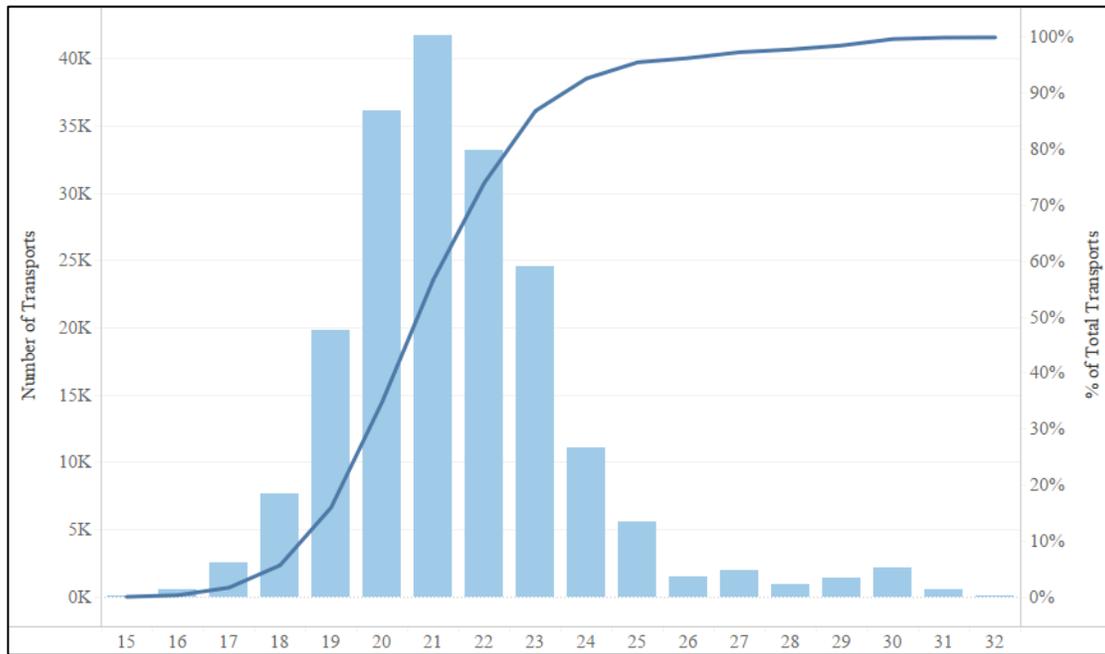
Figure 2 Transit time from Ningbo to Long Beach – The bar chart represents distribution of transit time in days, while the line chart represents the cumulative frequency of transit time in days

Of course, when looking at different port combinations, the transit time and its associated standard deviation vary as well but not necessarily in the same magnitude. Generally speaking, we observed that as transit time gets longer, the associated variability becomes greater (both in terms of days and percentage).

Furthermore, there is also variability regarding the time spent in either the port of departure or the port of arrival by the container. This variability depends on the performance of the ports. Figures 3 and 4 show the time distribution of the time spent by containers in the ports of Los Angeles, California and Yantian, China. As explained later in the report, Los Angeles is only considered a port of arrival within this project. The distribution shown in Figure 3 is only for containers arriving in Los Angeles. This time period represents the moment between when the vessel arrives at the port and when the container is unloaded from the vessel.
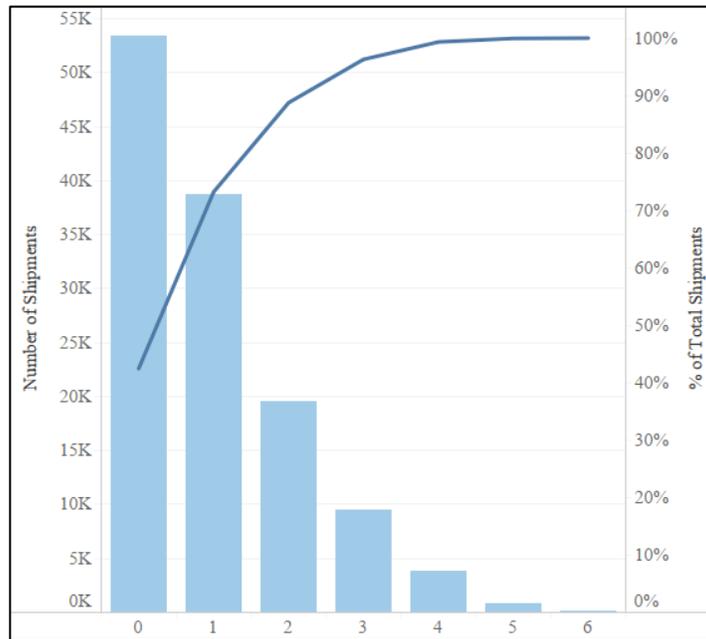
Figure 3: Transit time Port of Los Angeles in days - The bar chart represents the distribution of time from arrival at the port until the container is discharged, while the line chart represents the cumulative frequency of this time period

