INTELLIGENT DATA-DRIVEN METHODS FOR DEMAND AND PRICE PREDICTION IN THE SHIPPING INDUSTRY

Ayesha Ubaid

School of Computer Science

Faculty of Engineering and Information Technology

University of Technology Sydney, NSW, Australia

A Thesis submitted in fulfilment

Of the requirements of the degree of

Doctor of Philosophy

May 10th 2021

Certificate of Original Authorship

I, ***Ayesha Ubaid*** declare that this thesis, is submitted in fulfilment of the requirements for the award of ***Doctor of Philosophy*****, in the *FEIT* at the University of Technology Sydney.**

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 10th May 2021

SYNOPSIS

Machine Learning has found its applications in many industrial and commercial domains. However, there are few ignorant industries that still lack digitization and require fusion of AI into their processes. Once such industry is the container-shipping industry. The global supply chain is complex, with cargo volumes that are highly seasonal driven by events, consumer related, Christmas and Chinese New Year, agriculture harvests further impacted by extreme weather occurrences and changes to the Geo, Political regulatory environment affecting trade. In contrast, supply, the shipping capacity is fixed in the short run; this results in periods of mismatched supply and demand and therefore shipping price volatility. The Shipping lines are trapped by the current spot market pricing practice of using date validity and not the vessel voyage. This causes a disconnect between price, demand and supply, a problem that is compounded by the shipping line's enterprise systems. Entrenched operational silos within the lines are resulting in missed revenue opportunities through a lack of real time visibility into the availability of equipment and vessel space which are key inputs to price decisions. Forty percent of all shipping containers moved around the world are purchased on the short-term spot market; their commercial terms are set manually with emails and phone calls. Spot market price should be a product of supply and demand. However, the industry as a whole has little visibility into the state of the market in real time, and carriers are making sub optimal pricing decisions and their customer procurement decisions because of this.

Australian Linear Shipping industry is lacking digitization hence the visibility into the industry statistics is missing. Without real time visibility of the market as a whole, future spot pricing decisions are based mainly on a carrier's internal assessment of current and future booking build up and net contribution targets. This can, and usually does, present a different and misleading picture of market conditions. With the limited market wide

information and inability to use vessel voyage specific price as a lever to steer the cargo opportunity to the optimum vessel sailing, opportunity cost materializes.

In this research, we want to empower Australian Container Shipping Industry with machine learning capabilities to provide a market wide view of current and future demand (both for short-term and long-term), optimal spot pricing model and machine learning based prices prediction model that can predict spot pricing based on current demand and available capacity (supply).

To do so, the first step is to explore and model the relationship between demand, supply, and price. This can provide the current market scenario in which industry is operating. Based on the results inferred, a model is designed to set optimal spot pricing. In order to price effective spot pricing model, statistical analysis is done to discover the relationship between demand, capacity and price. Based on inferred results, historic data and spinning companies' limitations for price quotation, a novel mathematical model is designed that can calculate optimal pricing based on the mentioned factors. Finally, for making contract pricing more effective, the demand forecast performed earlier is used with the available capacity to predict optimal container pricing based on demand and supply using a regression based multivariate machine learning model. This predicted price will shorten the gap between contract and spot pricing in the container shipping industry. Hence, this research provides visibility into future demand and will allow shipping lines and their customers to make better-informed pricing and procurement decisions.

ACKNOWLEDGEMENTS

I would like to acknowledge the blessing of Almighty Allah, the most gracious and beneficent for giving me this opportunity to complete my Doctoral Dissertation under their supervision of Associate Professor Farookh Khadeer Hussain and Associate Professor Asif Qamar Gill.

Additionally, I would like to acknowledge the efforts of my family specially my husband Ubaid and mother, Shaheen without whose support and prayers I would have never been the person I am today. A special thanks to my kids, Suleman and Mustafa who despite their quite young age of 5 and 3 were able to understand that their mum has to study. Their positive and helping attitude made things easy for me. To all of my family, I hope this thesis completion will bring positive changes in our lives.

Moreover, I would like to thank my supervisors whose trust and faith in me made me achieve this goal quite comfortably. The never-ending support, compassion, encouragement and superb guidance are the key to my success. This thesis is as much of their work as mine. Finally, I would like to thank the Centre of Artificial Intelligence (CAI) for supporting me throughout this whole three-year tenure.

I would like to dedicate this thesis to my supervisor Dr. Farookh Khadeer Hussain and my beloved husband Ubaid-UR-Rehman who are my real mentors.

LIST OF PUBLICATIONS

JOURNAL PAPER

- Modeling Shipment Spot Pricing in the Australian Container Shipping Industry: Case of ASIA-OCEANIA trade lane Research Question 1

- Container Shipment Demand Forecasting in the Australian Shipping Industry: Case Study of Asia-Oceania Trade Lane Research question 2 and 3 (under review)

CONFERENCE PAPERS

- Framework for feature selection in health assessment systems. Research Question 1.

- Machine Learning-Based Regression Models for Price Prediction in the Australian Container Shipping Industry: Case Study of Asia-Oceania Trade Lane. Research Question 4.

TABLE OF CONTENTS

Table of Figures

Table of Tables

LIST OF ACRONYMS

ML	Machine learning
UI	User interface
DS	Dataset
FS	Feature selection
LR	Linear R
NNs	Neural networks
AR	Autoregressive
SVR	Support vector regressor
SVM	Support vector machines
GBR	Gradient boosted regression
RFR	Random forecast regressor
RMSE	Root mean squared error
R2 SCORE	R square score
EDA	Exploratory data analysis
MAPE	Mean absolute percentage error
LSTM	Long-term Short-term memory
ARIMA	Autoregressive moving average
SARIMA	Seasonal Autoregressive moving average
PACF	Partial auto correlation function

1 Introduction

1.1 Introduction:

In this chapter, we present an introduction to the thesis. We start by providing an overview of pricing decisions in the container shipping industry and the limitations they are posing on the proliferation of the container shipping industry due to lack of digitization. We further highlight the role that digitization can play in enabling the sector to flourish. Additionally, we explain the current pricing methodology in practice in the container shipping industry and factors affecting the pricing quotation decisions. We also discuss the role that machine learning can play in digitization of the Australian shipping industry.

This research work is part of a research project conducted in collaboration with industry partner 'Mizzen Group', a digital pricing and rate management solution. The Mizzen team combines innovative digital capability in the shipping industry. The company delivers software for freight sellers, shipping lines and freight forwarders to set and distribute prices dynamically in new ways to their customers in the digital channel. Their capabilities enable them to deliver new products with a range of valuable attributes to serve their customers' needs [1].

The project aims to provide a broad industry wide view of the Australian shipping industry’s statistics such as future container shipment demand, contract price prediction based on forecasted demand and available capacity, and spot prices based on container shipment demand and available capacity (supply).

The background and motivation behind this research are discussed in section1.2. Sections 1.3 explains factors affecting the shipment prices followed by research gaps present in the shipping industry in section 1.4. Research scope is highlighted in section 1.5. In section 1.6, significance of incorporating intelligent data driven methods in the shipping industry are identified. Section 1.7 specifies thesis objective followed by their significance in section 1.8. In section1.9, innovations out of this research and its contribution in the research area are discussed. In section 1.10, the stakeholders of this research work are identified followed by thesis organization in section 1.11. Section 1.12summarizes this chapter.

1.2 Background and Motivation:

The need for transportation arises with the need to move goods from one place to another as per consumer demand. Amongst many, sea transportation is one of the cheapest and oldest modes of transportation of goods. The economic growth of every country is directly connected to the magnitude of sea transportation. It steers the economy of the country[2]. Like any other service industry, seasonal events affect the operation of the shipping industry. A container vessel follows a schedule, a set route of port calls, and sailing times and departs a port on a scheduled date regardless of cargo utilization factor. Cargo volumes are very seasonal and are driven by consumer events like Christmas, Chinese New Year, agriculture product harvest, and are also affected by geopolitical events such as a US-China trade war and outbreak of global pandemic like Coronavirus. In contrast, shipping capacity is fixed in the short run as extra capacity cannot be rapidly deployed or withdrawn to match cargo demand; this mismatch in supply and demand creates price volatility.

Pricing is one of the most influential factors in the service industry, like container shipping. The shipping market continues to grow, but freight rates remain stagnant due to inflexible and ineffective pricing strategies [3]. The mismatch in price and market statistics results in revenue loss, and instead of increasing revenue, shipping companies end up with significant income loss. According to a study performed by Deloitte University [4], the shipping industry can increase profit margin by 2 to 7% in just 12 months and yield a 200 to 350% return on investment by setting defective pricing. In order to deal with volatility in shipment pricing, there are two types of pricing strategies currently employed in the shipping sector. Contract market pricing and spot pricing. The contract market offers a fixed price for a known cargo task over a set period, with secured booking space in periods of high demand. The spot market has a fluctuating price; you can benefit from lower than contract rate prices in the low season, but face escalating costs and less certainty of being able to secure the booking on the vessel voyage you want in peak season, as space is reserved for contracted customers [5]. Current industry pricing practices and commercial product offerings deliver sub-optimal outcomes for the shipping lines and, indeed, for their customers. Freight rates are set with a date validity and not by vessel voyage. Any vessel can be booked in that date range at the same price, regardless of the utilization factor.

Therefore, there is a miss-match between price, shipping capacity (supply) and demand. Besides, there is very little market-wide data on real-time cargo demand, shipping capacity, and price, resulting in pricing decisions based on a shipping line's internal data only. Pricing is a best estimate only and not driven by broader, real-time data. Moreover, shipping lines offer limited product choice for a customer. A customer can take the price flexibility of the spot market but then has no booking certainty in peak season. They can trade off certainty with a contract, but then they forego price flexibility. A lack of contractual obligation compounds these factors at the time of booking enabling shipping lines customers to book cargo, often the same shipment with multiple carriers and they can cancel at the last minute without penalty. To counteract the explained scenario, shipping lines have adopted a practice of overbooking. In peak season, cargo cancellation does not exceed overbooked vessel cargo-carrying capacity. Therefore, not all the shipments can be loaded on a vessel and are left behind or "rolled" until the next vessel departure [6].

The main objective of shipping line spot pricing is to maximize the profitability of each vessel voyage. Internal factors affecting spot pricing includes the vessel utilization factor in the coming weeks for the service (a measure of supply and demand), the port pairs, equipment deficit, and balanced or surplus and global network requirement impacting the key variable cost component of equipment imbalance charge, weight of the cargo, and several containers to be booked. External factors driving the spot pricing includes current market conditions, actions of competitors. In this process, shipping lines need to manage the seasonality and fluctuating cargo demand in the operating year and ensure they have a base load of contract business and fill the remaining allocation with the spot market. This cargo mix will vary with the shipping line to the shipping line and trade lane to trade lane based on each line's strategy and the inherent characteristics of a market's structure. However, regardless of the cargo mix, in the short run, the spot market is the only channel for a sipping line to improve its profitability because the rate has not been fixed for a time like the contract market [6]. The process flow diagram of price setting in the shipping industry is shown in Figure 1.1. From figure 1.1, it is clear that it is only a single person/consultant who handles the price setting. When a customer calls a shipping company, the company representative gets a quotation from the consultant after liaising with the customer. The consultant's capability to respond to the request optimally and promptly has a direct impact on the company's profitability. The consultant's response is based on experience, the judgment of demand and capacity at a particular time of the year and static price guide. Thus, such an important decision is made without sufficient data and involvement of other critical personnel in the company. Such practices increase the risk of missing opportunities to increase net profit margin.

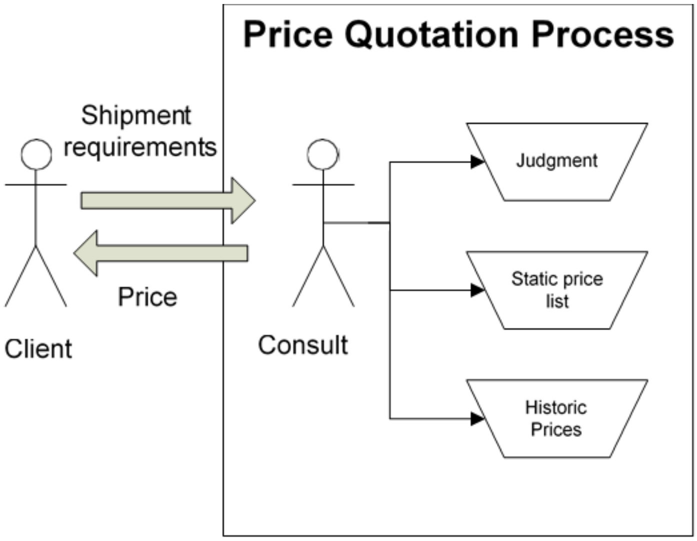


Fig. 1-1 Price Quotation Process in the Shipping Industry

1.3 Factors Affecting Shipment Pricing:

Traditional pricing strategies are quite ineffective and tangential. The industry has no digital view of current cargo demand, available capacity (supply) and pricing. The deep-rooted traditional system for pricing in the shipping industry is stagnant as fleet capacity and demand grow, but prices remain constant. Thus, despite continuous market growth, there is little capitalization. In the absence of flexible pricing methodologies based on current demand and capacity, shipping companies are unable to adjust their rates in response to market fluctuation (i.e., changing demand and capacity over time). To deal with the issue, companies tend to reduce shipment costs. This results in even smaller financial gains with every round of price deduction. Hence, companies need to shift their focus from cost-cutting towards growth. This can only be achieved via introducing modern technologies into the system and by having a market-wide view of all the factors affecting price setting. The following are some of the factors affecting shipment pricing.

1.3.1.1 The complexity of rate formulas:

The rate-setting is quite complicated and requires cause and effect modeling of many factors such as current demand, available capacity, fuel prices, seasonality, economic event occurrence, and global pandemic occurrences such as Coronavirus (COVID 19). This becomes humanly impossible to incorporate the effect of all such factor to fix the pricing [4, 7]

1.3.1.2 The complexity of individual contracts:

Every customer has different requirements, and so makes the contract. Providing discounts to one of the loyal customers can have a high chain impact on the company business, which may not be felt immediately. Visibility into such a contract can provide long-term advantages [5].

1.3.1.3 Poor visibility to shipping utilization:

Lack of visibility into the utilization of the shipment have quite long-term adverse effects on the company's performance. The consultants have no idea if the ship's usage is worthwhile or not. Thus quoting less pricing in case of low utilization is a wholesome loss [7, 8].

1.3.1.4 Inability to respond to market forces:

Fuel prices, change in seasonality of cargo movement due to port congestion or closure care beyond human imagination. This means quoting prices without considering such matters can lead to profit or loss.

1.3.1.5 Resistance to technology:

Numerous shipping companies struggle with technology boundaries. Some still work with spreadsheets while others' systems may not be dynamic or flexible to cope with contract variations; in order to keep an edge over the competitors, adoption of technologies proactively in a must.

1.3.1.6 The shift from cost to value:

Every customer requesting a service, wants it to be pain-free. This value provided to customers becomes a competitive differentiator. In this case, revenue can be increased without having an increase in cost. Offering excellent value for money can attract more customers [3, 5, 9].

1.4 Gaps in the Current Shipping Pricing Methods:

Keeping in view the current; old fashioned operations along with the hitches they are posing into the activities of the industry discussed in previous sections we can clearly identify a few gaps in the current shipping methodology. These include:

- **I.** There is a need to understand the relationship between shipment demand, capacity and pricing. Since the relationship is vague, shipping industry is lacking visibility into the industry statistics and is thus loosing worthwhile profit.
- **II.** There is a need to develop a data driven model that can calculate optimal spot price for a shipment container based on current shipment demand and available capacity. This calculation of spot pricing is vital for earning revenue and helping improve businesses.
- **III.** There is a need to develop shipment demand forecasting capabilities for the shipping industry in order to help stakeholders plan ahead with trusted industry statistics to take future supply chain and pricing decisions.
- **IV.** Furthermore, there is a need to equip a shipping company with price prediction models that can help the industry. Stakeholders could even quote optimal contract pricing closer to the spot pricing.

1.5 Research/Thesis Scope:

In the previous sections, we have discussed the pricing methodology currently in use in the shipping industry along with the limitations imposed on the shipping industry. In the previous section, we also highlighted the research gaps we identified in the existing system. Based on the mentioned findings, we aim to utilize data driven (machine learning and statistical) models for digitizing the Australian Shipping Industry. In this research work, we aim to:

- **I.** Explore the relationship between current shipment demand, available capacity, and spot pricing.
- **II.** Utilizing the relationship between shipment demand, capacity and spot pricing identified in (I), we further plan to mathematically model optimal spot pricing based on the inherent connection between demand and capacity.
- **III.** Furthermore, in our aim to provide industry transparency, we intend to implement a ML model that can predict future shipment demand for both the short and long-run. Using the expected demand, the industry stakeholders can foresee the future shipment demand for informed supply chain decisions.
- **IV.** Exhausting the visibility delivered by (III) in regards to shipment demand, we furthermore aim to design a contract price prediction model that can predict container shipment prices based on the current shipment demand and available shipping capacity.

In this thesis, we aim to contribute to the theoretical and practical body of knowledge by implementing intelligent data-driven methods for demand and price prediction in the shipping Industry. However, initially, we have limited our scope to the Australian shipping industry. We intend to collect real-time data from the Australian shipping industry's primary stakeholders. To further narrow down the scope of our research work, we have selected 1 out of 7 trade lanes to perform, analyze and demonstrates our hypothesis. These seven trade lanes include Asia-Oceania, America, Africa, Europe, Indiana, Middle East, Pacific Ocean islands. Since the Asia-Oceania trade lane is the busiest operating trade lane, it is ideal for performing analysis. To further narrow the research, we have selected a single trade lane, the Asia-Oceania trade lane (see figure 1.2).

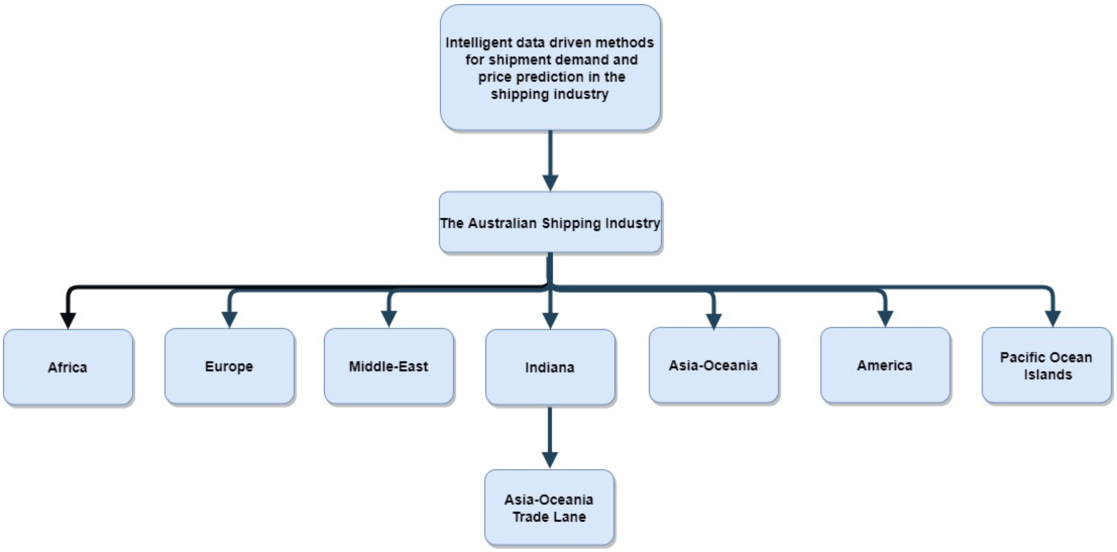

Intelligent data driven methods for shipment demand and price prediction in the shipping industry
The Australian Shipping Industry
Africa
Europe
Middle-East
Indiana
Asia-Oceania
America
Pacific Ocean Islands
Asia-Oceania Trade Lane

Fig. 1-2 Thesis Scope

The Australian Shipping industry is also lacking the digitization. Hence, the visibility into the industry operations is missing. Without real-time visibility of the market statistics as a whole, future pricing decisions are based mainly on carriers' internal assessment of current and future booking build up and net contribution targets. This can, and usually does, present a different and misleading picture of market conditions. With the limited market-wide information and inability to use vessel voyage specific price as a lever to steer the cargo opportunity to the optimum vessel sailing, opportunity cost materializes. To achieve the discussed goal, machine learning (a subfield of Artificial intelligence) can be used to provide a market-wide view of the factor-determining price and help in decision-making.