On the opposite side, the data from the port of Yantian represents the time it took for the cargo to be loaded on the vessel. The timeframe it covers ranges from the receive date to the vessel departure date. It is clear that we are facing a lot of variance at this step of the journey. This variance can be explained by the fact that only some shipments have to be consolidated. Likewise, certain shipments can be stored on the port's yard for a couple of days before being loaded, either because they arrived on the yard earlier than scheduled or because they missed their original vessel departure and are waiting for the next one.
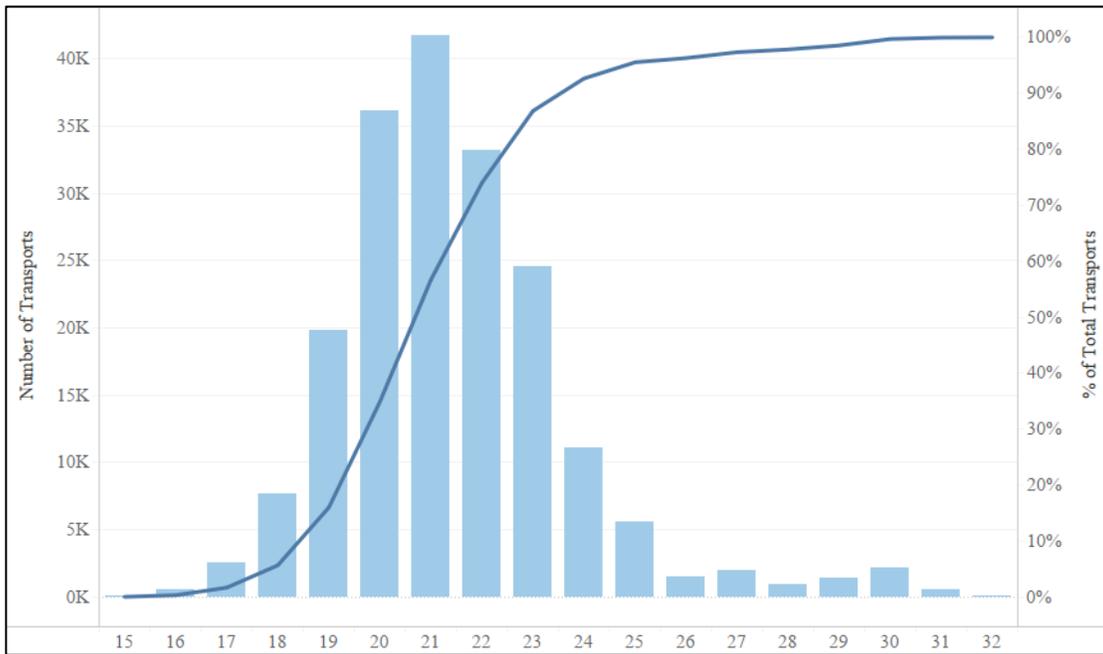
Figure 4: Transit time in days port of Yantian - The bar chart represents the days spent in the port from the time the shipment is received by the shipper to the departure of the container ship, while the line represents the cumulative frequency of this period

Lastly, when considering carriers in the dataset, it shows a rather consolidated profile, i.e. more than 50% of the records are associated with only four carriers, whereas the rest of the shipments are associated with 47 other carriers (see Figure 5).
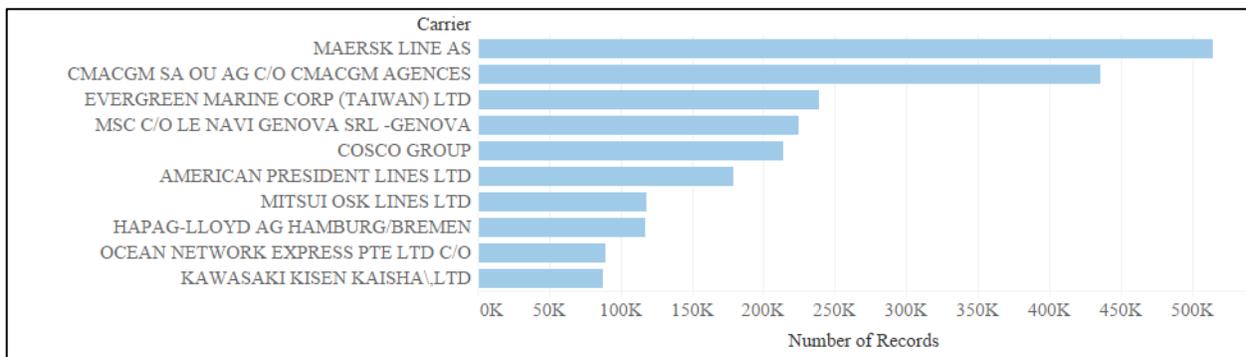
Figure 5: Top 10 Carriers - ranked by number of shipments in the dataset

As shown above, there is a great number of route-carrier combinations present in the dataset. In order to minimize the number of combinations to work with, after consulting the partner company, we agreed to limit the scope of our study to shipments loaded in South East Asia (China, Hong Kong and Vietnam) and

unloaded in the United States. This geographical choice was made because the amount of data for routes with European discharging ports is less substantial compared to that of the routes with US discharging ports. However, the geographic scale can be easily adapted by the company at a later stage. After defining this scope and dropping observations without an unload date, the final dataset used in the study 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.

3.3. Selecting relevant external factors

Chinese New Year

During the Chinese New Year festivities, ports are expected to work slower. Based on an interview with Maersk, we identified a 14-day time interval before the historical dates of Chinese New Year to model this effect. This period is used to model the slowdown in activity that happens prior to the actual celebration. In the model, we use a binary column: it takes the value one if the timestamp 'Expected Receipt Date' falls into this period. If not, the value is zero and therefore, there is no impact of Chinese New Year on the expected transit time for the shipment.

Port Congestion US

We obtained daily data for number of containers unloaded in a given port from Panjiva Inc. (source: https://panjiva.com/). The maximum number of containers unloaded over the period 2015-2018 sets a baseline (capacity). Then we compared that capacity to the flow on a given day using a rolling mean of three days. As an example, for the 24th of April, we will take the historical volumes unloaded on the 23rd, 24th and the 25th, calculate the mean of the 3 days and then compare it to that max capacity determined before. This helps to determine the historically used capacity for a port at this time of the year.

Weather

Weather data was not taken in account into our modelling. We made this decision because of the lack of free publicly available data. Another factor in our decision was the imperative to solve the dilemma of considering weather patterns around the port of departure and port of arrival versus weather throughout the entire route of the vessel (which would require to know approximately what route the vessel took).

Positioning of the Container on the vessel

Positioning of the container on the vessel refers to the slot the container has been assigned to on the vessel. After interviews with Maersk, it was determined that this could be a factor of influence as it reportedly has a great impact on the unloading time of the container. We even found out that there are so called "premium spaces" on the vessels: these are the slots where the containers are unloaded first at port of destination, which allows these containers to be out of the vessel and then the port of destination faster. However, we could not include that factor into our modelling for three reasons:

- First, detailed placement of a container (or a customer's containers) on a vessel is variable and might be proprietary data to the ocean carrier and the customer

- Second, while we might have been able to access such data for Maersk's operations, this would have been the only carrier for which we would have had this kind of information. This would create a model biased towards Maersk because the output would be more precise, thus influencing the user's decision to select Maersk as the carrier. Our goal here is to develop a carrier agnostic model, that will help to make the best decision that is in the shipper's best interests and not the carriers'.

- Third, this kind of data would have increased the amount of data necessary to run our model by a substantial magnitude, making it longer to train and run and more complex to handle in general.

3.4. Selecting features for the Machine Learning model

The following part describes the features used in the models and the data they are built on. Not all of the features are used in every one of the four different models. Appendix B includes a table that shows which features are used in which model.