1.6 Significance of intelligent data driven model for the shipping industry:

Machine learning (ML) is a sub-branch of artificial intelligence that is used to teach machines to learn from underline data, the hidden trend, and insight saved into it. ML has found its applications in many industrial and commercial domains, from medical to military[10]. A few of the business advantages ML can provide in the shipping industry are explained in the next section.

ML can be used to generate a digital view of the market that can provide end-to-end visibility into the market statistics and thus, full control over the operations. This data analysis to predict demand, prices, and customer behaviors, taking into account all rules and constraints that might affect the business. Such digitization of the business allows KPIs achievement as well as other benefits like:

1.6.1 Visibility into industry Stats:

ML can help provide an industry-wide view of container shipments' current demand, available shipment capacity.

1.6.2 Future Planning:

Forecasting capabilities achieved using ML models can help industry stakeholders plan ahead for future operations[6].

1.6.3 Substantial financial impact:

Price optimization can bring worthwhile financial benefits. Data-based solutions can benefit from pricing solutions quite fast[6, 7].

1.6.4 Productivity:

Time can be more productive whilst using technology. The system performs all the price optimization while the consultant can focus on other value-added services such as evaluation of the tender process, managing pricing campaigns, and reviewing pricing strategies.

1.6.5 Improved consistency and accuracy:

Since data-driven models are guided by historical data, the consistency and accuracy of the market view are very reliable. This promotes customer's trust in the business [4, 6, 7].

1.6.6 Faster and Assured decisions:

Consultants can make decisions faster and can create scenarios to see the consequences of the made decisions. This can improve the overall profitability of the business and help in informed decisions.

1.6.7 Consistent pricing:

Using data-driven methods for market analysis and operations allows businesses to quote consistent pricing based on the utilization factor. It has two-fold advantages; it earns the customer's confidence and provides maximized profit for each voyage.

1.7 Thesis Objectives:

This research aims to empower the current manually operating Australian shipping industry with data-driven capability for forecasting short-term and long-term demand, price prediction, and spot price calculation. This will embed digitization into the shipping industry and will fill the existing research gap. We have summarized the objectives of this thesis as follows:

- **I.** To model the relationship between demand, supply, and price using statistical methods. Moreover, to calculate spot prices based on historical data, demand, and supply.
- **II.** To forecast short-term demand for Australian Container Shipping while incorporating holiday and seasonality effects using Machine Learning Models
- **III.** To build a long-term demand forecasting model for Australian Container Shipping Industry using Machine Learning Models

IV. To suggest a price for spot market using historical and future supply and demand forecast using machine learning models.

V. To verify and validate the research questions as described earlier.

1.8 Significance of the Thesis:

This thesis presents data-driven methods for digitizing the container shipping industry using statistical and machine learning-based models. However, it should be noted that the research has been conducted specifically for the Australian Container shipping industry and data from only one trade lane i.e., Asia Oceania trade lane is used for verifying the hypothesis. Additionally, the only factors incorporated in this research are demand, capacity (supply) and the prices. No competition between shipping lines and other business-related issues have been considered.

To the best of our knowledge, there are no ML-based systems incorporated into the Australian Shipping Industry, does there exist any real-time visibility into the business. The research conducted in this novel study would help the Australian shipping industry gain a broad market wide view of the business. Following ae the significant contributions of this thesis:

1.8.1 Novel Feature Selection Framework:

This framework will allow the data scientist to select original features from the dataset efficiently by allowing domain experts and other relevant parameters incorporation into the framework. No such feature selection framework has been proposed so far that can incorporate user-defined parameters and domain expert input into the feature selection (FS) process.

1.8.2 Modelling of the relationship between price, demand, and supply:

The statistical modeling allows shipping companies to get a realistic view of current market situations and help them analyze what factors are negatively affecting their businesses identified by the study. The study will open up complete market statistics for a better understanding of the supply chain situation.

1.8.3 Model to forecast short-term demand in the Australian Shipping Industry:

The demand-forecasting model designed will provide a broad market view of future demand to shipping companies, allowing them to set spot prices based on current demand. There is no forecasting capability available until now to forecast demand for the Australian Shipping industry. The short-term demand-forecasting model will predict six weeks ahead of demand.

1.8.4 Model to forecast long-term demand in Australian Shipping Industry:

The year ahead demand forecasting will allow shipping companies to plan for their capacities and their price to achieve business advantage. This model will be the first-ever demand prediction model that will provide the AUS Shipping Industry with the capability of forecasting one year ahead of demand.

1.8.5 Model to predict the price based on current demand and available supply:

The very first price prediction model for the Australian shipping market will be designed. This will help provide shipping industry stakeholders to make informed decisions regarding pricing.

1.9 Innovation and Research Contribution:

I. A novel hybrid feature selection framework was also proposed, which is capable of selecting features in a minimum number of iterations with equitable classification accuracy and computation time. Furthermore, the framework can incorporate input from the domain expert for selecting feature subsets. This offers excellent flexibility in the proposed framework. Also, the proposed framework has the capability of discarding useless features from the remaining feature set.

II. The Australian Shipping industry is missing a demand forecasting tool that can forecast future demand for helping them make an informed pricing decision. The designed algorithm will be specific for their dataset and will help them forecast demand for both short term and long term.

III. The Australian shipping industry lacks intelligence in their pricing system. The research will develop an ML model for price predictions based on demand and supply. This can empower the shipping industry with AI and help businesses increase their revenue by making informed decisions.

1.10 Stakeholders:

This research is highly practical in nature and is conducted in collaboration with industry partner, Mizzen group. Based on the application of the conducted research, the industry stake holders includes shipping companies, their clients, ship builders, port terminal employees, ship owner, charters and operators, suppliers and business partners, insurers and local and indigenous communities.

1.11 Thesis Organization:

In this thesis, we present a complete roadmap for digitizing the Australian sipping industry. In order to achieve its objectives, we have organized this thesis in nine chapters. In this section, we give a summary of each chapter.

Chapter 2: This chapter provides an extensive literature review of existing feature selection methods for selecting useful features from a dataset to be used in ML models, a survey of the proposed spot price calculation for container shipments, an overview of forecasting methods presently employed for time series forecasting and ML-based multivariate prediction models. We have also identified the research gaps in the literature.

Chapter 3: Chapter 3 explains the research problem definition and presents the research overview.

Chapter 4: Chapter 4 explains research methodology and solution overview.

Chapter 5: Chapter 5 provides insight into the data sourcing process used for data collection for the thesis. Moreover, this chapter discusses various data-processing techniques for understanding and making datasets ready for data-driven methods.

Chapter 6: Chapter 6 provides statistical analysis to identify problems in price setting in the Australian shipping Industry. This chapter also describes a novel model to set optimal spot pricing in the container shipping industry based on demand, capacity, and historical data and pricing thresholds used by industry.

Chapter 7: This chapter describes short-term and long-term cargo demand forecasting models employed to forecast demand for the Australian container shipping industry.

Chapter 8: This chapter enlightens the price prediction model designed for predicting price based on demand and available capacity for the Australasian container shipping industry.

Chapter 9: Chapter 9 concludes this thesis by providing a summary of the results in this thesis, along with potential future work.

1.12 Summary:

In this chapter, we have provided the introduction, background and motivation behind the conducted research. We have explained the pricing methodology effective in the shipping industry and have highlighted the importance of application of DATA driven methods into the shipping industry.

Based on current operational practices, we have identified the potential research gaps present in the industry. Keeping in view of the identified gaps, research objectives are formalized and presented in the chapter. In addition to this, we have also formalized the scope of this thesis based on the objectives identified. Moreover, the significance of the thesis and the innovation out this thesis is also discussed. The relevant stakeholders are also identified in the next section. Finally, the plan of the thesis is presented. In the next chapter, we present a detailed overview of the existing literature review on data driven methods in the shipping industry.

1.13 References:

1. Mizzen Group Pty Ltd. Available from: https://www.mizzengroup.com/.

2. Y.H.V. Lun, K.-H.L., T.C.E. Cheng, Shipping and Logistics Management. 2010, London: Springer,.

3. Akman Biyik, C., Pricing in Liner Shipping Industry: A Review and Assessment. 2017.

4. Larry Montan, T.K., Julie Meehan, Getting Pricing Right The value of a multifaceted approach. Deloitte University Press.

5. Akman Biyik, C. and M. Tanyeri, Pricing Decisions in Liner Shipping Industry: A Study on Artificial Neural Networks. 2018.

6. Drewry. Technology to reduce freight rate volatility and capacity risks 2019; Available from: https://www.drewry.co.uk/white-papers.

7. Pricing for Profit in Container Shipping. April 2016.

8. Chen, R., J.-X. Dong, and C.-Y. Lee, Pricing and competition in a shipping market with waste shipments and empty container repositioning. Transportation Research Part B: Methodological, 2016. 85: p. 32-55.

9. Gelareh, S., S. Nickel, and D. Pisinger, Liner shipping hub network design in a competitive environment. Transportation Research Part E: Logistics and Transportation Review, 2010. 46(6): p. 991-1004.

10. Dey, A., Machine Learning Algorithms: A Review International Journal of Computer Science and Information Technologies, 2016. Vol. 7 (3)(1174-1179).

2 Literature Review

2.1 Introduction:

In this chapter, we present an overview of the existing literature on intelligent data driven methods applied in the container shipping industry. In the next section, for both discussion and evaluation purposes, we divide the existing literature into four different classes based on the identified thesis objective in chapter1. These include, (1) time series forecasting enlightened in section 2.5, (2)Feature selection methods to select required features from vast shipping dataset described in section 2.6,(3)spot price calculation based on cargo demand and capacity explained in section 2.7 and (4) Price prediction models employed in the container shipping industry to predict contract pricing expounded in section 2.7.

In addition to these descriptions, we have also identified the research gaps in their respective sections.

2.2 Classification of Literature Review:

Based on the identified research gaps, the literature review for this thesis can be divided into four different subparts. The literature review includes the study of (1) Time-series forecasting models, (2) Feature selection techniques for extracting features from the shipping dataset, (3) Shipment spot price calculation methods focus on demand and capacity as a deterministic factor and (4) ML-based prediction models, both generic and with a focus on the shipping industry. Figure 2.1 below shows the tree for the literature review.

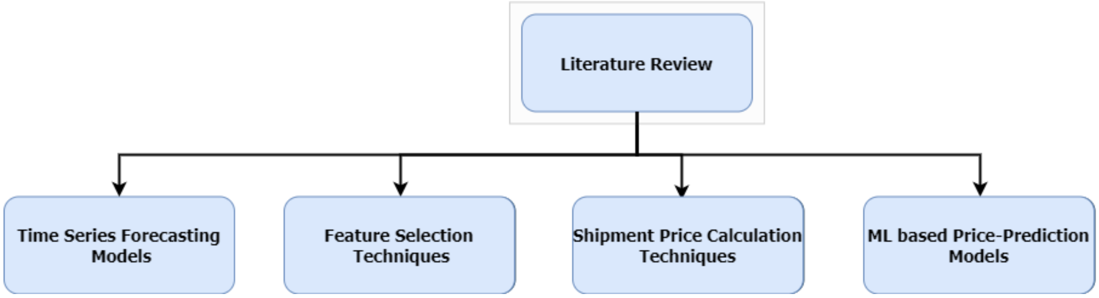

Literature Review
Time Series Forecasting Models
Feature Selection Techniques
Shipment Price Calculation Techniques
ML based Price-Prediction Models

Fig. 2-1 Literature Review Classification

2.3 Time Series Forecasting:

In literature, time series forecasting can be defined as:

“Prediction of events through a sequence of time”. [1]

Time series forecasting uses previous time-based values from the past to determine the probable future values. Time series forecasting has varied applications such as weather forecasting, earthquake forecasting, signal processing, pattern recognition and many more.

The literature regarding forecasting is quite vast and diverse. Researchers have investigated and developed different forecasting techniques and their combinations to achieve better performance and accuracy. Prediction algorithms have been used in industries ranging from business, industrial engineering, medicine, physics, and statistics to foresee future events [2]. A forecasting algorithm uses information from past experiences to anticipate future events. These algorithms can be used in situations where data from a specific sample space can be collected over a period. This makes prediction algorithms ideal for use within smart environments[1]. Forecasting has become an essential activity in business these days to make better and more informed decisions. The prediction of future events allows companies to proliferate by developing sustainable solutions [3]. Forecasting can be defined as "a planning tool that helps management in its attempts to cope with uncertainties of the future, relying mainly on data from the past and present and analysis of trends" [5]. Companies can improve their performance and competitive position and can also achieve high yield levels by implementing well-constructed forecasting models into their process [4].

2.4 Forecasting Process:

The success of predictive models is critically dependent on its accuracy. Low accuracy can lead to extremely misleading results, causing costly damage to the business. Hence, it is of great importance to monitor forecast errors by selecting suitable validation metrics. Forecasting must be done using a verified process in order to make valuable predictions.

In [5], authors presented a forecasting process framework with an emphasis on implementing forecasting techniques into fundamental processes of planning and assessment. According to Schultz, forecasting should be evaluated in areas of model building. The process starts from a corporate level to provide better visibility to decision making. In the second phase, the model is built, followed by its evaluation. The evaluation of the model is based on quantifiable measures such as sales, cost, and profit. Figure 2.2 shows the forecasting process, as outlined by Schultz.

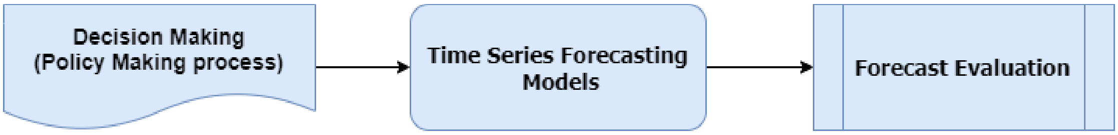


Fig. 2-2 Schultz Forecasting Model

Shima and Siegel [6] proposed a new forecasting process comprising six steps. In the first step, the purpose of forecasting is defined. In the second step, the time horizon is selected, followed by selecting the forecasting methods. In the fourth stage, data sourcing is undertaken, and in the fifth stage, forecasting is performed. The last stage of the framework is monitoring the forecasts as shown in figure 2.3.

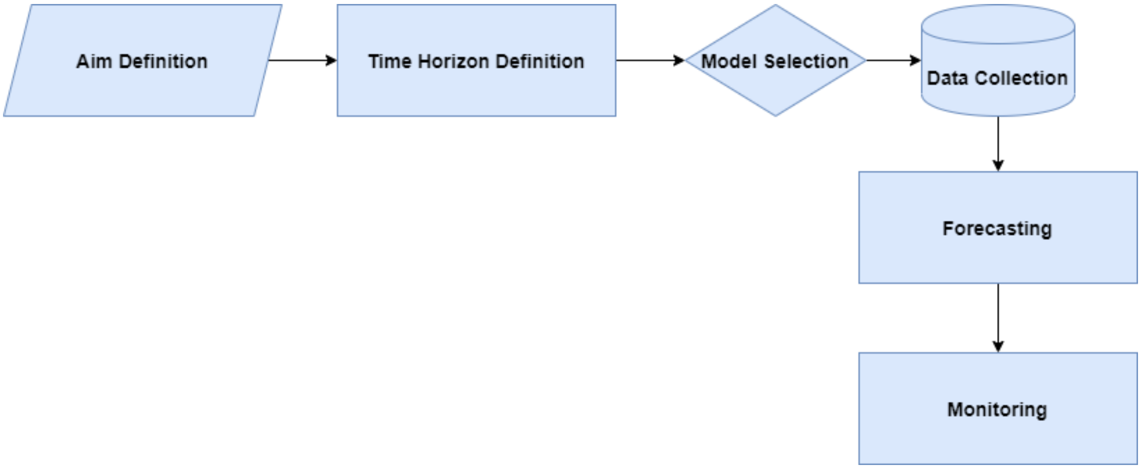


Fig. 2-3 Shima and Siegel’s forecasting Model

Brockwell and Devis in 2010, presented another approach specific to time series forecasting [7]. In the first step of their forecast process, data features are examined using plotting techniques to check trends, seasonality, and other data components to extract stationary residuals. This step is followed by fitting the model. Figure 2.4 shows an overview of Brockwell and Devis' forecasting process.

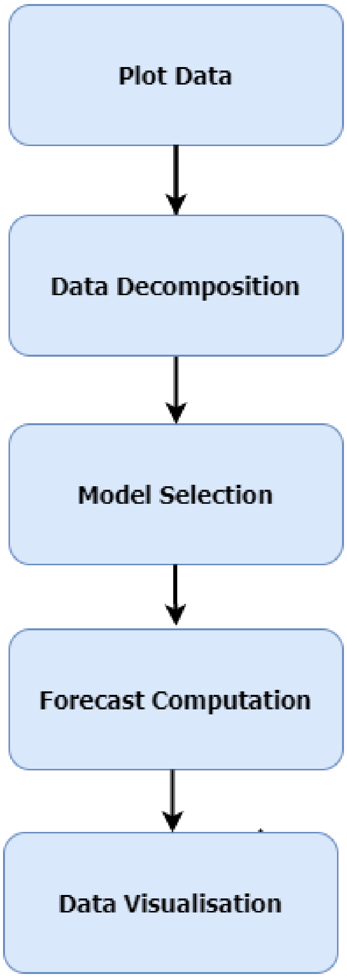


Fig. 2-4 Brockwell and Devis' forecasting model

2.5 Existing Time Series Forecasting Models:

Carrying out forecasts in time series data has been a general problem for a long time. A time series allows us to predict future values depending on the components of the series from historical data. Moving average is the simplest forecasting method. It calculates the average sample observation and provides forecasts for the next period based on the calculated average. For each new sample, there is a newly calculated average, and the previous one is removed. Thus, a forecast is computed for every new data observation [8]. The method can generate entirely accurate forecasts for time series with regular trends. A series where trends change with time may provide false forecasts [9]. The classification of time series forecasting algorithms is shown in figure 2.5.

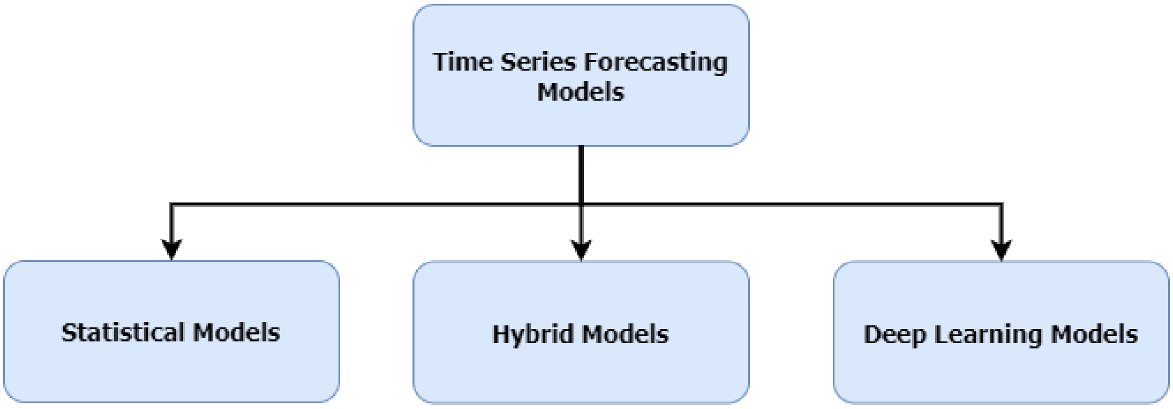


Fig. 2-5 Classification of time series forecasting models

2.5.1 Statistical Models:

The weighted moving average is a variant of a simple moving average. In this method, weights are assigned to the most critical period. The higher the weights, the more critical the data values. This method is more sensitive to trends [10]. Simple exponential smoothing (SES) assumes that forecasted data have fluctuations around a constant level over time [11] . A variant of exponential smoothing is the Holt-Winters non-seasonal method. It includes a trend term that measures the expected increase or decreases per unit period at the local mean level. The Holt-Winters seasonal method is an extension of the Holt-Winters non-seasonal method. A smoothing factor for each period of the year is added to adjust the forecast according to the expected seasonal fluctuation[12]. Box and Jenkins in the 1970s presented autoregressive (AR) and moving averages (MA) models for time series predictions [13]. AR considers the current values of a time series as the linear combination of its past values. However, MA is a function of random interference that affects the series. The proposed models proved to be quite useful for predictions in their initial era. As the research continued, it was noticed that there are situations where time series does not follow linear trends. Thus, a range of new models has been presented to cater to these needs[14]. The autoregressive integrated moving average (ARIMA) is most commonly used for time series forecasting[14, 15]. ARIMA exploits dependency between an observation and a residual error from a moving average model applied to a lagged observation. It does so by utilizing the relationship between observation and lagged observations. It makes time series stationary by subtracting an observation from previous observations. The ARIMA model has variants such as seasonal ARIMA (SARIMA), which caters to the seasonal variances in a time series and ARIMAX, which handles the data points' covariance in a time series.

2.5.2 Hybrid Models:

The class of hybrid models constitutes the models which are a combination of machine learning models and statistical models to more effectively cater to both linear and nonlinear data. Artificial neural networks (ANNs)[15-17] are also found to be very efficient for catering for the non-linearity of time series. A support vector regressor (SVR) can also handle the nonlinear part of the time series well. In [17], a novel method for energy demand prediction using SARIMA and support vector regression is performed. SARIMA handles the linear data component while SVR handles the nonlinear data components. Hybrid linear and nonlinear models are also employed for time series forecasting. Much focus was on ANN and ARIMA models. ARIMA models handle linear data components, while ANN models handle nonlinear parts. In [14], ANNs are used for one month ahead of price prediction in the liner shipping industry. The liner shipping industry is volatile and is impacted by seasonal variations, public holidays, and travel routes. According to [18], ANNs can handle the volatility of the shipping industry and provide promising forecasts. However, ANNs suffer from overfitting problems. In (Chou, Chu & Liang 2008), a new regression-based model is designed to forecast shipping container volumes, i.e., supply. The author claims that the designed regression model can cater for non-stationary parts of time series. In (Han et al. 2014), SVM is used to forecast the dry bulk freight index.

2.5.3 Deep Learning-Based Models:

A new area of deep learning has been explored for time series analysis [15]. Spot electricity prices are predicted using deep learning methods. The author proposed four deep learning models to perform time series analysis for spot electricity price prediction, which include deep neural networks (DNNs), hybrid long-term short-term memory DNN (LSTM-DNN), hybrid GRU-DNN and convolution neural networks (CNNs). The study infers that deep learning methods outperform statistical and ANN-based models. In [16], research was conducted on Bitcoin price prediction by comparing LSTM, RNN, and ARIMA. The results from the deep learning models as compared to other classes of models are more promising.

2.5.4 Comparative Analysis of existing time series techniques:

Based on the literature review in regard to a time series forecasting model, it is evident that plenty of work has been done in varied domains to perform predictions based on time series. Various models from different classes are designed and applied in different domains. However, the application of any of the time series models is still scant and there are a limited number of models that are applied/designed to this domain. To the best of our knowledge, there exist no forecasting capability in the industry that can predict future demand based on seasonality and trends from the past data. In order to cover this gap, we have applied existing time series forecasting models that can handle trends and seasonality present inherently in the dataset. The table 2.1 below shows the existing time series forecasting models and their application in the container shipping industry.

Table 2-1 Summary of Time Series Models Application in the Container Shipping Industry

Models	Application on shipping industry	Domain	Supporting Models
Simple exponential smoothing (SAS)	✕		
Holt Winter's non-seasonal method	✕		
Holt Winter's seasonal method	✕		
Auto Regressive (AR)	✕		
Moving Average (MA)	✕		
Auto Regressive Moving Average (ARIMA)	✓	Price prediction	ANNs
Seasonal Auto Regressive Moving Average (SARIMA)	✕		
Fb PROPHET	✕		
Artificial Neural Networks (ANNs)	✓	Price Prediction	ARIMA
Support Vector Regressor (SVR)	✓	Price Prediction	ARIMA
Support Vector Machines (SVM)	✓	Freight index.	Nil
Long Short Term Neural Networks (LSTM)	✕		
Recurrent Neural Networks (RNN)	✕		

2.6 Feature Selection Techniques:

Innovations in technology have led to the collection and storage of massive amounts of data. This data can be used for knowledge discovery. Based on this data, different hypotheses can be made. The stored data can be used as input to the ML algorithms to make predictions for future events. To use the data in the ML algorithms, we need to select the most influential or prominent data features that are termed features. Features may be defined as the quantitative properties of the process being observed. Feature selection (FS) is selecting subsets of relevant variables from the original dataset, representing the data effectively [19]. It helps in understanding data and reducing computation time, noise, and data dimensionality, resulting in increased prediction performance [19]. FS does not create new features; rather, it helps in obtaining a subset of existing features. If two features represent the same meaning of data, only one of them is selected. Figure 2.6 shows the generic feature selection process.

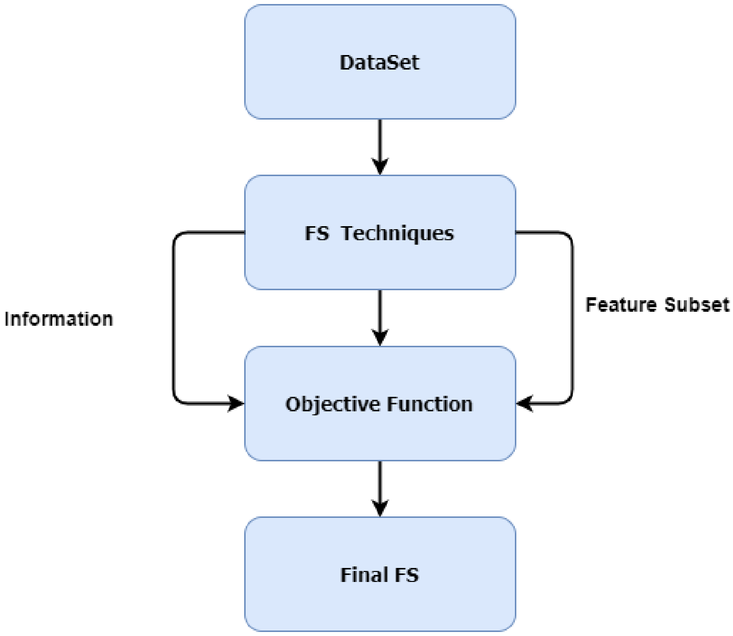


Fig. 2-6 Generic feature selection process

At the start of the feature selection process, the full dataset is fed into the FS algorithm which splits the dataset into subsets of the selected features. These subsets are used as input to the objective function to evaluate their effectiveness. If an objective function is achieved, the selected subsets become part of the final feature set. However, if objective function is not achieved not, the FS process continues until it is achieved. The final feature set is then used in several ML algorithms for algorithm training. There are different methods for feature selection. These include filters, wrappers, embedded methods, and classification algorithms [20]. A feature is said to be useful if it has unique information about the classes in the dataset [19]. Features can be discarded only if they do not influence class labels [21]. Filters are pre-processors for feature ranking. Features with a high ranking are selected as input to the prediction algorithms.

Wrappers, on the other hand, depend on the performance of the predictors [19]. In addition to these two methods, there are embedded and classification algorithms that are also useful for feature selection as shown in figure 2.7.

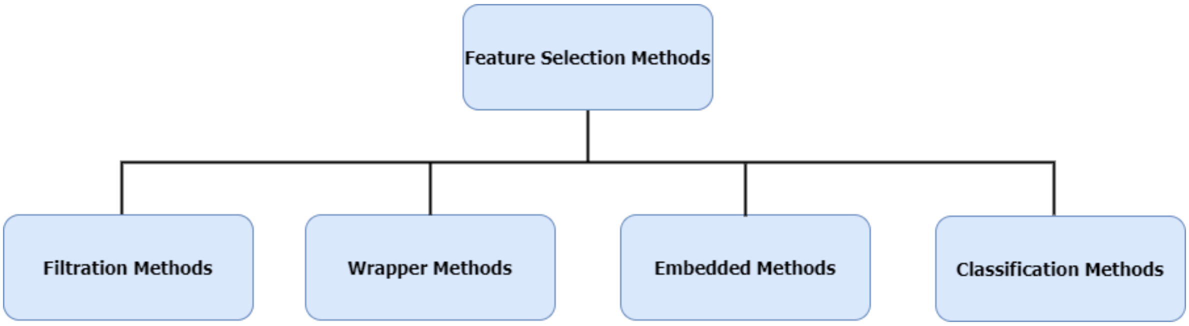


Fig. 2-7 Classification of feature selection methods

2.6.1 Filtration based methods:

Filter methods are used as primary selection criteria and are also known as ranking methods [22, 23]. They are used to rank a variable's usefulness. They set up a threshold, the features below, which are discarded by filtering out irrelevant data. Correlation criteria are the most well-known and commonly used filter method. The Pearson correlation coefficient is used to detect the linear dependencies between input (variable) and output (class). Mutual information (MI) is another filtration technique. It is a measure of the dependencies between two variables. MI can be calculated for both discrete and continuous variables. MI between X and Y is zero if they are independent and is greater if they are dependent. MI alone provides unsatisfactory results, so it is used in combination with embedded methods to achieve better results.

2.6.2 Wrapper Methods:

Wrapper methods rely on classification algorithms to produce results. Wrappers use classification models as a black box and their performance as the objective function to select a subset of features [19]. These can be divided into two classes, sequential selection algorithms and heuristic search algorithms [24]. Sequential selection algorithms are iterative. Sequential forward selection (SFS) starts with an empty feature set in the first iteration, and the empty feature set is filled with a single feature to obtain the maximum value of the objective function[25]. Starting from the second iteration, each feature is added into the subset, and its effectiveness concerning the objective function is measured. Added features take a permanent position in the feature set if they provide the maximum value of the objective function. Sequential backward selection (SBS) starts with a full dataset and sequentially removes from the feature set those features whose removal has a minimum effect on predictor performance. The sequential floating forward selection (SFFS) algorithm, proposed in [25], adds a new step in basic SFS, which is to remove a feature from the existing subset and check if it provides the maximum value of the objective function. Another variant of SFS is the adaptive SFFS algorithm (ASFFS), which reduces the correlation amongst the feature set. A parameter r is calculated, which specifies the number of features added in the next inclusion[19].

The parameter r is selected randomly. The plus-L-minus-r search algorithm is similar to naïve SFS. At each cycle, L variables are added, and r variables are removed from the dataset's subset. L and r are selected randomly. Heuristic search algorithms include genetic algorithms (GA) as explained in [26] and can be used for feature selection. In GAs, chromosome bits show if a feature should be included or not in the feature set. The parameters of GA can be changed based on the data and the application. A modification of GA called CHCGA is proposed in [14]. The CHCGA selects the best N individuals from a pool of parents. Better offspring replace less fit parents.

2.6.3 Embedded Methods:

The embedded methods explained in [20] reduce the computation required to classify different subsets. They combine feature selection and the training process. In[27], a greedy search algorithm is used to find MI between the features and output class and between the subsets. Max-relevancy min-redundancy (mRMR) is another embedded method, similar to the greedy search algorithm.

2.6.4 Classification Methods:

Classifiers can also be used for feature selection. Weights are assigned to the features, and the linear relation between input and output is established. Such a feature removal technique is called recursive feature elimination. It is used in SVM and is called SVM-RFE. The same methodology can be adopted in neural networks (NNs) and linear regression. Clusters, e.g., K-nearest neighbors (KNNs), help grouping features that are naturally related to each other. Details of unsupervised methods for feature selection can be found in [28, 29],[30-32]. Semi-supervised methods for feature selection are investigated [28],[33]. In addition to these techniques, the domain expert's role is also of great importance in certain areas, such as the medical and finance fields. Input from domain experts while selecting features can improve the selection process exponentially.

2.6.5 Comparative Analysis of existing feature selection techniques:

Grounded on the literature review in relation to feature selection methods designed, implemented and applied so far it is apparent that there exists no such feature selection method that can incorporate domain expert’s input into the feature selection process and that is required to extract the features from shipping data set having vast and varied features. Hence in this research we aim to design a new feature selection method that can extract features quickly and as directed by domain expert. In table 2.2 below, we have summarized the existing literature review for the readers of this thesis.

Table 2-2 Summary of Feature Selection Method's Application in the Container Shipping Industry

Methods	Application in shipping industry
Pearson Correlation	✕
Mutual Interference (MI)	✕
Sequential selection algorithms	✕
heuristic search algorithms	✕
Sequential forward selection	✕
Sequential backward selection	✕
Sequential floating forward selection (SFFS)	✕
Plus-L-minus-r search algorithm	✕
Genetic algorithms	✕
Greedy search algorithm.	✕
Support vector machines- recursive feature elimination (SVR-RFE)	✕
Neural networks (NNs)	✕
K-nearest neighbors (KNNs)	✕

2.7 Optimal Price Calculation Techniques:

Estimations are critical concerning making business decisions. Researchers have shown that shipment price estimation is quite complicated in the shipping industry [12]. Hence, the critical idea behind shipment price estimation for the container shipping industry is not to trust estimations for price-setting; instead, it has more to do with developing plans [13]. Modeling optimal pricing in the shipping industry is an understudied area. However, in the recent past, a few studies have been conducted to address the issue. Based on these studies, the shipment pricing methods can be classified into two major classes, Scenario-based pricing methods, and Algorithmic pricing, as shown in Figure 2.8.

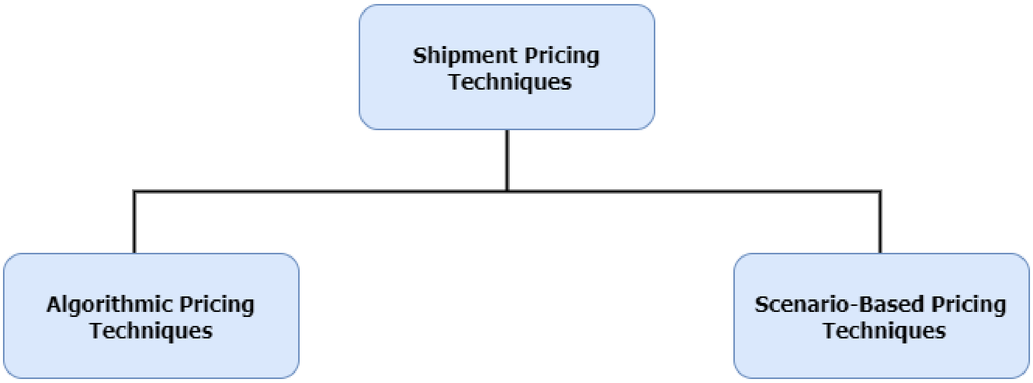

Shipment Pricing Techniques
Algorithmic Pricing Techniques
Scenario-Based Pricing Techniques

Fig. 2-8 Classification of shipment pricing methods

2.7.1 Scenario-based Pricing:

In this group of pricing studies, authors have made assumptions of the shipment routes, number of carriers/forwarders and have proposed the pricing strategies based on supposed specific shipping scenarios. In [34], the authors studied pricing decisions in the container industry by taking into account empty container repositioning costs and obtained some analytics properties. They worked on two closed- port systems to simplify their research problem. Xu et al. [27] developed the work in [34] by extending the research scenario to one carrier, two forwarders, and one shipper. The author built a mathematical model to study the price determination method followed by carriers and forwarders. According to Chen, minimum and maximum pricing thresholds exist, and pricing is quoted within the threshold limit. The forwarder pays the shipment price if the shipment price estimation is below the threshold. However, if the shipment price exceeds that threshold, it has to be given by the carrier. Shah et al. in [35] proposed a novel price-setting model. This model determines the price of transport for a shipment and the inventory holding costs, although a varied number of deterministic variables are considered to set shipment prices in the above studies. However, in real life, shipments are not operated under ideal scenarios. There can be various carriers, forwarder ports, and shipment lines that may or may not become part of any cargo transportation. Hence other approaches are also explored for price setting in the shipping industry.

2.7.2 Algorithmic Pricing:

Keeping in view the limitations of scenario-based studies, researchers have also studied algorithms for pricing. In this group of research studies, game theory has been used along with Nash equilibrium to determine the price for shipment containers keeping demand or capacity as deterministic factors. Lee et al. [36] adopted a game theory approach. The authors modeled three players who compete with each other for pricing and routing choices. The selection criterion used in this method is the Nash equilibrium. Wang et al. conducted groundbreaking research in the pricing domain in 2014. The authors used the game theory as a research setting and used freight rate, service frequency, and capacity as critical price-setting criteria[37]. In 2016, a pricing method based on game theory was proposed using Nash equilibrium and ship capacity[38]. However, the research above takes shipment demand as a deterministic factor. In the research studies presented in [36],[38],[37], varied pricing decisions are proposed using shipment demand as a non-deterministic factor. In these studies, the authors do not use competition between shipping lines as a deterministic variable.