Average time spent at Port of origin (Port of origin mean)

This is the average time shipments spend in the port of origin from the moment when they are received by the shipper until the vessel leaves the port. The average is based on the different shipper and port combinations.

Standard deviation of time spent at Port of origin (Port of origin SD)

This is the standard deviation of the average time shipments spend in the port of origin from the moment when they are received by the shipper until the vessel leaves the port. The standard deviation is based on the different shipper and port combinations.

Average time spent at Port of destination (Port of destination mean)

This is the average time shipments spend at the port of destination from the moment when the containership arrives at the port of destination until the containers are discharged at the port. The average is based on the different customer and port combinations.

Standard deviation of time spent at Port of destination (Port of destination SD)

This is the standard deviation of the time shipments spend at the port of destination from the moment when the containership arrives at the port of destination until the containers are discharged at the port. The standard deviation is based on the different customer and port combinations.

Average time spent per Route (Route mean)

This is the average transit time a containership needs to travel from one port to another. This average is based on the different carrier and route combinations.

Schedule

This is a refined route mean. This average also depends on the day of the week the containership left the port of origin. It helps simulating the different schedules and loops that carriers provide for the same port combination. Carriers operate vessels on different loops that include a fixed number of ports. The vessels will arrive at the same port on the same day of the week but may travel on a different route.

Origin Service

This is a binary column indicating if the shipment has to be consolidated at the port of origin after it is received by the shipper. Consolidation may add to the overall transit time.

Holiday

This is a binary column indicating if the shipment falls within a two-week period before Chinese New Year.

Quarter

This is the z-score of the transit time for every shipment in the dataset normalized by the carrier route combination:

$$zscore = \frac{x - \overline{x}}{\sigma}$$

To be able to compare different routes, the mean and standard deviation used to calculate the z-score are not the mean and standard deviation of all transit times but for every carrier and route combination. We then calculate the average z-score for every quarter. This indicates a higher average transit time in one quarter compared to another.

Expected Time to port

The time period between the booking date and the expected time the shipper will receive the shipment at the port of origin. This expected time is given by the customer.

Port of origin late

This is a binary column indicating if the shipment is received after the expected received date.

Port of origin latest

This is a binary column indicating if the shipment is received after the latest possible received date given by the shipper.

Capacity of port of destination

This is the percentage of total capacity historically occupied at the port of destination in a two-day window around the expected arrival date of the containership.

Late departure

This is a binary column indicating if the containership left the port of origin later than the expected date of departure.

3.5. Selecting a Machine Learning algorithm

Building on the research from Fancello et al (2011) and their Neural Network model, we experimented with Linear Regression and Random Forest to see if these algorithms perform better on our dataset than a Neural Network. After comparing the results of these three algorithms, we decided to build the final prediction model with a Random Forest algorithm. Final results and choice of hyperparameters for the Random Forest are discussed further in chapter 4.1.

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 and saves the final model. Users can choose which one of the four models they want to train and if the model should be trained on a whole new data set or add new data to an existing model. The third script is used for prediction in a production environment. The program can read input data like the choice of carrier, shipper and route and then predict the corresponding estimated date of unloading at the port of destination (see Appendix C). The output also contains a 90% confidence interval with the earliest and latest unloading date. The input is currently made by choosing the variables via a dropdown menu in an Excel sheet, but can be changed by the company to any other input method.

3.6. Coupling our models with Harmony

Our main goal was to be able to add a prediction capability to the existing descriptive tool Harmony. As our models build on the same historical data as Harmony, the two can be coupled with minor adjustments. The company would have to add data only about port congestion and about Chinese New Year dates.

Right now, Harmony is a descriptive analysis tool: it takes historical data and then describes what happened in the past at every milestone along the way for a shipment on a certain shipping lane. Coupled with our program, the tool would also be able to look ahead and predict the unloading time of a container at the port of destination.

If Maersk wants to extend the capability of the prediction tool to other routes, they can follow the same methodology and just add to our model. The historical data will be used in the same format for different routes. Additionally, port congestion data for the new ports touched by these new routes would also have to be supplied in order to train the extended model. From there, predictions can be made in the same way as they are done in our original program.

4. RESULTS AND ANALYSIS

In this section, we will present the results obtained by the different Machine Learning algorithms we have tested. Then we will compare them and justify our choice of using the Random Forest algorithm.

4.1. Results using the Random Trees algorithm

The Random Forest algorithm is an ensemble method that consists of multiple Regression Trees used either for classification or prediction. The algorithm uses "if-then" statements to split the data into subgroups that minimize the variance of the target value in these sub groups. The difference from Regression Trees is that the Random Forest algorithm produces multiple Regression Trees by randomly sampling the training data with replacement. The splits are found randomly by looking at only a subsample of all features and splitting on the feature that reduces the variance the most. In order to yield better results, Random Forest algorithms grow much bigger trees than a simple Regression Tree. If a Regression Tree is grown until the end, it is certain to overfit the data and only perform well on the training set. Because the Random Forest randomly samples multiple training sets and predicts the results by averaging all the trees, it will not overfit as much as the simple Regression Tree. Ensemble methods typically perform better on predicting an outcome but lose the interpretability of the model. Shmueli et al. (2018) use the example of an investor in the financial market who creates a portfolio of assets to reduce his risk to explain why ensemble methods typically improve predicting power and reduce variance in the predicted outcome.

In the model used for this project, the splitting criterion is reducing the mean squared error after the split. The number of features available for a split in this model is the square root of the number of all features. The algorithm used is the RandomForestRegressor from the popular Python library "Scikit-learn" (Buitinck et al., 2013 and Pedregosa et al., 2011).

Figure 6 represents a Decision Tree on the model *booked*. The first split is on the average time the carrier needs for this specific route. The next split is on the expected time from booking date to received by shipper. All of these splits reduce the mean squared error of the samples in the leaves of the tree. This is a representation of how a Decision Tree works. The final Random Forest model is deeper, splits on random selected variables and consists of 500 of these regression trees.

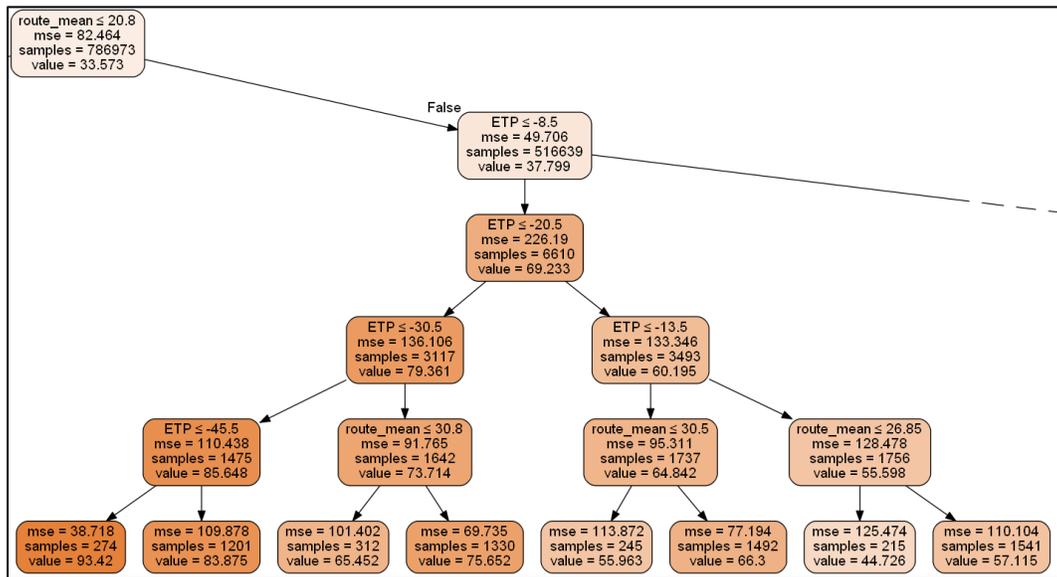*Figure 6: Sample of a Decision Tree - This figure displays one branch of one of the trees used in the booked model*

We ran all the models multiple times with different depths of the trees in order to decide the proper tree depth that results in the most stable parameters. In Figure 7, the model *booked* performance, measured by mean absolute percentage error (MAPE), on the training and test data sets is presented.
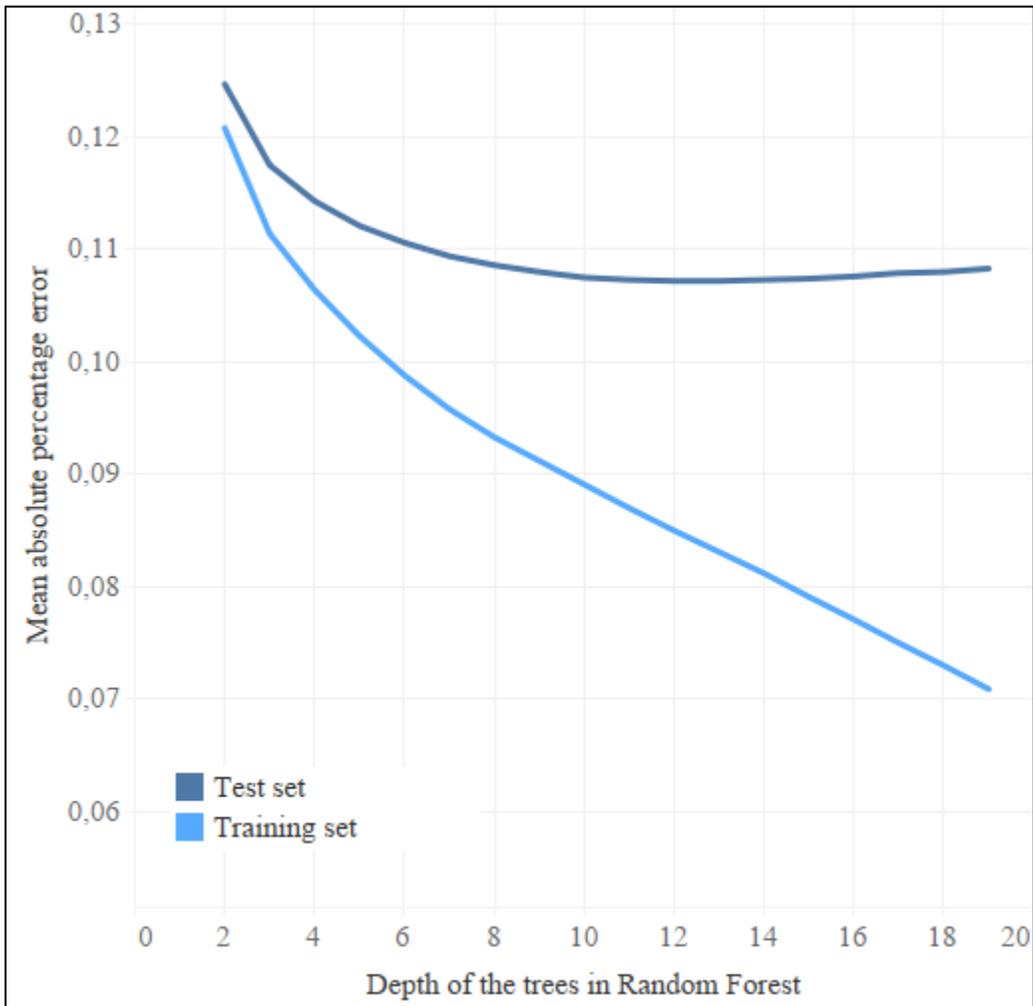
Figure 7: Performance regarding to hyperparameter depth of the trees

It shows that in terms of MAPE, the model applied to the training dataset significantly outperforms that on the test set, suggesting overfitting, when the depth of the trees becomes deeper. As a result, we chose 10 as the tree depth because it provides the most stable results. In Table 1 the performance of the most stable models is represented with their test set performance indicators.

Table 1: Key performance indicators of the Random Forest models - Aggregated results for all carriers on all routes of the test set

Model	R²	Mean absolute error	Mean absolute percentage error	Root mean squared error	Depth of the Trees
Booked	0.64	3.74	0.1075	5.44	10
Received	0.74	3.36	0.0981	4.68	7
Gate_in	0.82	3.0	0.1004	4.02	9
Depart	0.88	2.1	0.0939	2.76	5

As expected, the accuracy gets better as the shipments move forward on the route. The model *depart* has the best performance indicators. The predictive power of all of the models could be improved by adding more features. For example, after consulting industry specialists, we concluded that the position of the container on the vessel has an impact on unloading time spent in the port of destination. The information about the positioning on the vessel would impact the performance of all the models.

Table 2 shows the performance indicators of our baseline model. The baseline model tests the mean of the target variable by carrier on a specific route from the first two years in the dataset against the third year. We can see that the Random Forest performs better for all models. However, as the cargo gets closer to its destination, the performance difference between the baseline model and the Random Forest decreases. Especially the Random Forest model *depart* does not perform significantly better than the baseline model.