2.7.3 Comparative Analysis of existing spot pricing techniques:

Compounded by the literature review in regards to optimal spot price calculation for the shipment containers, it is apparent that there exist no studies that can help in deciding container shipment spot pricing based on shipment demand and available shipping capacity. Table 2.3 below summarizes the literature review we have done.

Thus, in this thesis we aim to model the shipment spot pricing based on container shipment demand and available shipping capacity to fill out this research gap.

Table 2-3 Summary of spot pricing techniques designed so far in the Container Shipping Industry

Models	Application in shipping industry	Domain	Supporting Models
Simple exponential smoothing (SAS)	✕		
Holt Winter's non-seasonal method	✕		
Holt Winter's seasonal method	✕		
Auto Regressive (AR)	✕		
Moving Average (MA)	✕		
Auto Regressive Moving Average (ARIMA)	✓	Price prediction	ANNs
Seasonal Auto Regressive Moving Average (SARIMA)	✕		
Fb PROPHET	✕		
Artificial Neural Networks (ANNs)	✓	Price Prediction	ARIMA
Support Vector Regressor (SVR)	✓	Price Prediction	ARIMA
Support Vector Machines (SVM)	✓	Freight index.	Nil
Long Short Term Neural Networks (LSTM)	✕		
Recurrent Neural Networks (RNN)	✕		

2.8 Machine Learning-based prediction Models:

The implementation of forecasting techniques ranges from relatively simple to complex methods. A considerable amount of work has been done to perform forecasting based on time series [17]. The selection of a forecasting model is solely dependent on the dataset available, the forecasting context, the time horizon of the forecast, the degree of desirable accuracy and the time availability to make an analysis. These elements must be weighed to select the technique that can best benefit the businesses. Forecasting techniques are of various types, which include regression, classification, statistical, and time series models as seen below in figure 2.9.

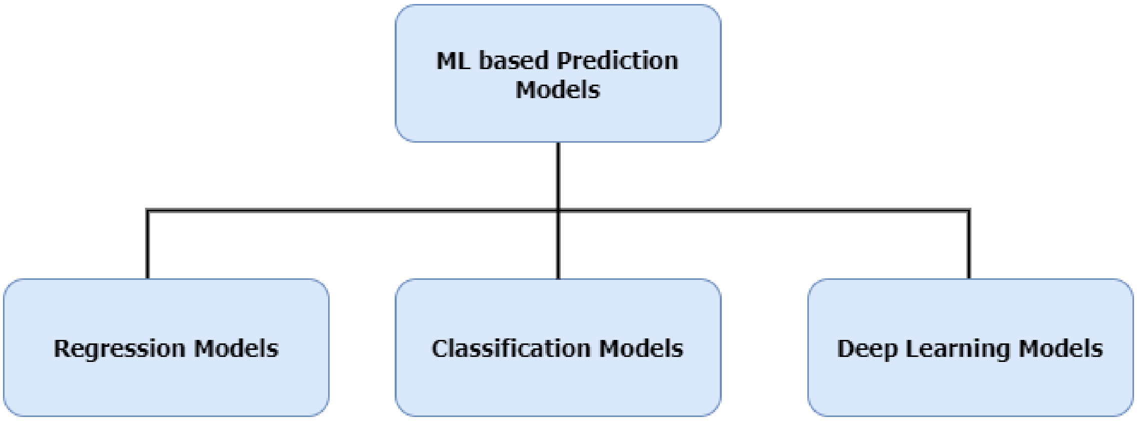


Fig. 2-9 ML base Prediction Models

2.8.1 Regression-based models:

Regression-based models are used to assess the relationship between the dependent and independent variables. These models can be linear or nonlinear. Examples of regression-based models include linear regression, support vector regression, kNN regression, random forest regression, and gradient boosted regression. Linear regression is an ML algorithm that comes under the umbrella of supervised models. It is used to understand the relationship between variables and the forecasting of target variables for forecasting. Support vector regression uses the same principle as that of SVM, but with a minor difference. It tends to minimize the error by maximizing the hyperplane while tolerating the part of the error [39]. K-nearest neighbor regression uses feature similarity from a training dataset to predict a new data point's values. The new data point is assigned a value based on how closely it resembles the training dataset [40].

On the other hand, random forest regression is an ensemble technique and uses begging as an underline method. Rather than relying on individual trees, it uses multiple decision trees to determine the final output. Gradient boosted regression is a tree-based ensemble method, but unlike begging, predictions are made sequentially (Rodriguez-Galiano et al. 2015).

2.8.2 Classification-based models:

This class of models aims to classify the data and predict new instances based on previous groupings. The methods in this group include naïve Bayes, neural networks (NNs), recurrent neural networks, k-nearest neighbors, k-means, clustering, and ensemble methods. New techniques for forecasting include deep learning methods. Deep learning is an advanced form of supervised learning, and it has its roots in NNs. These NNs have multiple layers and are capable of handling multivariate and large datasets [41]. Below is a summary of the research done in the recent past on making predictions. House price prediction is undertaken using different algorithms such as C4.5, RIPPER, Naïve Bayesian, and AdaBoost [42]. According to [42], RIPPER outperforms the other algorithms in comparison. In [43], SVM is used to predict water levels. NNs have been in the limelight for the past decade for making valuable predictions. Crude oil pricing is predicted using NNs with multivariate datasets. [43] Showed that NNs outperform regression models as regression-based models are prone to overfitting.

2.9 Comparative Analysis of existing feature selection techniques:

From the studies conducted in section 2.7, it is evident that there exist no price prediction models that can predict shipment prices (both spot and contract) based on available shipping capacity also known as supply and current shipment demand. Summary of the literature review over this research issue is presented in table 2.4.

To fill this research gap, in this research we aim to implement a price prediction model that consider shipment demand and available capacity for the price calculation.

Table 2-4 Summary of regression- based machine learning model's application in the Container Shipping Industry

Models	Application in shipping industry
Linear regression	✕
Support Vector Machines	✕
kNN regressor	✕
Gradient boosted regression	✕
Random Forest regression	✕
Support Vector Regressor (SVR)	✕
Naïve Bayes	✕
Neural Networks (NNs)	✕
Recurrent Neural Networks (RNNs)	✕
K-means clustering	✕
Ensemble Methods	✕
Support Vector Machines (SVM)	✕
Long Short Term Neural Networks (LSTM)	✕
RIPPER	✕
C4.5	✕
AdaBoost	✕

2.10 Overall identified research gaps:

Based on the summaries presented in Table 2.1-2.4, following are the few research gaps that are found from the literature and are the focus of this thesis:

I. There is no existing study on the relationship between demand, supply and pricing which highlights the factors affecting pricing decisions.

II. There exist no feature selection methods that can use input from the domain expert to speed up the feature selection method from varied and vast real-time datasets.

III. There is no existing mathematical model based on demand, supply and price to calculate the optimal price (opportunity cost) of shipments.

IV. There is no existing demand forecasting model that can predict future demand for both the short- and long-term which incorporates seasonality and holiday effects.

V. There is no existing price prediction machine learning-based model that incorporates the relationship between demand and supply to predict the price of a shipment.

2.11 Summary:

In this chapter, we have presented an extensive literature review over the research directions we identified in chapter 1. From the domain study it can be seen that the shipping industry is financially in a downward state, according to a study published in 2017, because it does not have industry-wide visibility of current demand and supply, resulting in uninformed pricing decisions. However, there is a shift in the application of AI-based solutions to solve significant problems. The literature review in this chapter has provided us with the information that there is a need to source real time data in order to design data driven methods for digitizing the shipping industry. In the next chapter, we explain the data sourcing from the Australian shipping industry's stakeholders and it's preprocessing for designing the data driven methods for Australian shipping industry.

2.12 Reference:

1. Wu, S., et al., Survey on Prediction Algorithms in Smart Homes. IEEE Internet of Things Journal, 2017. 4(3): p. 636-644.

2. Dey, A., Machine Learning Algorithms: A Review International Journal of Computer Science and Information Technologies, 2016. Vol. 7 (3)(1174-1179).

3. Sullivan, W.G., A review of: “APPLIED FORECASTING METHODS” by N. T. Thomopoulos, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, iiv + 369 pages, list $24.50. The Engineering Economist, 1979. 25(4): p. 327-327.

4. Daim, T. and A. Hernandez, A framework for managing the forecasting process. Vol. 12. 2008. 597-627.

5. L. Schultz, R., Fundamental aspects of forecasting in organizations. Vol. 7. 1992. 409-411.

6. SHIM, J.K., & SIEGEL, J. G. , Handbook of financial analysis, forecasting & modeling. 1988, Englewood Cliffs, N.J., .

7. Brockwell, P.J.a.D., R.A Introduction to Time Series and F*orecasting*. 2nd Edition ed. Texts in Statistics. 2010, New York: Springer

8. Cortinhas, C.B., Ken (Kenneth Urban), *Statistics for business and economics,* . 2012,: European ed, Wiley, Chicheste.

9. Jarreter, J., *Business Forecasting Methods*. reprint ed., UK: Blackwell (1987).

10. Taylor, J.W., Exponentially weighted information criteria for selecting among forecasting models. International Journal of Forecasting, 2008. 24(3): p. 513-524.

11. Ostertagova, E. and O. Ostertag, Forecasting Using Simple Exponential Smoothing Method. Acta Electrotechnica et Informatica, 2012. 12: p. 62–66.

12. Chatfield, C., Time Series Forecasting. illustrated ed. 2000: CRC Press.

13. Stellwagen, E. and L. Tashman, ARIMA: The Models of Box and Jenkins. Foresight: Int. J. Appl. Forecast., 2013: p. 28-33.

14. Akman Biyik, C. and M. Tanyeri, Pricing Decisions in Liner Shipping Industry: A Study on Artificial Neural Networks. 2018.

15. Lago, J., F. De Ridder, and B. De Schutter, Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. Applied Energy, 2018. 221: p. 386-405.

16. McNally, S., J. Roche, and S. Caton. Predicting the Price of Bitcoin Using Machine Learning. in 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP). 2018.

17.Ebrahimian, H., et al., *The price prediction for the energy market based on a new method.* Economic Research-Ekonomska Istraživanja, 2018. **31**(1): p. 313-337.

18.Voyant, C., et al., Machine learning methods for solar radiation forecasting: A review. Renewable Energy, 2017. **105**: p. 569-582.

19.Chandrashekar, G. and F. Sahin, *A survey on feature selection methods.* Comput. Electr. Eng., 2014. **40**(1): p. 16-28.

20.Blum, A.L. and P. Langley, *Selection of relevant features and examples in machine learning.* Artificial Intelligence, 1997. **97**(1): p. 245-271.

21.H C Law, M., M. Figueiredo, and A. K Jain, *Simultaneous feature selection and clustering using mixture models.* IEEE transactions on pattern analysis and machine intelligence, 2004. **26**: p. 1154-66.

22.Vergara, J. and P. Estevez, *A Review of Feature Selection Methods Based on Mutual Information.* Neural Computing and Applications, 2014. **24**.

23.Rodriguez-Galiano, V.F., et al., Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. Science of The Total Environment, 2018. **624**: p. 661-672.

24.Vieira, S.M., et al., Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. Applied Soft Computing, 2013. **13**(8): p. 3494-3504.

25.Reunanen, J., Overfitting in making comparisons between variable selection methods. J. Mach. Learn. Res., 2003. 3: p. 1371-1382.

26.Goldberg, D.E., Genetic Algorithms in Search, Optimization and Machine Learning. 1989: Addison-Wesley Longman Publishing Co., Inc. 372.

27.Xu, L., et al., Pricing and balancing of the sea–cargo service chain with empty equipment repositioning. Computers & Operations Research, 2015. **54**: p. 286-294.

28.Inbarani, H.H., A.T. Azar, and G. Jothi, *Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis.* Computer Methods and Programs in Biomedicine, 2014. **113**(1): p. 175-185.

29.Mitra, P., C.A. Murthy, and S.K. Pal, *Unsupervised feature selection using feature similarity.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002. **24**(3): p. 301-312.

30.P. Xing, E. and R. Karp, CLIFF: Clustering of High-Dimensional Microarray Data Via Interative Feature Filtering Using Normalized Cuts. Bioinformatics, 2001. **17**.

31.Pal, S.K., R.K. De, and J. Basak, *Unsupervised feature evaluation: a neuro-fuzzy approach.* IEEE Transactions on Neural Networks, 2000. **11**(2): p. 366-376.

32 Pudil, P., et al., Feature selection based on the approximation of class densities by finite mixtures of special type. Pattern Recognition, 1995. **28**(9): p. 1389-1398.

33.Zhao, Z. and H. Lu, *Semi-supervised feature selection via spectral analysis.* Proceedings of the 7th SIAM International Conference on Data Mining, 2007: p. 641-646.

34.Chen, R., J.-X. Dong, and C.-Y. Lee, *Pricing and competition in a shipping market with waste shipments and empty container repositioning.* Transportation Research Part B: Methodological, 2016. **85**: p. 32-55.

35.Shah, N. and J.K. Brueckner, *Price and frequency competition in freight transportation.* Transportation Research Part A: Policy and Practice, 2012. **46**(6): p. 938-953.

36.Lee, H., et al., Modeling the Oligopolistic and Competitive Behavior of Carriers in Maritime Freight Transportation Networks. Procedia - Social and Behavioral Sciences, 2012. **54**: p. 1080-1094.

37.Liu, D. and H. Yang, *Joint slot allocation and dynamic pricing of container sea–rail multimodal transportation.* Journal of Traffic and Transportation Engineering (English Edition), 2015. **2**(3): p. 198-208.

38.Yin, M. and K.H. Kim, *Quantity discount pricing for container transportation services by shipping lines.* Computers & Industrial Engineering, 2012. **63**(1): p. 313-322.

39.Pereira, F.C. and S.S. Borysov, *Chapter 2 - Machine Learning Fundamentals*, in *Mobility Patterns, Big Data and Transport Analytics*, C. Antoniou, L. Dimitriou, and F. Pereira, Editors. 2019, Elsevier. p. 9-29.

40.Zhang, S., et al., *A novel kNN algorithm with data-driven k parameter computation.* Pattern Recognition Letters, 2018. **109**: p. 44-54.

41.Tealab, A., Time series forecasting using artificial neural networks methodologies: A systematic review. Future Computing and Informatics Journal, 2018. **3**(2): p. 334-340.

42.Park, B. and J.K. Bae, Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 2015. **42**(6): p. 2928-2934.

43.Kisi, O., et al., A survey of water level fluctuation predicting in Urmia Lake using support vector machine with firefly algorithm. Applied Mathematics and Computation, 2015. **270**: p. 731-743.

3 Problem Definition and Overview

3.1 Introduction:

In the previous chapter, we highlighted the problems the shipping industry is facing and the research to be conducted to solve the explained problems. In this chapter, we have outlined problem definition and have provided its overview.

In section 3.1, the key terminologies that we use in this thesis are outlined to help readers understand the research problems that, we intend to solve. In section 3.2, problem overview and problem definition is explained. In section 3.4, research issues are discussed followed by the definition of the research scope in section 3.5. The research approach to the problem in discussed in section 3.6. Section 3.7 summarizes this chapter.

3.2 Key Concepts:

In this section we explain the key terminologies and their context in which they are used in this thesis.

3.2.1 Spot pricing:

To deal with volatility in shipment pricing, there are two types of pricing strategies currently employed in the shipping sector, contract market pricing and spot pricing. The spot market has a fluctuating price, where one can benefit from lower than contract rate prices in the low season, but face escalating prices and less certainty of being able to secure a booking on a particular vessel in peak season, as space is reserved for contracted customers [1] [2]. Current industry pricing practices and commercial product offerings deliver sub-optimal outcomes for the shipping lines and also for their customers.

3.2.2 Contract pricing:

Contract pricing is the term used to describe the price quotation provided to the customers for booking cargo ahead of time without information of future demand and capacity [1].

3.2.3 Cargo demand:

The requirement by customers to transport cargo containers via sea from one location to another [1].

3.2.4 Time series forecasting:

Forecasting is the process of making predictions of the future events based on past and present data. The term forecasting is specifically used for predicting future events based on similar time of the previous and present year. Time series forecasting is the term used to model predict future values based on previously observed values [3].

3.2.5 Prediction:

Prediction can be termed as prophecy i.e., a statement about a future event. Often forecasting and prediction are used interchangeably. However, prediction can be without time dependency based on data from past year [4].

3.2.6 Multivariate prediction model:

Prediction of any target variable based on more than a single variable.

3.2.7 Relationship:

Relationship can be termed as interdependencies between two entities. In this thesis, relationship refers to the dependencies of variables among each other.

3.2.8 EDA:

EDA refers to exploratory data analysis. It is a statistical term that refers to analyzing the data sets to summarize their main characteristics.

3.2.9 Correlation:

Correlation is a statistical term used to demonstrate and quantify the statistical relationship between two variables. It refers to the degree of association between the variables.

3.2.10 Model:

Model refers to the proposed structure typically in a smaller scale than original. In other words, it refers to the example that imitates the concept. In this thesis, we use model in a dual context, (1) Mathematical model; (2) Machine learning model [5].

3.2.11 Machine learning model:

Machine learning models are output by algorithms and are composed of model data and a prediction algorithm [6, 7].

3.2.12 Mathematical model:

It describes a system by a set of variable and a set of equations that establish the relationship between the variables.

3.3 Problem Overview and Problem Definition:

In this research, we want to introduce digitization into the Australian Container Shipping Industry with machine learning capabilities to provide a market-wide view of current and future demand (both for short-term and long-term), prices prediction model that can predict spot pricing based on current demand and available capacity (supply). We also intend to design a model to calculate optimal spot pricing for container shipment based on current cargo demand for increased revenue.

The first step is to explore the relationship between demand, supply, and price. This can provide the current market scenario in which the industry is operating. To calculate valid spot pricing, we have done the statistical analysis to discover the relationship between demand, capacity, and price. Based on inferred results, historical data, and stakeholder's limitations for a price quotation, a novel mathematical model is designed that can calculate optimal pricing based on the mentioned factors. In the second step, we have designed a short-term demand prediction model to predict short-term cargo demand i.e., six weeks. Also, we have designed an algorithm for long-term demand forecasting i.e., six months ahead. The cargo demand models designed (both for short-term and long-term) can incorporate volatility in the shipping industry i.e., seasonality in historical data and previous and upcoming holidays effects in their forecast. Finally, for making contract pricing more active, the demand forecast performed earlier is used with the available capacity to predict optimal container pricing based on demand and capacity-using a regression-based multivariate machine learning model. This predicted price will bridge the gap between contract and spot pricing in the container shipping industry. Hence, this research provides visibility into future demand and will allow shipping lines and their customers to make better-informed pricing and procurement decisions.

This research aims to provide shipping industry participants with live trade lane level visibility of short-run cargo demand and the supply of shipping capacity and empty equipment availability to service this demand. It will allow shipping lines and their customers to make better and more informed pricing and procurement decisions. It will also help avoid situations that regularly arise where demand outstrips supply, yet freight rates fall. By showing the actual shipping capacity in the trade for the next six weeks and forecast the cargo demand for it, we overlaid it with 'spot price' data to get a composite picture to make reliable data-driven decisions. This research will bring a price prediction capability to the Australian shipping industry by using machine learning models. The machine learning models will also account for price fluctuations due to seasonal variations and holidays. As a result of these developed methods, businesses will be able to grow their revenue by setting prices using a global view of the Australian shipping market using demand, supply, and other aiding factors. The primary research question (RQ) that will be addressed in this research is as follows:

How can we intelligently model the relationship between demand and supply so that business can reliably predict the future demand and container shipment prices?