Table 2: Key performance indicators of the baseline models - Aggregated results for all carriers on all routes of the test set

Model	R²	Mean absolute error	Mean absolute percentage error	Root mean squared error
Booked	0.51	4.3	0.12	6.57
Received	0.73	3.72	0.111	5.24
Gate_in	0.8	3.1	0.1013	4.24
Depart	0.88	2.14	0.0941	2.84

4.2. Results using the Neural Network algorithm

A Neural Network is a black box algorithm that tries to mimic the activity of the human brain. The advantage of a Neural Network over other Machine Learning algorithms like Linear Regression is that it automatically finds non-linear relationships between the features (Shmueli et al., 2018). Neural Networks perform very well on highly complex problems. For example, they are used in practice to classify pictures of tumors in the medical sector or to predict the steering pattern in self driving cars. This algorithm can be used not only for classification, but also for regression. Fancello et al. (2011), used a Neural Network with one or two hidden layers to predict the arrival time of shipments in ports (see Figure 8).
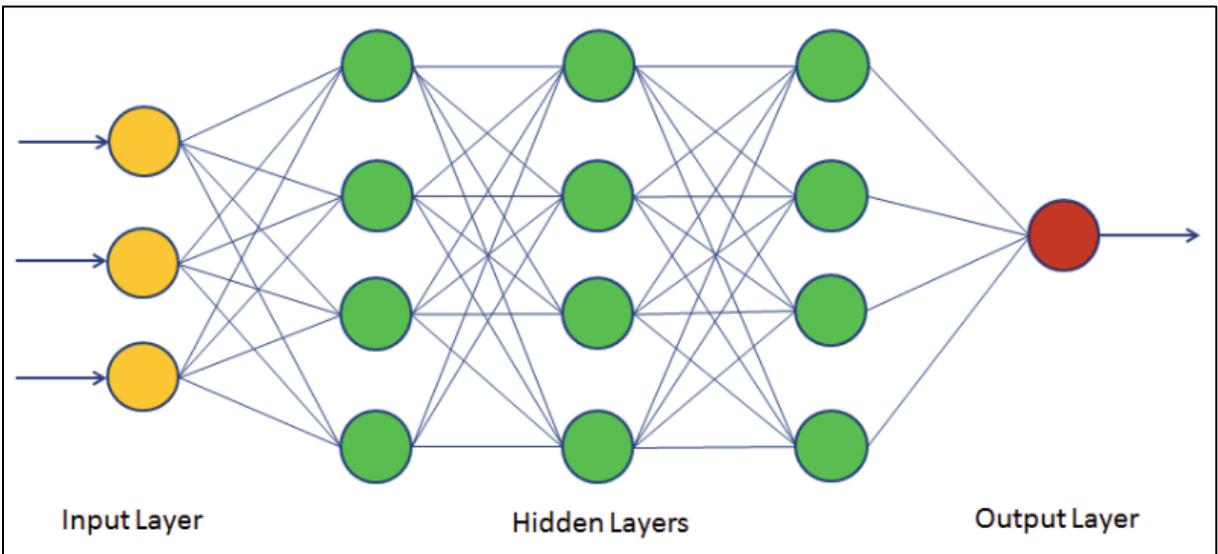
Figure 8: Representation of a Neural Network - Source: www.datacamp.com

The Neural Network Model uses the same input variables as the Random Forest. The difference is that it is a black box and it is not possible to visualize what happens inside the model. For this project, we used the widely used open source Python library Keras, which is an interface for the Tensorflow backend (Chollet, 2015).

The final parameters for the algorithm were found by trial and error. The final model uses one hidden layer with 50 neurons. The hidden layer also uses 20% dropout regularization to minimize the risk of overfitting the data. The neurons use a rectified linear unit activation function and the model minimizes the mean

squared error of the prediction. Before training the model, all variables are standardized to transform all features to the same scale.

As we saw with the Random Forest models, the Neural Network performs better than the baseline model on the first two models where the transit time and overall variability is larger (see Table 3). The Neural Network actually performs slightly worse than the naive mean on the models *gate-in* and *depart*. Another downside of the Neural Network models is their need for big computational resources. Training a Neural Network takes much more time and resources than any other model. The model could be improved by adding more hidden layers, but the accuracy gains would be marginal and the computational power needed to train the model would increase even more.

Table 3: Key performance indicators of the Neural Network models - Aggregated results for all carriers on all routes of the test set

Model	R²	Mean absolute error	Mean absolute percentage error	Root mean squared error
Booked	0.61	3.99	0.1157	5.71
Received	0.73	3.33	0.095	4.72
Gate_in	0.8	3.21	0.1115	4.23
Depart	0.85	2.33	0.0997	3.12

4.3. Comparing results between Neural Network and Random Forest

Here we are not only comparing the results of the Random Forest and the Neural Network, but we are also comparing both of them against the baseline model. This baseline model is the average of the target variable for every route carrier combination, calculated on the training set with a naïve mean.