3.4 Research Issues:

In this section, we aim to further discuss the major research issue identified in the above section. In order to solve the aforementioned research issue, it is further broken down into the following sub-research questions:

I. How to model the relationship between demand, supply, and price in the Australian shipping industry? And use this relationship to determine optimal shipment pricing.
II. How to design a feature selection method that can incorporate input from domain expert into its feature selection process?

III. How to forecast short-term container demand using historical demand data incorporating holiday and seasonality effects in the Australian shipping industry?

IV. How to predict long-term container demand in the Australian shipping industry, which incorporates holiday and seasonality effects?

V. How to build a reliable price prediction model given actual demand and supply in the shipping industry?

VI. How to verify and validate the research above questions RQ1 to RQ4?

3.5 Research Scope:

Based on research issues identified in section 3.4, the scope of the thesis is defined. The thesis scope is as follows:

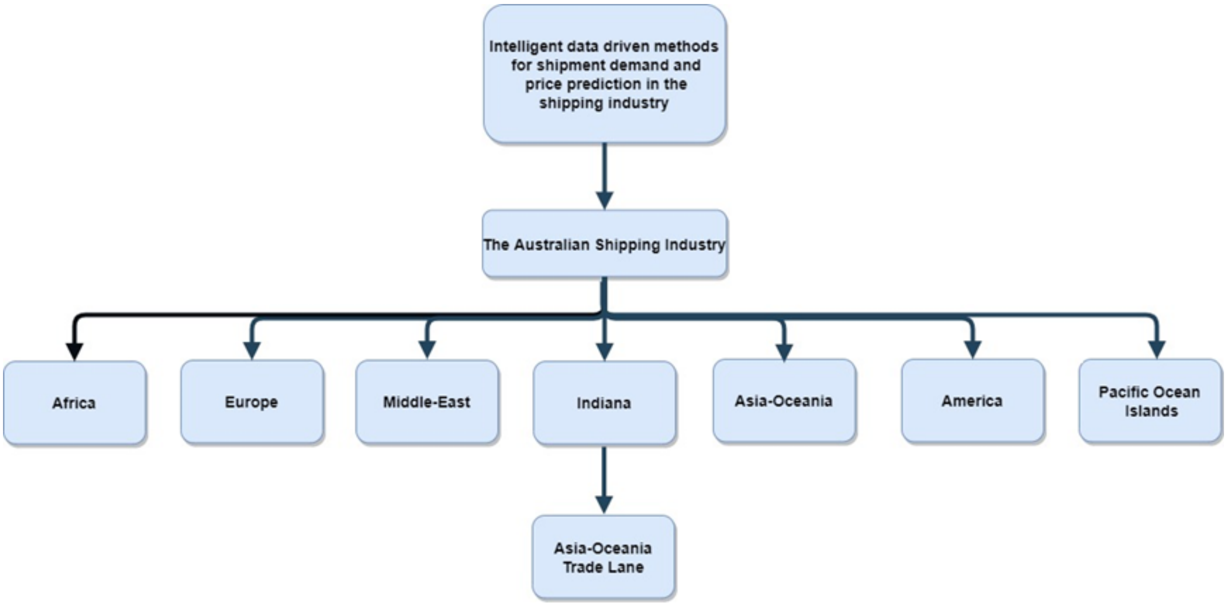

Intelligent data driven methods for shipment demand and price prediction in the shipping industry
The Australian Shipping Industry
Africa
Europe
Middle-East
Indiana
Asia-Oceania
America
Pacific Ocean Islands
Asia-Oceania Trade Lane

Fig. 3.1 Model relationship between container shipment demand, capacity, and price.

3.5.1 Determine relationship between shipment demand, available shipping capacity and shipment pricing to determine optimal shipment spot pricing:

The first research issue is to determine what relationship exist between the container shipment demand, available capacity and the prices quoted for the container shipments based on the current demand and capacity.

As explained in the previous chapters, currently shipping industry has no digital view to determine what the market statistics are. The shipping industry stakeholders are quoting prices independently of shipment demand and capacity. Hence, it is losing much revenue [2]. In order to understand the current relationship between the said variables it is vital to understand market operations and to determine where the shortfall lies.

3.5.2 Forecasting short-term container demand using historical demand data incorporating holiday and seasonality effects in the Australian shipping industry:

The shipping industry has no visibility into the container shipment demand. This lack of knowledge effects their price quotation capabilities in the short-term. There exists no method that can help them visualize container shipment demand ahead, even for short-term even. There is a concerning need to develop such capability that can help them quote prices such that it can earn revenue for them [8]. Solving this research issue, will enable the shipping industry to look into the near future demand and will help them increase their business and will create an impact in the developing shipping industry [9]

3.5.3 Forecasting long-term container demand using historical demand data incorporating holiday and seasonality effects in the Australian shipping industry:

Although the research issue explained in 3.5.2, has its impact on pricing decisions, the importance of long-term demand forecasting can also not be denied. To help the shipping industry maintain consistent supply chain availability as well as make informed economic and infrastructure decisions, long term container demand is of utmost importance. Thus, to solve this vital research issue, we also design a ML model that can predict future demand for 52 weeks in order to provide future demand requirements one year ahead.

3.5.4 Developing ML based price prediction model:

Pricing independent of container shipment demand and capacity is the primary cause of industry instability and hence tends to be one of the most important research issues identified in the shipping industry. There exists no such model that can predict shipment price based on current shipment demand and capacity. Therefore, we aim to develop a machine learning based model that can provide shipment price prediction capability to the shipping industry.

This model will not only aid in predicting optimal price for the spot market but will also aid in quoting the optimal contract pricing to the shipping industry stakeholders.

3.6 Research approach to problem solving:

Research methodology is the “collection of problem solving methods governed by a set of principles and a common philosophy for solving targeted problems”[10]. A number of research methodologies have been proposed and applied in the information systems domain such as case studies, field studies, design research, field experiments, laboratory experiments, surveys, and action research [10].

3.6.1 Choice of Research Method:

In this research, we consider design science research (DSR) as the most appropriate research methodology to achieve our research objectives [11]. The reason for this is that this research aims to produce a new series of algorithms in the form of a software decision-making tool. This software decision-making tool will help various stakeholders in the shipping industry. This is consistent with the philosophy of DSR which aims to develop a new product or artefact as output. The methodology, as illustrated in figure 3.2, comprises five stages [10].

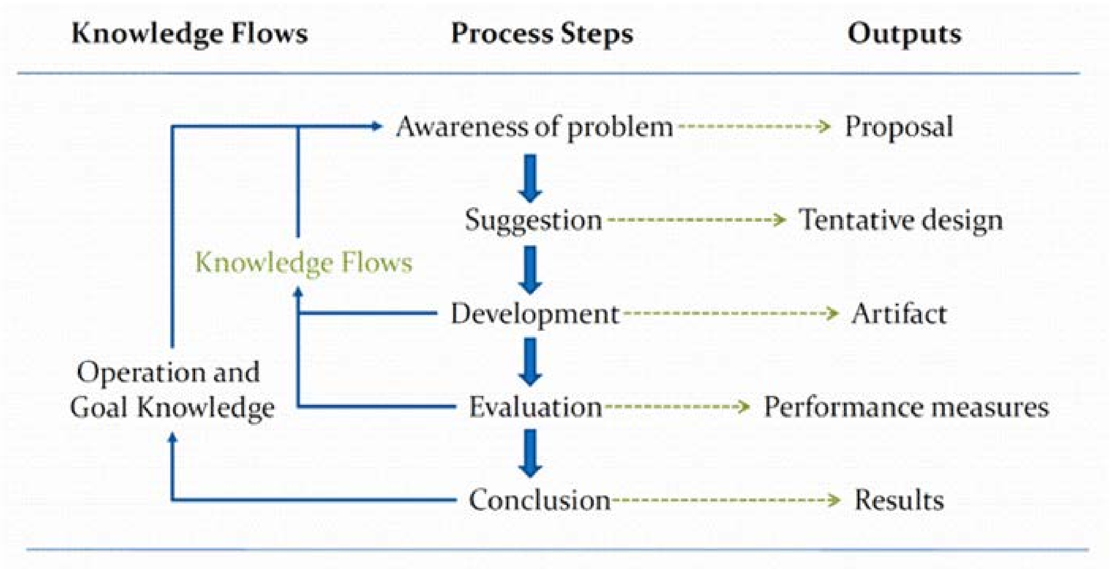


Fig. 3.2 General Methodology for Design Research

3.6.1.1 Awareness of problem:

This is the first step where the limitations of the existing literature and relevant software applications are analyzed, and significant research problems are acknowledged. The research problems reflect a gap between existing applications and the expected status. Research problems can be identified from different sources: industry experience, observations on practical applications and a literature review. A clear definition of the research problem provides a focus for the research throughout the development process. The output of this phase is a research proposal for a new research effort.

We began with identifying the research topic in order to gain awareness of the problem. The choice of a research topic can arise from personal interest, from the observation, or from the literature describing previous theory and research in the area, from social concern or as an outcome of some currently popular issues. The topic of this research is chosen from a combination of personal research and a business concern from the Australian Shipping Industry. The research will influence Australia's adoption of digital freight solutions. Aiming to bring transparency to the containerized cargo demand and shipping capacity supply, the research will allow shipping lines and their customers to make better-informed pricing decisions.

3.6.1.2 Suggestion:

This phase immediately follows the identification of the research problems where a tentative design is suggested. The tentative design describes the prospective artefacts and how they can be developed. The suggestion phase is a creative process during which new concepts, models and the functions of artefacts are demonstrated. The resulting tentative design achieved in this step is usually one part of the research proposal. Thus, the output of the suggestion step is also feedback on Step I, whereby the research proposal can be revised.

In the second step, we suggested some solutions to the identified problems. To do so, we exploited various ML algorithms.

3.6.1.3 Development:

This phase considers the implementation of the suggested tentative design of the artefacts. The techniques for implementation will be based on the artefact to be constructed. The implementation itself can be simple and need not involve novelty; the novelty is primarily in the design not the construction of the artefact. The development process is often an iterative process in which an initial prototype is first built and then it evolves as the researcher gains a deeper comprehension of the research problems.

Subsequently, we applied our concept using the selected algorithm for development of the solution to the problem. Following the development, a range of evaluation measure are studied. Amongst many, the following matrices are selected for evaluation of identified research problems.

3.6.1.4 Evaluation:

This phase considers the evaluation of the implemented artefacts. The performance of the artefacts is evaluated according to the criteria defined in the research proposal and the suggested design. The evaluation results may or may not meet the expectations and are fed back to the first two steps. Accordingly, the proposal and design might be revised, and the artefacts might be improved.

Root mean squared error (RMSE) and mean absolute percentage error (MAPE) are used to evaluate the performance of the time-series models[12]. RMSE can be computed as follows:

Equation 3.1

$$\mathrm{RMSE} = \sqrt{\overline{(\mathrm{f}-\mathrm{o})^2}} \qquad (1)$$

Where *f* is the forecast and *o* is the observed value. MAPE measures forecast accuracy as a percentage. MAPE can be calculated as follows:

$$MAPE = \frac{1}{N}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \qquad (2)$$

Where A_t is the actual value and F_t is the forecast.

3.6.1.5 Evaluation of Research Issues 1:

I. No of features extracted per iteration.
II. Regression Analysis [13].
III. Density plots [14, 15].
IV. Correlation analysis [16, 17].
V. Current pricing versus prices (suggested by proposed model) visualization.

3.6.1.6 Evaluation of Research Issues 2 and 3:

For the evaluation of the research issues 2 and 3 we have used

I. RMSE
II. MAPE

3.6.2 Evaluation of Research Issues 4:

For evaluation of research issue 4, we have used the following metrics:

I. RMSE
II. R^2 Score. [18, 19]
III. Accuracy [20]

3.6.2.1 Conclusion:

This is the final phase of a design research effort. Typically, it is the result of satisfaction with the evaluation results of the developed artefacts though there still may be deviations in the behavior between the suggested proposal and the artefacts that are actually developed. However, the design research effort concludes as long as the developed artefacts are considered 'good enough' wherein any anomalous behavior may well serve as the subject of further research.

3.7 Summary:

In this chapter, we presented a formal definition of the problem we intend to address in this thesis. The identified problem was subsequently decomposed as a set of four key cohesive issues, which need to be solved in order to address the problem being addressed in this thesis. Each of the identified four research issues were explained in depth in relation to the existing literature and were defined formally.

In the next chapter, we provide an overview of the solution for the problem being addressed in this thesis. Additionally, we present an overview of the solution for each of the four research issues that encompass the problem being addressed in this thesis.

3.8 References:

1. Akman Biyik, C. and M. Tanyeri, Pricing Decisions in Liner Shipping Industry: A Study on Artificial Neural Networks. 2018.

2. Ubaid, A., F. Hussain, and J. Charles, Modeling Shipment Spot Pricing in the Australian Container Shipping Industry: Case of ASIA-OCEANIA trade lane. Knowledge-Based Systems, 2020. **210**: p. 106483.

3. Brockwell, P.J.a.D., R.A *Introduction to Time Series and Forecasting*. 2nd Edition ed. Texts in Statistics. 2010, New York: Springer

4. Chatfield, C., *Time Series Forecasting*. illustrated ed. 2000: CRC Press.

5. Taylor SJ, L.B.P.P.e.v.h., *Forecasting at scale*. 2017, Facebook, Menlo Park, California, United States.

6. Dey, A., *Machine Learning Algorithms: A Review* International Journal of Computer Science and Information Technologies, 2016. **Vol. 7 (3)**(1174-1179).

7. Burkov, A., The Hundred-Page Machine Learning Book. 2019: Andriy Burkov.

8. Pricing for Profit in Container Shipping. April 2016.

9. Using technology to tame freight rate volatility and reduce capacity risks (white paper). 2019: Drewry Supply Chain Advisors.

10.Niu, L., J. Lu, and G. Zhang, Cognition-Driven Decision Support for Business Intelligence: Models, Techniques, Systems and Applications. 2009: Springer.

11.Abutabenjeh, S. and R. Jaradat, Clarification of research design, research methods, and research methodology:A guide for public administration researchers and practitioners. Teaching Public Administration, 2018. **36**(3): p. 237-258.

12. Brassington, G. Mean absolute error and root mean square error: which is the better metric for assessing model performance? in EGU General Assembly Conference Abstracts. 2017.

13. Brook, R.J. and G.C. Arnold, *Applied regression analysis and experimental design*. 2018: CRC Press.

14.Spencer, C., C. Yakymchuk, and M. Ghaznavi, Visualising data distributions with kernel density estimation and reduced chi-squared statistic. Geoscience Frontiers, 2017. **8**(6): p. 1247-1252.

15. Geenen, J.W., et al., Increasing the information provided by probabilistic sensitivity analysis: The relative density plot. Cost Effectiveness and Resource Allocation, 2020. **18**(1): p. 1-10.

16. Makowski, D., et al., *Methods and algorithms for correlation analysis in R.* Journal of Open Source Software, 2020. **5**(51): p. 2306.

17. Yang, X., et al., *A survey on canonical correlation analysis.* IEEE Transactions on Knowledge and Data Engineering, 2019.

18. Mooi, E. and M. Sarstedt, A concise guide to market research: The process, data, and methods using IBM SPSS Statistics. New York: Springer. 2011. 307.

19. Moksony, F. and R. Heged, *Small is beautiful. The use and interpretation of R2 in social research.* Szociológiai Szemle, Special issue, 1990: p. 130-138.

20. *Accuracy_Score*. [cited 2020 24 December]; Accuracy measures]. Available from:https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html.

21. Bairagi, V. and M.V. Munot, Research methodology: A practical and scientific approach. 2019: CRC Press.

4 Solution Overview

4.1 Introduction:

As explained in chapter 2, several research works have made advances to solve real life problems using machine learning algorithms. However, as is evident from the discussions in chapter 2 and 3, the issue of how we can intelligently model the relationship between demand, supply and prices so that the shipping industry's stakeholders can reliably predict the future shipment prices remains unresolved. Therefore, in chapter 3 we highlighted four primary research issues that are aimed at solving this pivotal problem.

In this chapter, we present an overview of the solution to each of the four research issues starting from Section 4.2 up to section 4.7. Finally, section 4.8 concludes this chapter.

4.2 Overview of the Solution for implementing intelligent data driven methods for demand and price prediction in the shipping industry:

As pointed out in the last chapters, not much work has been done to incorporate intelligent data driven methods into the shipping industry's processes. In this section, we present an overview of the overall solution for incorporating various data driven methods into the shipping industry's processes so as to provide market-wide visibility into the industry statistic. Figure 4.1 below depicts the research questions originated from the primary research question to explain the solution of research issues identified so far.

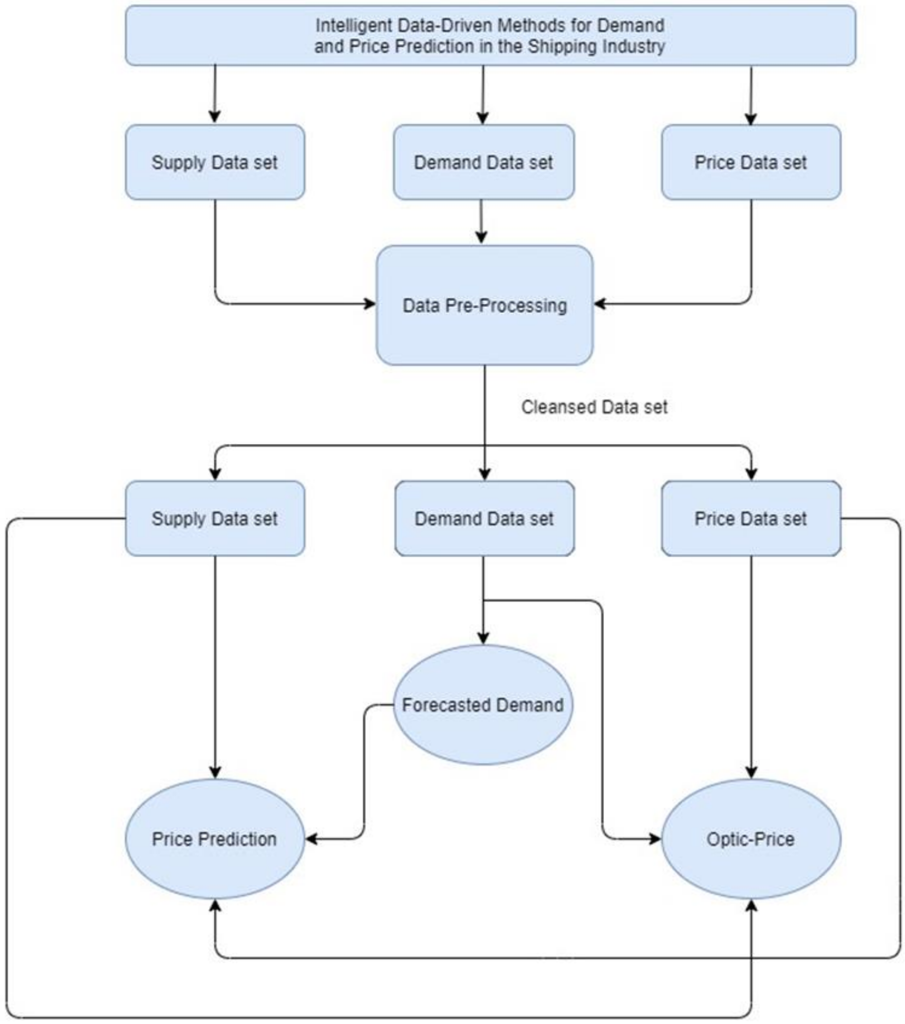


Fig. 4-1 Solution Overview of Intelligent Data Driven Models for shipment demand and price prediction in the Australian Shipping Industry.

4.3 Proposed solutions:

In this step, we outline and discuss the proposed approach for each of the five sub-questions.

4.4 *Solution Overview for RQ 1:* How to model the relationship between demand, supply and price in the Australian container shipping industry?

In the previous chapters, it was pointed out that there exists no study that can explain the relationship between shipment demand, available shipping capacity and the respective shipment prices. Moreover, we also discussed and he need to understand this this vital relationship to gain a business edge. Hence, in this research issue, we aim to investigate the aforementioned pivotal relationship between the said variables. To do so, we have divided the research issue into the following parts.

I. Sourcing of dataset from various sources and no consolidated Dataset for shipping industry exists consisting of demand, supply and pricing. For this research objective, datasets from three different sources are used. Initially, demand data from five major Australian ports, Port Botany [55], Port of Melbourne [56], Port of Brisbane [57]; Flinders Port [58] and Fremantle Port [59] are gathered from their respective websites. Supply data is gathered from Sea-Intelligence maritime analysis [60]. Sea-Intelligence publishes weekly data at the end of each week to provide information on the capacity provided by carriers. Pricing data for imports is gathered from the Shanghai Containerized Freight Index [61].

II. The mentioned datasets have a number of variables along with pricing information. In order to select the key variables for price prediction in the Australian liner shipping industry, there is a need for a novel framework for feature selection. Hence, we aim to design a novel feature selection method that can incorporate input and customization into feature selection process by domain expert.

III. Following the feature selection comes data cleansing. The demand data statistics on the websites are presented as trade summaries. Region-based monthly proofread imports and exports are given for both empty and full containers in twenty-foot equivalent units (TEUs). From this data, we selected imports for both empty and full imports for the Asia Oceania region. For demand data, the first step is to separate the trade line demand for Asia Oceania. The second step is to add up all the empty and filled container demands for Asia-Oceania trade lines. This gives us the total demand for the Asia Oceania trade line. This demand dataset is a combination of both exports and imports. In the next step, import and export data are segregated using the port data ratios. The demand data is in a monthly view. The monthly data is then converted into weekly data.

IV. The supply data contains weekly capacity (TEUs) with respect to shipping companies which are offering services to and from Australia along the Asia Oceania trade line. Weekly data for the Asia-Oceania trade line is selected in the first step. The second step is the selection of the time frame for which the supply needs to be analyzed. In this research, we use data from the past three years i.e. , 2016-2019. In the third step, missing values are handled using the back-fill technique [62].

V. The pricing information gathered from SCFI has pricing records for all trade lines. The first step is the selection of trade line pricing data. The price data are also available in weekly format. Hence, the next step is imputing the missing values using the backfill technique. Data integration involves combining supply, demand and price datasets into a single cohesive dataset. The consolidated data contains supply, demand and pricing data on a weekly basis. Table 4.1 shows the variables used in the research.

Table 4-1 Features used in the research

Region	Name of attributes	Description
Asia Oceania Import Trade Line	Total Supply	Total Supply for Asia Oceania trade line
	Total Demand	Total demand for Asia Oceania trade line (sum of empty and full containers)
	Price	Prices quoted for shipping containers for import to Australia for the Asia Oceania trade line
	Date	Starting date of week

In order to study the relationship between the selected features, we aim to perform exploratory data analysis (EDA) followed by correlation analysis to quantify the relationship between the selected variables (i.e., demand, supply and price).

Once the relationship between the shipment demand, capacity and prices are determined, in the next step, we intend to model optimal shipment spot price based on historical data trends and values. **Solution Overview for RQ 2 and RQ 3:** In order to find a solution for RQ 2 AND RQ3, we select three algorithms which can incorporate seasonality. These are seasonal autoregressive integrated moving average (SARIMA), long short-term recurrent neural networks (LSTM) and Prophet.

4.5 Solution overview RQ 2 AND RQ3: How to forecast short-term demand using historical demand data which incorporates holiday and seasonality effects?

In this section, we present solution overview of the research issue 2 and 3 together. In order to forecast demand for both short and long-term, the following is steps are followed:

I. Segregation of demand data with respect to time already consolidated and cleansed for solution of research issue 1.

II. Time horizon selection for the forecast. We have selected 6 weeks for short-term and 52 weeks for long-term demand forecast. This horizon selection is done by taking into the consideration of industry needs and as specified by industry partners [Mizzen].

III. Custom defined holiday's selection. This step will be done in collaboration with the industry partner, as including holidays effecting industry operations are quite important for forecasts.

IV. In chapter 2, we have explained various ML models that can provide forecasting on temporal data. Hence, in this step, we select already existing time series models that can provide forecasts while incorporating seasonality and trends inherently present in iii above, data. For this purpose, we have short-listed SARIMA, PROPHET and Holt Winter's method for fitting into the shipment demand dataset.

4.6 Solution Overview for RQ 4: How to build reliable price prediction model given actual shipment demand and available shipping capacity?

To develop a price prediction model for a shipment based on demand and supply, following are the steps which we intend to follow:

I. Data collection from the research issue 1 in order to have demand, capacity and prices based on them.

II. Selection of appropriate model to predict price. For this purpose, we have selected regression-based models as they have the capability to materialize the hidden relationship of features into their prediction results. Five models are short listed to find the best fitting model amongst them. These include, (1) linear regression (LR); (2) support vector regression (SVR); (3) K-nearest neighbor regression (kNN-R); (4) random forest regression (RFR); and (5) gradient boosting regression (GBR). All of these models will be evaluated after the necessary parameter settings. After evaluation, the best fit model will be selected to perform the prediction.

4.7 Solution Overview of RQ5: Evaluation of RQ1, RQ2, RQ3 and RQ4:

Determination of metrics to evaluate the identified research issue is of vital importance. This evaluation lets us decide the best model to be implemented in the real world problem. Following are the metrics we aim to use for evaluation of our research issues:

I. In order to determine the relationship between shipment demand, available capacity and shipment price, regression and correlation analysis is used for evaluation. However, for evaluating a feature selection method's effectiveness, number of iteration vs features extracted is observed. For the later of research question 1, i.e., modelling of optimal price calculation based on shipment demand and capacity, again correlation analysis, regression analysis and data point plotting is presented.

II. To evaluate RQ2 and RQ3, we have used MAPE, and RMSE to measure the effectiveness of the models. Detailed explanation of these terms are presented in chapter 7.

III. For evaluation price prediction modes, we have selected accuracy, RMSE and R^2 score of models.

4.8 Summary:

In this chapter, we proposed a solution overview for each of the four cohesive research issues that we intend to implement in this thesis. In the next chapter we will explain the data sourcing and pre-processing process that we have performed over the shipping dataset to make it usable for our data driven models. The final cleansed dataset achieved by performing various operations in chapter 5 will be used in our research methods explained in chapter 6, 7 and 8 respectively.

5 Domain Driven Selective Wrapping

5.1 Introduction

The first chapter presents an introduction to the thesis followed by an extensive literature review in chapter 2. In chapter 3, we explained the problem definition and overview of this thesis followed by solution overview in chapter 4.

It is evident from the above chapters that data collection, feature selection and data pre-processing are the primary steps prior to carrying out data analytics. In this chapter we outline and explain in detail the data pre-processing involved in our work.

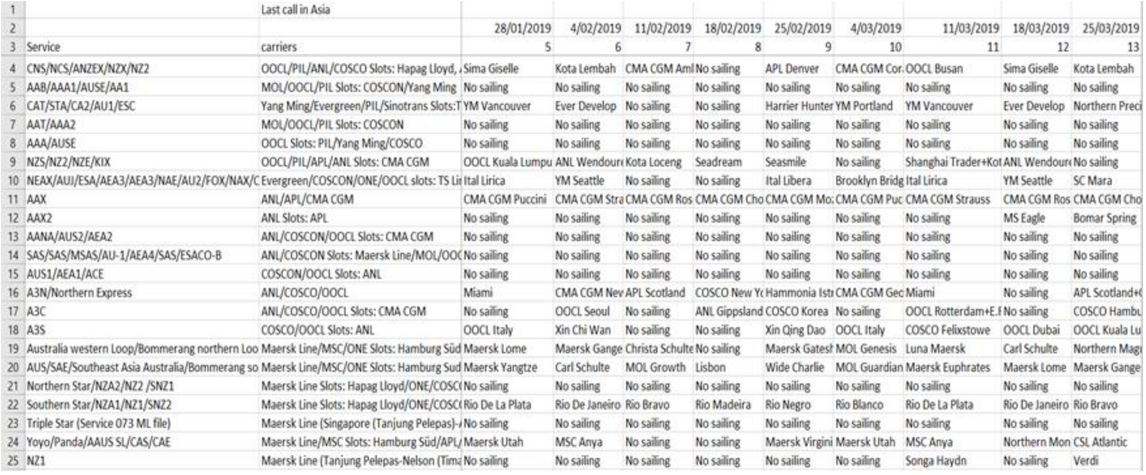
This chapter is organized as follows. In Section 5.2, we explain the dataset sourcing process. In section 5.3, we have explained the feature extraction process. Section 5.4 describes data pre-processing and cleansing.

5.2 Data Set Sourcing:

In this section, we explain the data sourcing process adopted to gather the dataset for this thesis. In the first step, we collected shipment demand data. To our knowledge, there exists no such dataset that can provide insights into the shipment demand, available shipment capacity and shipment pricing (both spot price and contract pricing) in the Australian Shipping industry. Hence, the methodological preparation of such data set is a critical aspect.

5.2.1 Capacity (Supply) Dataset Sourcing:

The available shipping capacity (supply) data set has been sourced from Sea-Intelligence. Sea-Intelligence is an analytical reporting company that publishes weekly available shipping capacity (supply) trade summaries [1, 2]. The weekly outlook is based on published schedules provided by ocean carriers. This data contains weekly total capacity (in TEUs) along with the vessel names. Fig 5.1 shows a snapshot of the dataset sourced from Sea-Intelligence for the Asia Oceania trade lane.

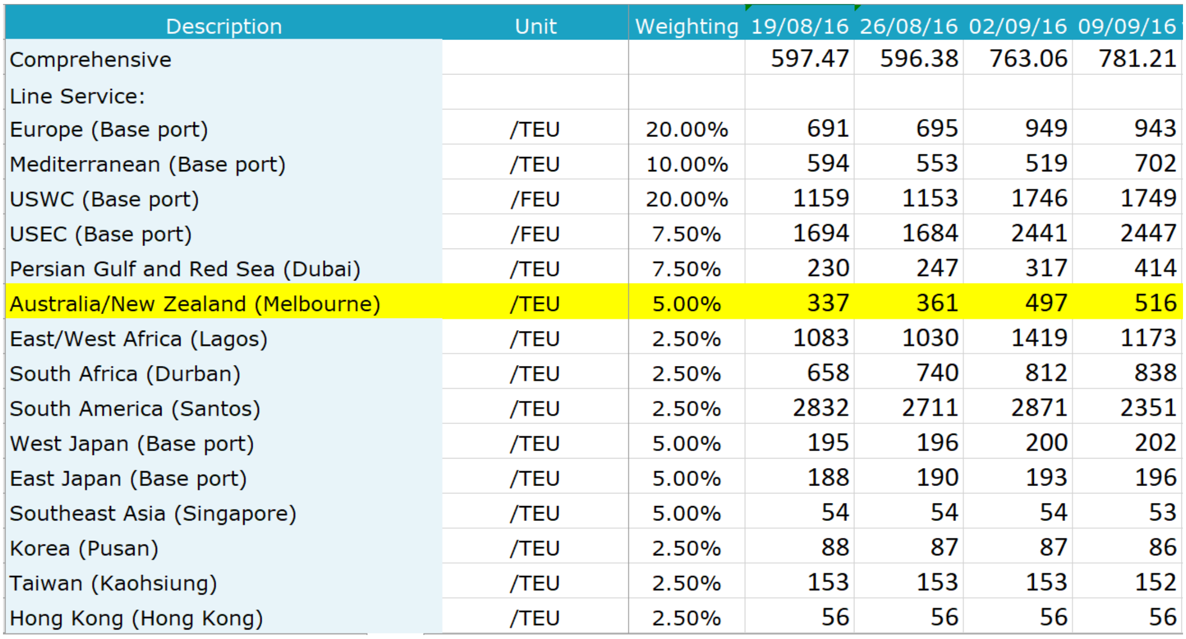


1		Last call in Asia									
2			28/01/2019	4/02/2019	11/02/2019	18/02/2019	25/02/2019	4/03/2019	11/03/2019	18/03/2019	25/03/2019
3	Service	carriers	5	6	7	8	9	10	11	12	13
4	CNS/NCS/ANZEX/NZX/NZ2	OOCL/PIL/ANL/COSCO Slots: Hapag Lloyd, /	Sima Giselle	Kota Lembah	CMA CGM Aml	No sailing	APL Denver	CMA CGM Cor	OOCL Busan	Sima Giselle	Kota Lembah
5	AAB/AAA1/AUSE/AA1	MOL/OOCL/PIL Slots: COSCON/Yang Ming	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing
6	CAT/STA/CA2/AU1/ESC	Yang Ming/Evergreen/PIL/Sinotrans Slots:T	YM Vancouver	Ever Develop	No sailing	No sailing	Harrier Hunter	YM Portland	YM Vancouver	Ever Develop	Northern Preci
7	AAT/AAA2	MOL/OOCL/PIL Slots: COSCON	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing
8	AAA/AUSE	OOCL Slots: PIL/Yang Ming/COSCO	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing
9	NZS/NZ2/NZE/KIX	OOCL/PIL/APL/ANL Slots: CMA CGM	OOCL Kuala Lumpu	ANL Wendour	Kota Loceng	Seadream	Seasmile	No sailing	Shanghai Trader+Kot	ANL Wendour	No sailing
10	NEAX/AUJ/ESA/AEA3/AEA3/NAE/AU2/FOX/NAX/C	Evergreen/COSCON/ONE/OOCL slots: TS Lir	Ital Lirica	YM Seattle	No sailing	No sailing	Ital Libera	Brooklyn Bridg	Ital Lirica	YM Seattle	SC Mara
11	AAX	ANL/APL/CMA CGM	CMA CGM Puccini	CMA CGM Stra	CMA CGM Ros	CMA CGM Cho	CMA CGM Mo:	CMA CGM Puc	CMA CGM Strauss	CMA CGM Ros	CMA CGM Cho
12	AAX2	ANL Slots: APL	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	MS Eagle	Bomar Spring
13	AANA/AUS2/AEA2	ANL/COSCON/OOCL Slots: CMA CGM	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing
14	SAS/SAS/MSAS/AU-1/AEA4/SAS/ESACO-B	ANL/COSCON Slots: Maersk Line/MOL/OOC	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing
15	AUS1/AEA1/ACE	COSCON/OOCL Slots: ANL	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing
16	A3N/Northern Express	ANL/COSCO/OOCL	Miami	CMA CGM Nev	APL Scotland	COSCO New Yc	Hammonia Istı	CMA CGM Gec	Miami	No sailing	APL Scotland+(
17	A3C	ANL/COSCO/OOCL Slots: CMA CGM	No sailing	OOCL Seoul	No sailing	ANL Gippsland	COSCO Korea	No sailing	OOCL Rotterdam+E.I	No sailing	COSCO Hambu
18	A3S	COSCO/OOCL Slots: ANL	OOCL Italy	Xin Chi Wan	No sailing	No sailing	Xin Qing Dao	OOCL Italy	COSCO Felixstowe	OOCL Dubai	OOCL Kuala Lu
19	Australia western Loop/Bommerang northern Loo	Maersk Line/MSC/ONE Slots: Hamburg Süd	Maersk Lome	Maersk Gange	Christa Schulte	No sailing	Maersk Gatesł	MOL Genesis	Luna Maersk	Carl Schulte	Northern Magı
20	AUS/SAE/Southeast Asia Australia/Bommerang so	Maersk Line/MSC/ONE Slots: Hamburg Sud	Maersk Yangtze	Carl Schulte	MOL Growth	Lisbon	Wide Charlie	MOL Guardian	Maersk Euphrates	Maersk Lome	Maersk Gange
21	Northern Star/NZA2/NZ2 /SNZ1	Maersk Line Slots: Hapag Lloyd/ONE/COSC(	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing
22	Southern Star/NZA1/NZ1/SNZ2	Maersk Line Slots: Hapag Lloyd/ONE/COSC(	Rio De La Plata	Rio De Janeiro	Rio Bravo	Rio Madeira	Rio Negro	Rio Blanco	Rio De La Plata	Rio De Janeiro	Rio Bravo
23	Triple Star (Service 073 ML file)	Maersk Line (Singapore (Tanjung Pelepas)-/	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing
24	Yoyo/Panda/AAUS SL/CAS/CAE	Maersk Line/MSC Slots: Hamburg Süd/APL/	Maersk Utah	MSC Anya	No sailing	No sailing	Maersk Virgini	Maersk Utah	MSC Anya	Northern Mon	CSL Atlantic
25	NZ1	Maersk Line (Tanjung Pelepas-Nelson (Tima	No sailing	No sailing	No sailing	No sailing	No sailing	No sailing	Songa Haydn	No sailing	Verdi

Fig. 5-1 Sea Intelligence weekly capacity dataset snapshot[3]

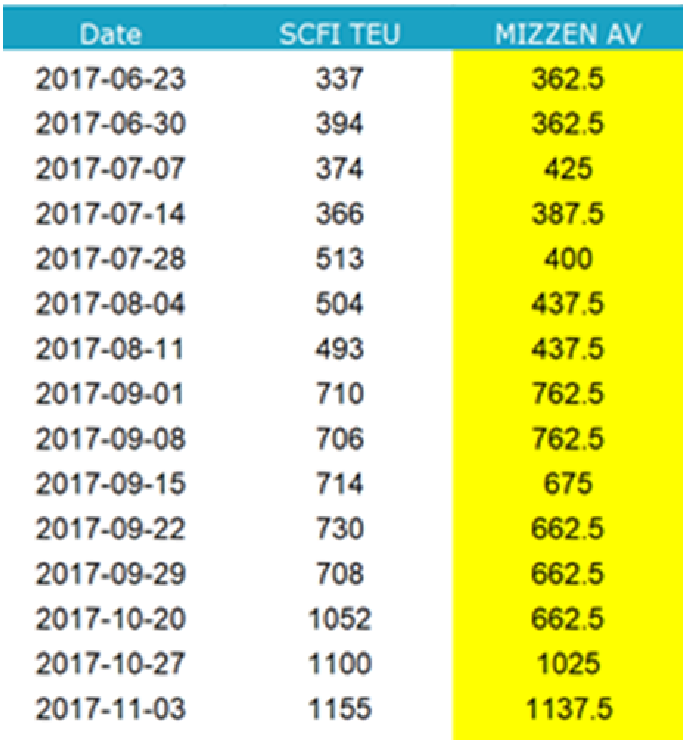
5.2.2 Shipment Price Dataset Sourcing:

The pricing data for the Asia Oceania trade lane has been gathered from two different sources: (1) Shanghai Freight Index (SCFI) [4] and (2) Mizzen group propriety data [5]. The SCFI publishes weekly shipment prices per TEU for different trade lanes. However, Mizzen records weekly shipment prices quoted in real-time for shipments. Hence, we have expected and actual pricing information for an individual week. In order to get real-time shipment prices, we have selected Mizzen prices where SCFI and Mizzen prices are not the same. Fig 5.2 (a) and (b) shows snapshots of SCFI and Mizzen price lists.