Although the baseline model for the model depart yields similar results as the Machine Learning models, the Random Forest performs best on all four models (see figure 9).
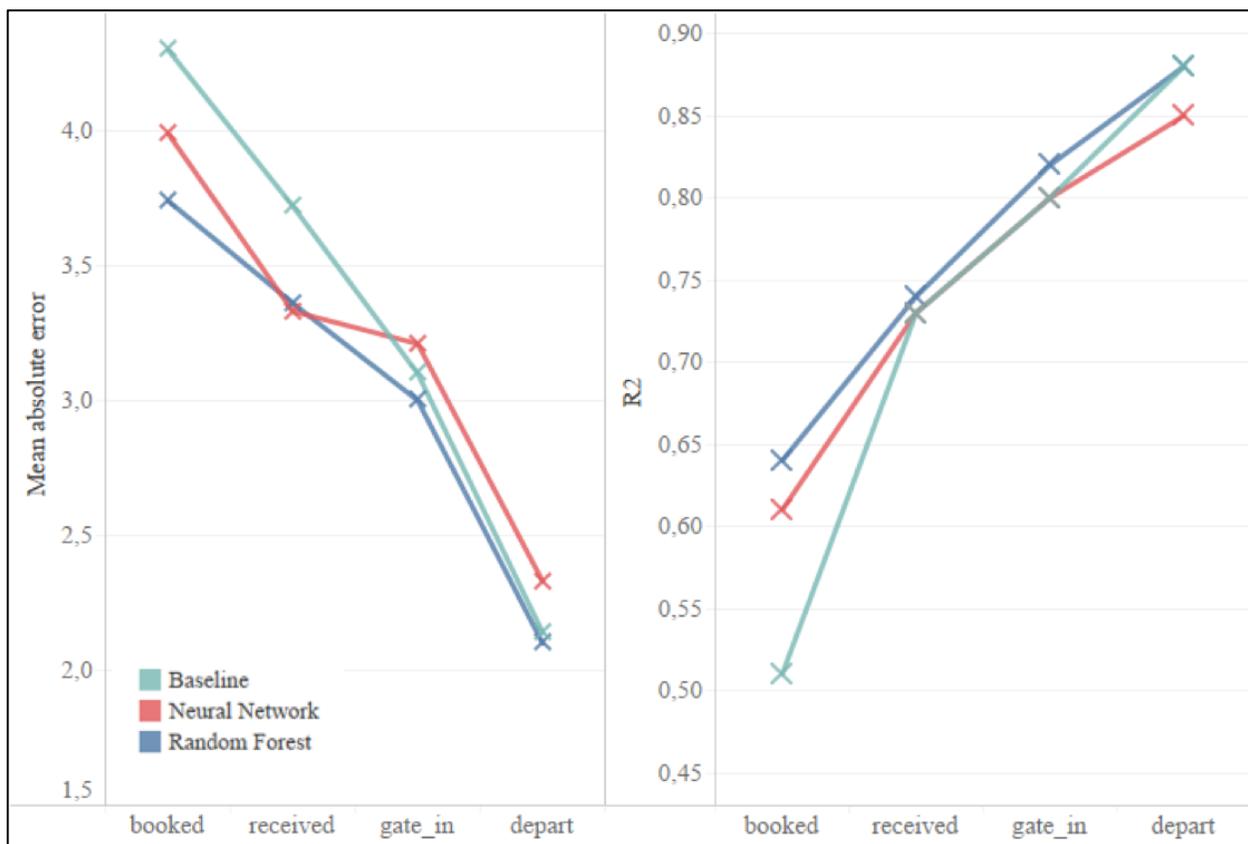
Figure 9: Comparison of Metrics: Random Forest vs Neural Network vs Baseline

Not only does the Random Forest model perform 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.

As mentioned before, 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.

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. Moreover, the Random Forest algorithm has a good interpretability, especially compared to Neural Networks, which helps monitoring it. It also has the benefit of shorter computing time.

5. DISCUSSION

In this section, we will discuss the results as well as the performance of the models. We will evaluate results and performance from an operations' point of view as well as from a modelling standpoint. The latter will explore the limitations of using such models to make predictions as opposed to using "simple" and more traditional approaches.

5.1. Quality of our model

In order to assess the quality of the predictions made using our models from an operational standpoint, Maersk shared with us their time interval error metric, which they use to assess prediction accuracy regarding ETA. According to this metric, the prediction can be two days too late or one day early compared to the actual recorded time of unloading of the cargo at the port of destination. Our Random Forest model does an acceptable job with its predictions on the test set. Figure 10 shows the percentage of all predictions that fall into that time interval, for each of the four models.
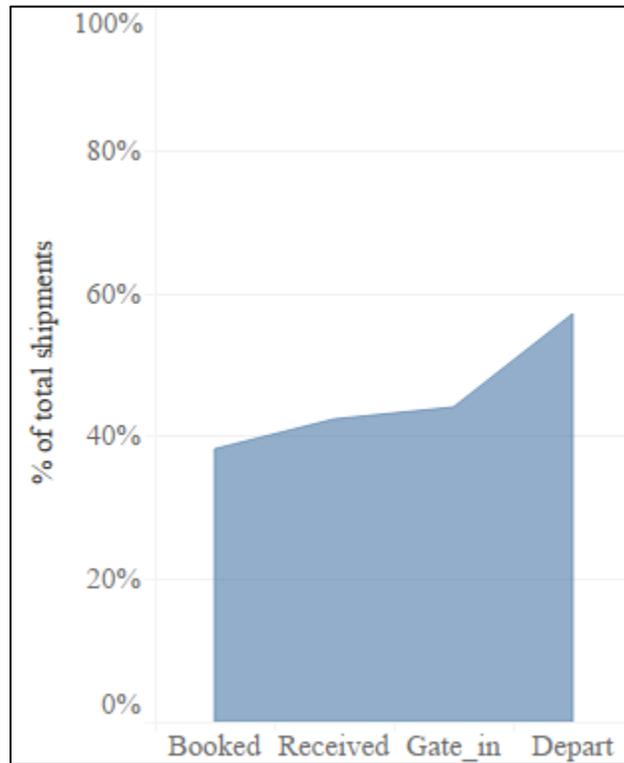
Figure 10: Percent of predictions falling in the time interval between two days before and one day after the actual arrival in the test set

Figure 11 shows the mean absolute error comparison and the R2 statistic between all the models we tested: Random Forest, Linear Regression and our baseline model. We can see that the Random Forest performs better, but the gap between the Random Forest and the simpler models is decreasing, and it becomes almost non-existent as the shipment progresses along. This observation leads us to believe that one option to exploit our model might be only to use it for the first 3 milestones: booked, received and gate-in. Essentially, the model *depart* would not be used for predicting the last leg of the journey, but could be replaced by the baseline or a Linear Regression. The upside for this configuration would be the lower maintenance of the overall system with one Machine Learning model replaced by a more traditional modelling approach.