Description	Unit	Weighting	19/08/16	26/08/16	02/09/16	09/09/16
Comprehensive			597.47	596.38	763.06	781.21
Line Service:						
Europe (Base port)	/TEU	20.00%	691	695	949	943
Mediterranean (Base port)	/TEU	10.00%	594	553	519	702
USWC (Base port)	/FEU	20.00%	1159	1153	1746	1749
USEC (Base port)	/FEU	7.50%	1694	1684	2441	2447
Persian Gulf and Red Sea (Dubai)	/TEU	7.50%	230	247	317	414
Australia/New Zealand (Melbourne)	/TEU	5.00%	337	361	497	516
East/West Africa (Lagos)	/TEU	2.50%	1083	1030	1419	1173
South Africa (Durban)	/TEU	2.50%	658	740	812	838
South America (Santos)	/TEU	2.50%	2832	2711	2871	2351
West Japan (Base port)	/TEU	5.00%	195	196	200	202
East Japan (Base port)	/TEU	5.00%	188	190	193	196
Southeast Asia (Singapore)	/TEU	5.00%	54	54	54	53
Korea (Pusan)	/TEU	2.50%	88	87	87	86
Taiwan (Kaohsiung)	/TEU	2.50%	153	153	153	152
Hong Kong (Hong Kong)	/TEU	2.50%	56	56	56	56

(a)



Date	SCFI TEU	MIZZEN AV
2017-06-23	337	362.5
2017-06-30	394	362.5
2017-07-07	374	425
2017-07-14	366	387.5
2017-07-28	513	400
2017-08-04	504	437.5
2017-08-11	493	437.5
2017-09-01	710	762.5
2017-09-08	706	762.5
2017-09-15	714	675
2017-09-22	730	662.5
2017-09-29	708	662.5
2017-10-20	1052	662.5
2017-10-27	1100	1025
2017-11-03	1155	1137.5

(b)

Fig. 5-2 Snapshot of Price dataset (a) SCFI price summary (b) Mizzen price summary

5.2.3 Demand Dataset Sourcing

We started by collecting real-time shipment demand datasets from five of the international ports to form a consolidated demand dataset. These ports include Port Botany [6], Port of Melbourne[7], Fremantle Port[8], Flinders Port [9], and Port of Brisbane [10]. Figure 5.3 shows the data source for the demand dataset.

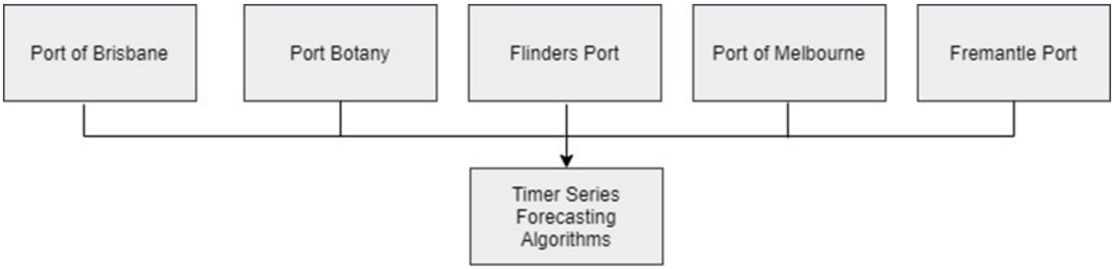
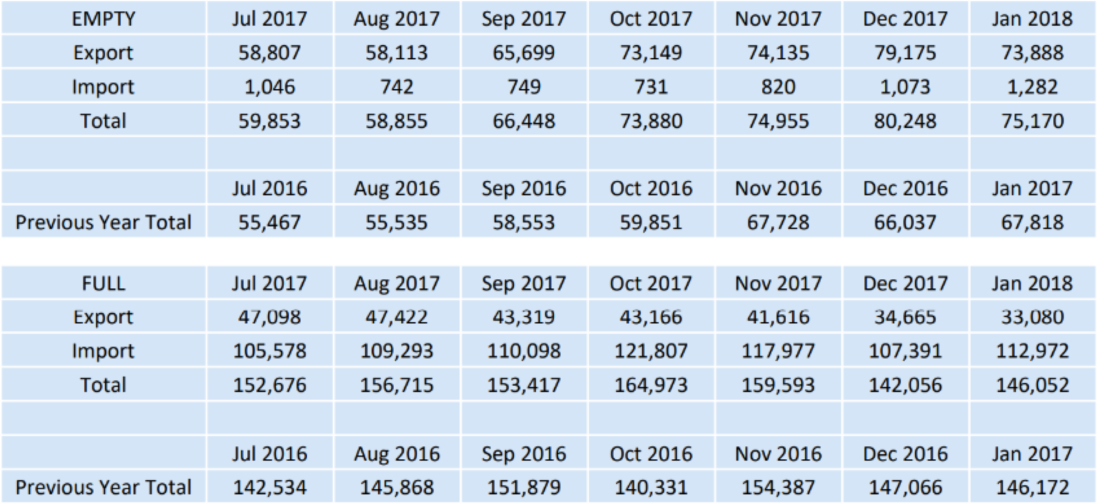


Fig. 5-3 Demand Source for Shipment Dataset

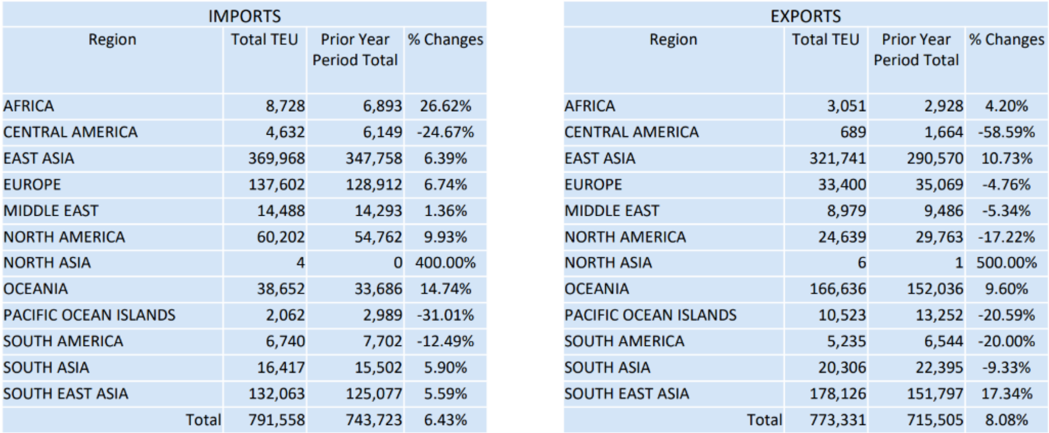
The detailed trade summary from Port Botany is shown in Fig 5.4. It consists of monthly trade summaries in TEUs for both empty and full containers (see Fig.5.4 (a)), region-based trade summaries (see Fig. 5.4 (b)) and top ten shipped commodities (see Fig 5.4 (c)). Fremantle Port and Port of Melbourne provide monthly trade summaries of all empty and full containers (see Fig 5.5 (a) and (b) respectively). However, the Flinders Port provides a region-wise break down for cargo shipment demand (see Fig.5.5 (c).) similar to Port Botany. On the other hand, the Port of Brisbane provides a commodity-based trade summary (see Fig. 5.5 (d)). From the trade summaries, it is evident that Port Botany has the maximum transparency in its trade summary.



EMPTY	Jul 2017	Aug 2017	Sep 2017	Oct 2017	Nov 2017	Dec 2017	Jan 2018
Export	58,807	58,113	65,699	73,149	74,135	79,175	73,888
Import	1,046	742	749	731	820	1,073	1,282
Total	59,853	58,855	66,448	73,880	74,955	80,248	75,170
	Jul 2016	Aug 2016	Sep 2016	Oct 2016	Nov 2016	Dec 2016	Jan 2017
Previous Year Total	55,467	55,535	58,553	59,851	67,728	66,037	67,818

FULL	Jul 2017	Aug 2017	Sep 2017	Oct 2017	Nov 2017	Dec 2017	Jan 2018
Export	47,098	47,422	43,319	43,166	41,616	34,665	33,080
Import	105,578	109,293	110,098	121,807	117,977	107,391	112,972
Total	152,676	156,715	153,417	164,973	159,593	142,056	146,052
	Jul 2016	Aug 2016	Sep 2016	Oct 2016	Nov 2016	Dec 2016	Jan 2017
Previous Year Total	142,534	145,868	151,879	140,331	154,387	147,066	146,172

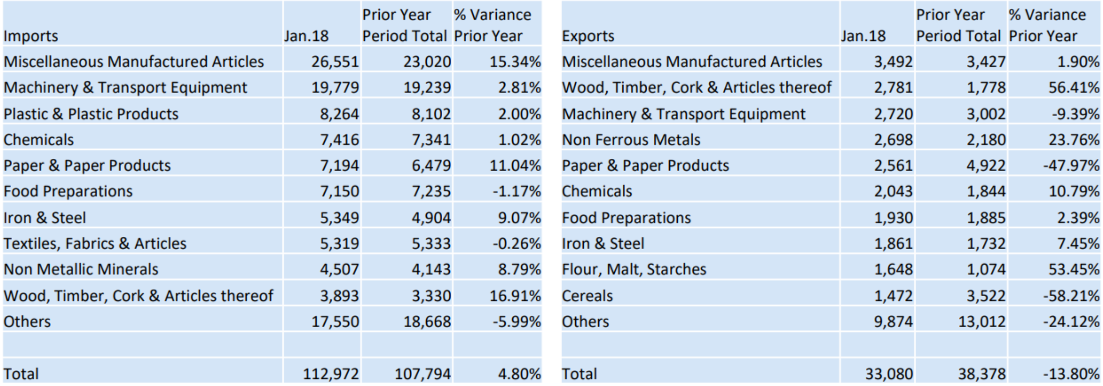
(a)



IMPORTS			
Region	Total TEU	Prior Year Period Total	% Changes
AFRICA	8,728	6,893	26.62%
CENTRAL AMERICA	4,632	6,149	-24.67%
EAST ASIA	369,968	347,758	6.39%
EUROPE	137,602	128,912	6.74%
MIDDLE EAST	14,488	14,293	1.36%
NORTH AMERICA	60,202	54,762	9.93%
NORTH ASIA	4	0	400.00%
OCEANIA	38,652	33,686	14.74%
PACIFIC OCEAN ISLANDS	2,062	2,989	-31.01%
SOUTH AMERICA	6,740	7,702	-12.49%
SOUTH ASIA	16,417	15,502	5.90%
SOUTH EAST ASIA	132,063	125,077	5.59%
Total	791,558	743,723	6.43%

EXPORTS			
Region	Total TEU	Prior Year Period Total	% Changes
AFRICA	3,051	2,928	4.20%
CENTRAL AMERICA	689	1,664	-58.59%
EAST ASIA	321,741	290,570	10.73%
EUROPE	33,400	35,069	-4.76%
MIDDLE EAST	8,979	9,486	-5.34%
NORTH AMERICA	24,639	29,763	-17.22%
NORTH ASIA	6	1	500.00%
OCEANIA	166,636	152,036	9.60%
PACIFIC OCEAN ISLANDS	10,523	13,252	-20.59%
SOUTH AMERICA	5,235	6,544	-20.00%
SOUTH ASIA	20,306	22,395	-9.33%
SOUTH EAST ASIA	178,126	151,797	17.34%
Total	773,331	715,505	8.08%

(b)

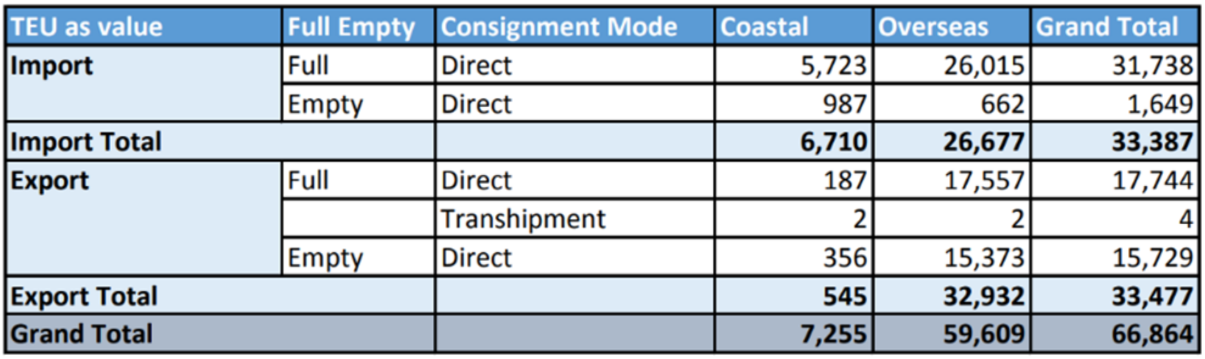


Imports	Jan.18	Prior Year Period Total	% Variance Prior Year
Miscellaneous Manufactured Articles	26,551	23,020	15.34%
Machinery & Transport Equipment	19,779	19,239	2.81%
Plastic & Plastic Products	8,264	8,102	2.00%
Chemicals	7,416	7,341	1.02%
Paper & Paper Products	7,194	6,479	11.04%
Food Preparations	7,150	7,235	-1.17%
Iron & Steel	5,349	4,904	9.07%
Textiles, Fabrics & Articles	5,319	5,333	-0.26%
Non Metallic Minerals	4,507	4,143	8.79%
Wood, Timber, Cork & Articles thereof	3,893	3,330	16.91%
Others	17,550	18,668	-5.99%
Total	112,972	107,794	4.80%

Exports	Jan.18	Prior Year Period Total	% Variance Prior Year
Miscellaneous Manufactured Articles	3,492	3,427	1.90%
Wood, Timber, Cork & Articles thereof	2,781	1,778	56.41%
Machinery & Transport Equipment	2,720	3,002	-9.39%
Non Ferrous Metals	2,698	2,180	23.76%
Paper & Paper Products	2,561	4,922	-47.97%
Chemicals	2,043	1,844	10.79%
Food Preparations	1,930	1,885	2.39%
Iron & Steel	1,861	1,732	7.45%
Flour, Malt, Starches	1,648	1,074	53.45%
Cereals	1,472	3,522	-58.21%
Others	9,874	13,012	-24.12%
Total	33,080	38,378	-13.80%

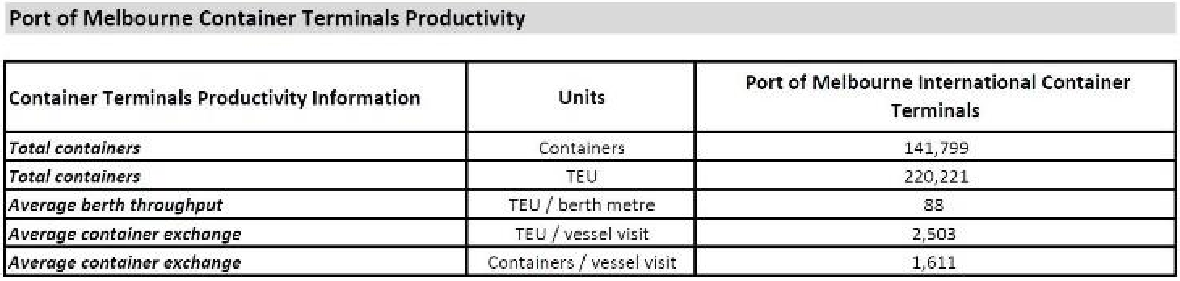
(c)

Fig. 5-4 Snapshot of Port Botany trade summaries (a) empty and full container demand (b) region-based demand (c) commodity-based demand



TEU as value	Full Empty	Consignment Mode	Coastal	Overseas	Grand Total
Import	Full	Direct	5,723	26,015	31,738
	Empty	Direct	987	662	1,649
Import Total			**6,710**	**26,677**	**33,387**
Export	Full	Direct	187	17,557	17,744
		Transhipment	2	2	4
	Empty	Direct	356	15,373	15,729
Export Total			**545**	**32,932**	**33,477**
Grand Total			**7,255**	**59,609**	**66,864**

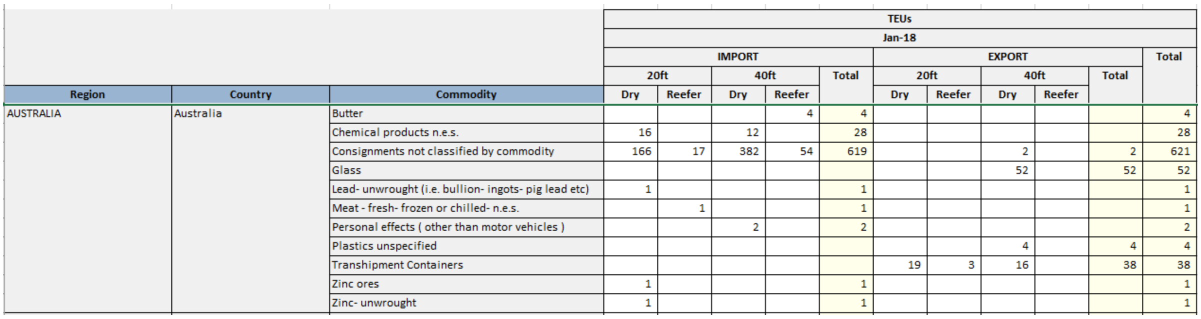
(a)