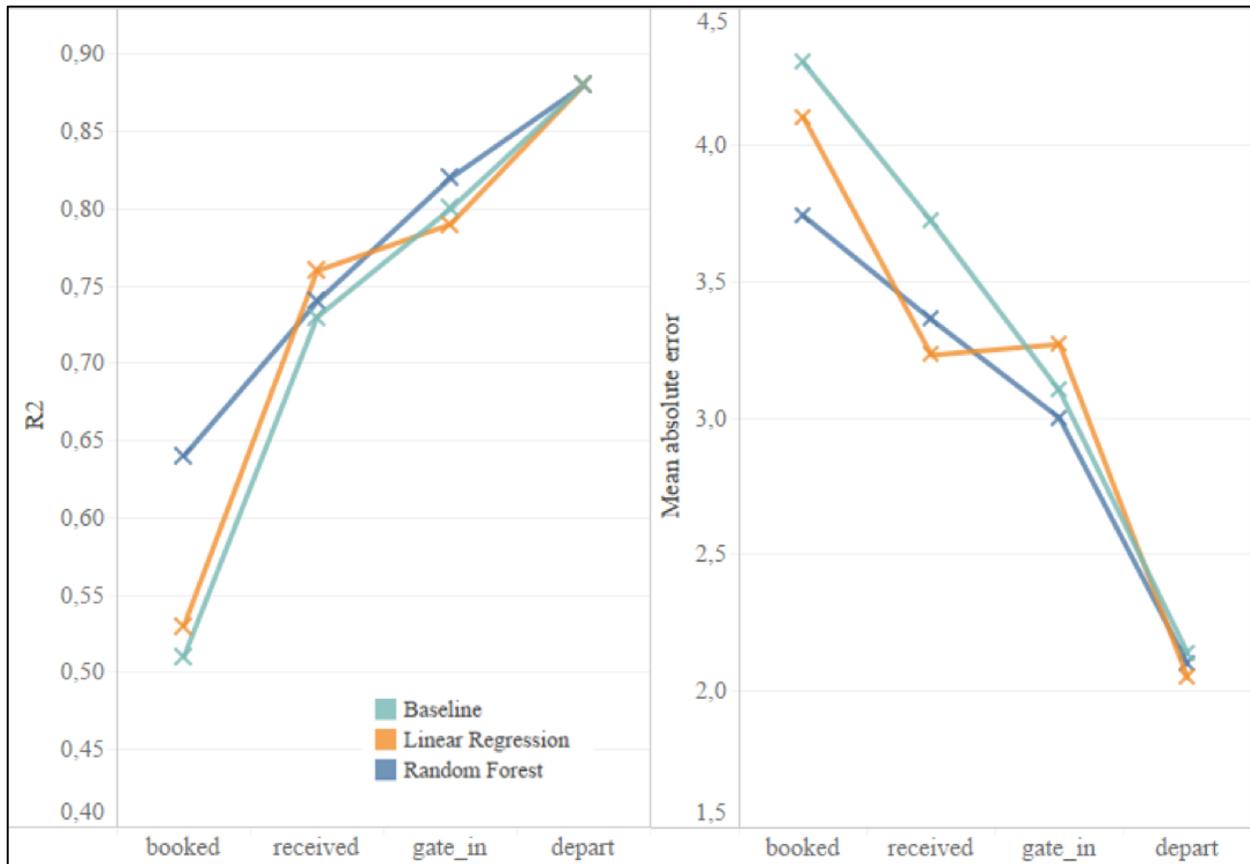*Figure 11: Model comparison: Random Forest vs Baseline vs Linear Regression*

This decreasing difference in performance can be explained by the fact that more stakeholders are involved from booking to gate-in , compared to after the vessel has left the port of origin. More stakeholders involved suggest higher friction and uncertainty in the system, thus it is more difficult to predict. This last leg of the journey is, to a large extent, under the control of the carrier. Moreover, it has been the carriers' focus to improve the service reliability.

To further improve the models, especially the *depart* model, more features could be added. For example, container positioning on the vessel, GPS positioning of the vessel and the corresponding timestamps along its journey to track its progress could be added. However, in the scope of our project, we did not have this kind of data available to be able to add it to our modelling.

Other models, like the one developed by Mohd Salleh et al. (2017) also include features regarding the weather conditions, at the port of arrival and en-route. In our case, we did not have access to such data but

it could be interesting in the future to include such features and measure the subsequent improvement of our models.

5.2. Limitations of our approach

Machine Learning is a method that is rather data intensive in two ways: it requires a lot of data in order to kick-start the analysis and modelling and it requires that this data be of at least moderate and at best great quality. For great quality to be achieved, this means there should be no missing or wrong data points in the dataset, as well as consistent and useable formatting of the data. If not, the length of the data cleaning part of the project might be considerably extended, which could pose a timing problem, especially if the project is bound in time.

Furthermore, developing such a model demands analytics and coding skills. These two skills, even if required, are not enough: having subject-matter experts providing input on the industry practices and interpreting results and data is crucial to the success of such an endeavor.

Looking further down the road, since the end goal of a project like ours is to supply our partner company with a functioning tool, this means Maersk will have to work on the models in order to make them a part of their computing environment. For example, the models we are delivering draw their data from several Excel files but in the future, we can imagine a direct link to a database instead.

6. CONCLUSION

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. However, we also found this is a valid use if lead times are long. As we have demonstrated in this report, of the four models we built, the one that performs the "least well" compared to more traditional methods is the *depart* model, as this is 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 word.

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 have more detailed data that provide more timestamps along the way of the shipment as opposed to having a timestamp only at the beginning and one at the end of each process. We are confident that this would drive up the accuracy of the model even though this seems difficult to achieve. It implies that data is collected everywhere along the way and then centralized for a later use. However, with the advent of the Internet of Things (IoT) and the increasing demand for instant tracking, this aim might be achieved in the near future.

Lastly, the ultimate goal for the evolution of our model would be to cover the entire journey of the shipment as opposed to stopping at the unloading from the vessel at the port of arrival. Indeed, in our study, we focused on the journey from transportation booking to the unloading of the vessel due to a lack of reliable data for subsequent steps. We believe that our model covers most of the variance that can happen for a shipment but with access to more data for the following steps, all the way to the final delivery at the customer's facility, it would be possible to make an actual prediction of when a load will be delivered.

REFERENCES

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., … Varoquaux, G. (2013). API

    design for machine learning software: experiences from the scikit-learn project. *ArXiv:1309.0238 [Cs]*.

    Retrieved from http://arxiv.org/abs/1309.0238

Chung, C.-C., & Chiang, C.-H. (2011). The Critical Factors: An Evaluation of Schedule Reliability in Liner

    Shipping, (4), 8.

de Langen, P., & Nijdam, M. (n.d.). NEW INDICATORS TO MEASURE PORT PERFORMANCE, 14.

Chollet, F. (2019a). *Deep Learning for humans.* Keras. Retrieved from https://github.com/keras-

    team/keras (Original work published 2015)

Fancello, G., Pani, C., Pisano, M., Serra, P., Zuddas, P., & Fadda, P. (2011). Prediction of arrival times and

    human resources allocation for container terminal. *Maritime Economics & Logistics*, *13*(2), 142–

    173. https://doi.org/10.1057/mel.2011.3

Find out everything about Damco. (n.d.). Retrieved December 19, 2018,

    from http://www.damco.com/en/about-damco

Kepaptsoglou, K., Fountas, G., & Karlaftis, M. G. (2015). Weather impact on containership routing in closed

    seas: A chance-constraint optimization approach. *Transportation Research Part C*, *55*, 139–

    155. https://doi.org/10.1016/j.trc.2015.01.027

Lee, H., Aydin, N., Choi, Y., Lekhavat, S., & Irani, Z. (2018). A decision support system for vessel speed

    decision in maritime logistics using weather archive big data. *Computers and Operations Research*, *98*,

    330–342. https://doi.org/10.1016/j.cor.2017.06.005

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Cournapeau, D. (n.d.).

    Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, 6.

Port Performance Annual Report to Congress 2017.pdf. (n.d.).

Salleh, N. H. M., Riahi, R., Yang, Z., & Wang, J. (2017). Predicting a Containership's Arrival Punctuality in Liner Operations by Using a Fuzzy Rule-Based Bayesian Network (FRBBN). *The Asian Journal of Shipping and Logistics*, *33*(2), 95–104. https://doi.org/10.1016/j.ajsl.2017.06.007

Sánchez, R. J., Hoffmann, J., Micco, A., Pizzolitto, G. V., Sgut, M., & Wilmsmeier, G. (2003). Port Efficiency and International Trade: Port Efficiency as a Determinant of Maritime Transport Costs. *Maritime Economics & Logistics*, *5*(2), 199–218. https://doi.org/10.1057/palgrave.mel.9100073

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (n.d.-a). DATA MINING FOR BUSINESS ANALYTICS: Concepts, Techniques, and Applications in R, 577.

UNCTAD. (2016). REVIEW OF MARITIME TRANSPORT 2016. *REVIEW OF MARITIME TRANSPORT*, 15.

van Merrienboer, B., Bahdanau, D., Dumoulin, V., Serdyuk, D., Warde-Farley, D., Chorowski, J., & Bengio, Y. (n.d.). Blocks and Fuel: Frameworks for deep learning, 5.

Vernimmen, B., Dullaert, W., & Engelen, S. (2007). Schedule Unreliability in Liner Shipping: Origins and Consequences for the Hinterland Supply Chain. *Maritime Economics & Logistics*, *9*(3), 193–213. https://doi.org/10.1057/palgrave.mel.9100182

Vettor, R., & Guedes Soares, C. (2017). Characterisation of the expected weather conditions in the main European coastal traffic routes. *Ocean Engineering*, *140*, 244–257. https://doi.org/10.1016/j.oceaneng.2017.05.027

APPENDIX

APPENDIX A

Description of each columns contained in the original dataset provided by Maersk.

Column Name	Description
PO Line Uploaded	Date when data is uploaded to the system (not relevant)
POH Client Date	(not relevant)
POH Upload Date	(not relevant)
Book Date	Date when customer books the transport
Receipt Date	Date when shipment is received by the shipper
Consolidation Date	Date of consolidation: "CY" should be same date,"CFS" could be 2-6 days later
ETD	Estimated time of departure from port of origin
ETA	Estimated time of arrival at port of destination
ATD	Actual time of departure from port of origin
ATA	Actual time of arrival at port of destination
ConsigneeName	(not relevant)
PO Number	(not relevant)
Origin Service	Indicates if shipments have to be consolidated at port of origin; Either CFS (Container Freight Station) or CY (Container Yard)
Destination Service	Indicates that shipment has to be consolidated
Consignee	(not relevant)
Carrier	Name of Carrier
VOCC Carrier	If carrier is a non-vessel operating carrier then the real carrier name is here
Carrier SCAC	Standard Carrier Alpha Code - 4 letter code for the carrier
CBL Number	Commercial Bill of Lading number
Booking Number	Reference number for booking
Shipper	Name of the shipper; the factory who sends the goods
Original Port Of Loading	Name of the city of the port of origin
Original Port Of Loading Site	Country of port of origin
Final Port Of Discharge	Name of city of the port of destination

Final Port Of Discharge Site	Country of port of destination
Actual Measurement	Measurement of the container
Earliest Receipt Date	Shipping window start
Expected Receipt Date	Customer will tell this date when booking a shipment
Latest Receipt Date	Shipping window end
Actual Receipt Date	Date when shipment is actually delivered at port of origin
Empty Equipment Dispatch-Actual	(not relevant)
Gate In Origin-Actual	Date of entry at port of origin
Container Loaded On Vessel-Actual	Date when container is loaded on vessel
Container Unload From Vessel-Actual	Date when container is unloaded from vessel at port of destination
Gate Out Destination-Actual	Date when container leaves port of destination
Container Empty Return-Actual	(not relevant)
Equipment Number	Reference number of the container
Confirmation Date	(not relevant)

Table summing up which features are used in each of the four models developed.

Name	*booked*	*received*	*gate-in*	*depart*
Port of origin mean	x	x	x	
Port of origin SD	x	x	x	
Port of destination mean	x	x	x	x
Port of destination SD	x	x	x	x
Route mean	x			x
Schedule		x	x	x
Origin service	x	x	x	
Holiday	x	x	x	
Quarter Z-score	x	x	x	x
Expected time to port	x			
Port of Origin late		x		
Port of Origin latest		x		
Capacity of Port of Destination		x	x	x
Late departure				x

APPENDIX C

List of the data field from the user to input into one of the four model to then obtain a prediction. Please note that not all the fields are required to be populated in the input file to run all the models.

Column Name	Description
Customer	Name of the end customer who will receive the shipment
Carrier	Carrier's Standard Carrier Alpha Code
Shipper	Name of the shipper
Original Port of Loading	Name of the port of departure
Final Port of Discharge	Name of the port of arrival
Origin Service	Either CFS (Container Freight Station) or CY (Container Yard)
Expected Receipt Date	Expected receipt date of the shipment by the shipper in the port of origin
Book Date	Date on which the transport was booked
Gate-in Origin Actual	Actual date of when the container on the port's yard
Actual Receipt Date	Actual date of when the shipper receives the goods
Latest Receipt Date	Latest receipt date, given by the shipper
ETD	Expected Time of Departure of the vessel
ATD	Actual Time of Departure of the vessel
ETA	Expected Time of Arrival of the vessel