Port of Melbourne Container Terminals Productivity

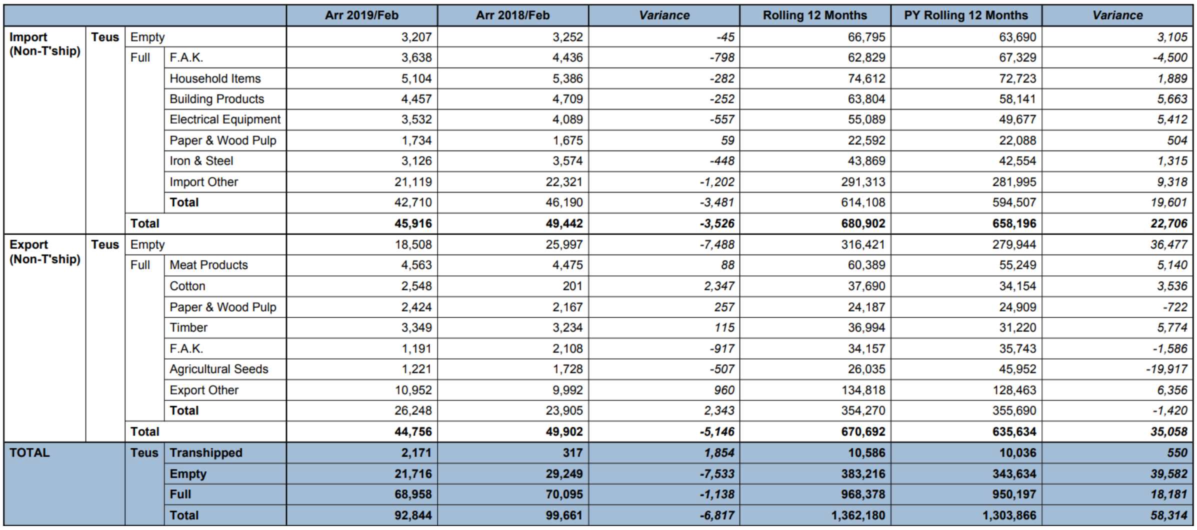
Container Terminals Productivity Information	Units	Port of Melbourne International Container Terminals
Total containers	Containers	141,799
Total containers	TEU	220,221
Average berth throughput	TEU / berth metre	88
Average container exchange	TEU / vessel visit	2,503
Average container exchange	Containers / vessel visit	1,611

(b)



			TEUs										
			Jan-18										
			IMPORT					EXPORT					Total
			20ft		40ft		Total	20ft		40ft		Total	
Region	Country	Commodity	Dry	Reefer	Dry	Reefer		Dry	Reefer	Dry	Reefer		
AUSTRALIA	Australia	Butter				4	4						4
		Chemical products n.e.s.	16		12		28						28
		Consignments not classified by commodity	166	17	382	54	619			2		2	621
		Glass								52		52	52
		Lead- unwrought (i.e. bullion- ingots- pig lead etc)	1				1						1
		Meat - fresh- frozen or chilled- n.e.s.		1			1						1
		Personal effects (other than motor vehicles)			2		2						2
		Plastics unspecified								4		4	4
		Transhipment Containers						19	3	16		38	38
		Zinc ores	1				1						1
		Zinc- unwrought	1				1						1

(c)



				Arr 2019/Feb	Arr 2018/Feb	*Variance*	Rolling 12 Months	PY Rolling 12 Months	*Variance*
Import (Non-T'ship)	**Teus**	Empty		3,207	3,252	*-45*	66,795	63,690	*3,105*
		Full	F.A.K.	3,638	4,436	*-798*	62,829	67,329	*-4,500*
			Household Items	5,104	5,386	*-282*	74,612	72,723	*1,889*
			Building Products	4,457	4,709	*-252*	63,804	58,141	*5,663*
			Electrical Equipment	3,532	4,089	*-557*	55,089	49,677	*5,412*
			Paper & Wood Pulp	1,734	1,675	*59*	22,592	22,088	*504*
			Iron & Steel	3,126	3,574	*-448*	43,869	42,554	*1,315*
			Import Other	21,119	22,321	*-1,202*	291,313	281,995	*9,318*
			Total	42,710	46,190	*-3,481*	614,108	594,507	*19,601*
		Total		**45,916**	**49,442**	***-3,526***	**680,902**	**658,196**	***22,706***
Export (Non-T'ship)	**Teus**	Empty		18,508	25,997	*-7,488*	316,421	279,944	*36,477*
		Full	Meat Products	4,563	4,475	*88*	60,389	55,249	*5,140*
			Cotton	2,548	201	*2,347*	37,690	34,154	*3,536*
			Paper & Wood Pulp	2,424	2,167	*257*	24,187	24,909	*-722*
			Timber	3,349	3,234	*115*	36,994	31,220	*5,774*
			F.A.K.	1,191	2,108	*-917*	34,157	35,743	*-1,586*
			Agricultural Seeds	1,221	1,728	*-507*	26,035	45,952	*-19,917*
			Export Other	10,952	9,992	*960*	134,818	128,463	*6,356*
			Total	26,248	23,905	*2,343*	354,270	355,690	*-1,420*
		Total		**44,756**	**49,902**	***-5,146***	**670,692**	**635,634**	***35,058***
TOTAL		**Teus**	**Transhipped**	**2,171**	**317**	***1,854***	**10,586**	**10,036**	***550***
			Empty	**21,716**	**29,249**	***-7,533***	**383,216**	**343,634**	***39,582***
			Full	**68,958**	**70,095**	***-1,138***	**968,378**	**950,197**	***18,181***
			Total	**92,844**	**99,661**	***-6,817***	**1,362,180**	**1,303,866**	***58,314***

(d)

Fig. 5-5 Trade Summaries (a) Fremantle Port (b) Port of Melbourne (c) Flinders Port (d) Port of Brisbane

This demand data gathered from multiple ports have too many features and there is a necessity to perform feature selection from the vast number of variables present. In the next section, we provide a detailed overview of the feature selection methodology we have designed specifically for this research work.

5.3 Feature Extraction Process:

As discussed in chapter 1 and 2, implementing any data driven model relies heavily on the features extracted for unfolding the hidden insights in it [11, 12]. From the literature review and shipment demand dataset, it is clear that the existing feature selection methods are not suitable for the dataset under observation. This is because the demand data sources have varied features and requires an expert opinion from the shipment domain. To overcome this gap, we have proposed a novel feature selection algorithm named *Domain Driven Selective Wrapping* (DSW) [13]. DSW is a unique framework that combines the advantages of filters and wrappers while reducing their disadvantages. The proposed framework uses an expert's domain knowledge for the selection of subsets. DSW is a hybrid framework that offers a twofold method for the selection of data in conjunction with domain knowledge. DSW also allows the incorporation of user defined parameters in FS to increase feature selection efficiency. Figure 5.6 below shows the block diagram of the proposed DSW framework.

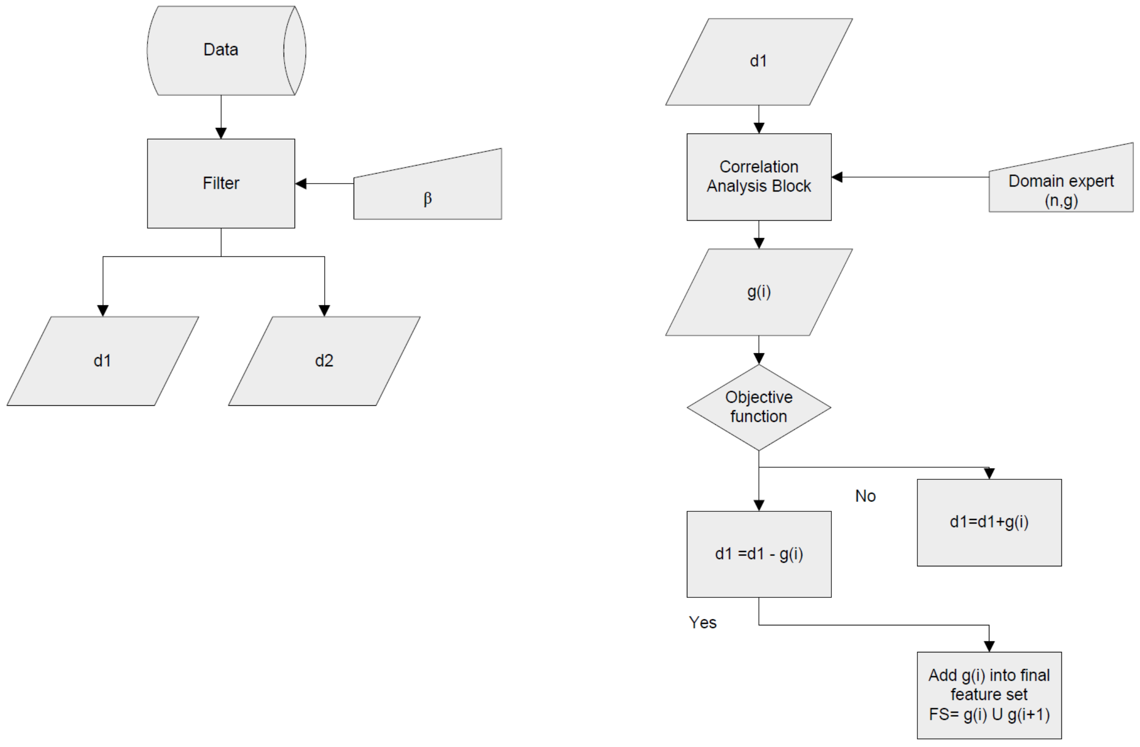


Fig. 5-6 Domain Driven Selective Wrapping (a) Phase-1 (b) Phase-2

5.3.1 Phase 1: Domain Knowledge-Based Filtration:

In this phase, preliminary filtration is performed based on the domain expert's knowledge. At the start of this phase, the whole data set is fed into the filtration block for analysis. Depending on the data content, a parameter β is also initialized and is given as input to the filtration phase. The value of β is set by the domain expert depending on the application and depicts the number of groups in which the data has to be filtered for further processing. Let D be the original data set and d represents the filtered subsets, then

$$D = d_1 \cup d_2 \cup \ldots \cup d_\beta \tag{5.1}$$

$$d_i \cap d_j = \phi, \qquad \mathrm{i} \neq \mathrm{j}, \qquad \mathrm{i}, \mathrm{j} = 1,2, \ldots, \beta$$

In equation (5.1), D represents the original dataset and $d_1 \cup d_2 \cup \ldots \cup d_\beta$, on the right-hand side of equation (5.1) depicts that it is the union of all the data sets having information of varied domain attributes. β depicts the total number of attributes whose datasets are merged into the main dataset. In other words, D is the union if all the subsets of the datasets having varied attributes of the domain under study.

5.3.2 Phase 2: Domain Knowledge-Based Correlation Analysis and Wrapping:

The second phase is wrapping one of the filtered datasets d using parameters defined by the domain expert. These parameters include n, representing the total number of features in a subset where $n \geq 2$ and $'g_i$ representing the subset identifier. The subset g_i depicts a correlated group of features in the subset. Thus, in an i^{th} iteration, n number of features are selected related to the same g_i. n and g_i are decided by analyzing the correlation between features by the domain expert. A subset is selected from the input dataset d for the i^{th} iteration. Hence, dataset d can be expressed as

$$d = \{g_{1,}, g_2, \ldots . g_n\} \tag{5.2}$$

Also

$$\{g_{1,}, g_2, \ldots . g_n\} \subseteq d \tag{5.3}$$

Features belonging to the same group become part of the same subset. The selected subset is used to check the effectiveness of the objective function. If the subset is able to provide the maximum value of the objective function, all the features forming the subsets are removed from the original dataset. Hence, the remaining dataset can be described as

$$d = d - g_i \tag{5.4}$$

If the objective function is not achieved, the subset remains part of the original dataset. In this case, Equation (5.4) becomes equation (5.5).

$$d = d + g_i \tag{5.5}$$

The selected subset is described by Equation (5.6), where x represents the features in the i^{th} subset derived from d.

$$g_i = (x_{1,}, x_2, \ldots . x_n), \text{where } g(i) = \begin{cases} x \in g_i, & 0 < r(x) < 1 \\ x \notin g_i, & r(x) < 0 \end{cases} \tag{5.6}$$

In Equation (5.6), $r(x)$ represents the Pearson correlation coefficient, which can be calculated using Equation (5.7). Equation (5.6) shows that all the questions with a positive correlation belong to the same question group. Any value between 0 and 1 indicates a positive correlation between the set of questions. Values less than or equal to 0 indicate a negative correlation between the questions. If no data is available, the $r(x)$ calculation is highly dependent on the experience of the domain expert. Once the data is generated by the initial setup, it can be incorporated into the system. This combination of domain knowledge and data is important because the domain expert may not be able to predict hidden relationships in the data.

$$r(x) = \frac{cov(x_i, Y)}{\sqrt{var(x_i) * var(Y)}} \tag{5.7}$$

5.4 Application of DSW on the shipment demand dataset:

As discussed in the previous section, it is of utmost importance to incorporate domain expert's knowledge into the feature selection process of source data. This has led to the design of a novel and unique FS methods. We have applied DSW on the shipping data set to select features of our interest.

5.4.1.1 Phase 1: Domain Knowledge-Based Filtration:

The demand data from each port is provided to phase-1 of DSW to filter out overall demand data into different categories as advised by domain expert [5] . According to Mizzen [5], the shipping industry which is working with us as the domain experts from the Australian shipping industry, the demand data consist of both export and import demand. Thus, in the first phase, the dataset is preliminarily filtered and is partitioned into subsets defined by β. In shipping dataset $= 2$. Phase 1 of DSW can be modified as figure 5.7 and equation (5.1) can be written as equation (5.8).

$$D_{total} = \text{d(E)} \cup \text{d (I)} \tag{5.8}$$

Where d(E) refers to the shipment demand of export for both full and empty containers. Likewise, d (I) refers to shipment imports demand for full and empty containers. Hence, at the end of phase1, total demand data is segregated into 2 sub parts. Figure 5.7 shows an overview of the process for phase-1 of DSW for preliminary filtration of shipment dataset.

5.4.1.2 Phase 2: Domain Knowledge-Based Correlation Analysis and Wrapping:

Each of the subsets of the data set contains data for all of the trade lanes operating from Australia. To separate the trade-lane wise demand, the subsets are forwarded to Phase 2 of DSW. Since we have selected import demand prediction, we will place import subsets from Phase 1 into Phase 2. Figure 5.8 shows the processing of DSW Phase 2. Since we have selected Asia-Oceania trade lane import only dataset, shipment demand data for import only is provided to phase 2 of DSW. Domain expert can perform correlation analysis to identify the participating regions in the selected trade lane. To do so, the domain expert sets the n (i.e., number of trade lanes in the dataset) and $g_{(i)}$ (total regions belonging to the similar trade lanes are identified) for analysis.

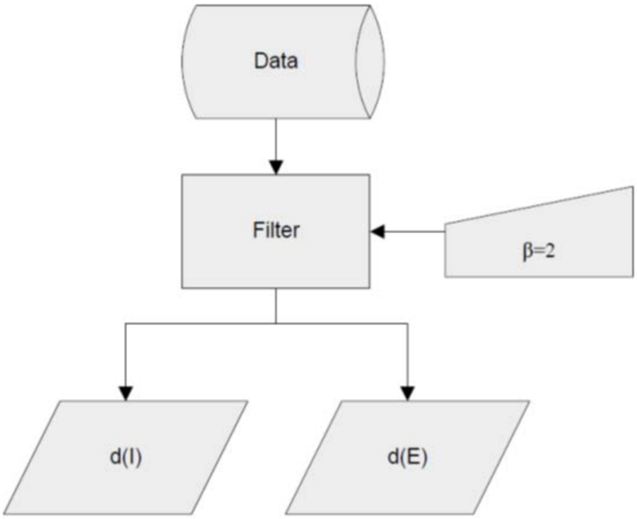


Fig. 5-7 DSW phase-1

Once the trade lane related import demand variables are selected, they are passed through the objective function. The objective function can be any selection criteria that can determine if the features selected are correlated or not. In this thesis, domain expert is the only resource that can serve the function of objective function. For this applied research, we have' Mizzen Group'[5] on board with us who are providing us every insight into the dataset, data sourcing and the useful variables for our research. They are providing approval for each and every variable of the dataset that are necessary for conducting this research to achieve the aim of this research. However, in other application, ML based models can be used as objective function.

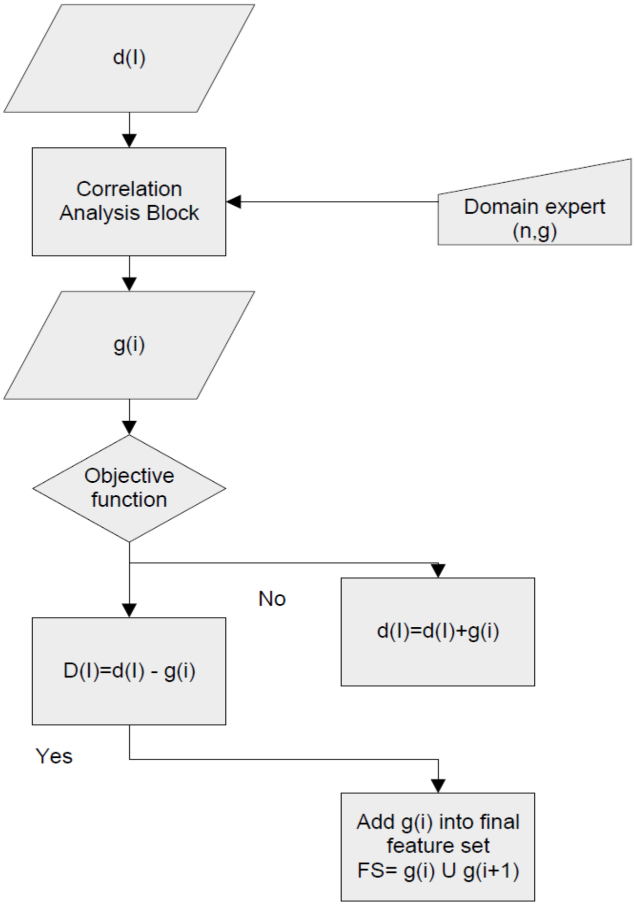


Fig. 5-8 DSW Phase-2

In figure 5.8, the phase-2 of DSW for import dataset is shown. The first step is to perform correlation analysis over the import dataset $d(I)$. In the correlation analysis, the value pair $(n\,(i), g\,(i))$ is set by experts from Mizzen group. Using the value of $n\,(i)$, total number of features for the iteration $g\,(i)$ are selected which are strongly dependent. This $g\,(i)$ is approved/disapproved by the objective function. As mentioned earlier in this chapter, the objective function is based on data analyst choice; it can be either an already existing algorithm, a customized algorithm [13] or the approval of domain expert. In our case, we have the Australian shipping industry experts from Mizzen group to approve the feature set. If the objective function (Mizzen employee) approves the $g\,(i)$, then it is removed from the main feature set i.e. in this case $d(I)$ and added into the final FS. The process continues for $g\,(i+1)$ iteration until all the features are added into to the final FS.

It can be seen that the entire process is extremely domain expert dependent and equally has an option of using already existing data-based algorithms. The purpose of the framework is to provide a skeleton in which all options can be incorporated and can be applied to the complex shipping industry's dataset.

If the region g (i) belongs to the same trade lane (i.e. trade lane under observation. Asia-Oceania in this case), it is removed from the initial dataset and is placed into the final feature set (FS) for use. However, if the data is not related to the trade lane under consideration, the data is added back to the main dataset i.e., $d(I\,)$ and hence, to be assessed again for other possible trade lane participation. We have selected n= 1 and g= 4. Thus, it can be concluded that for a single trade lane i.e. , Asia Oceania there should be four regions that are participating in the said trade lane. Hence for this research, equation (5.2) becomes equation 5.9 as below:

$$d(Asia - Oceania) = g_{East\ Asia}, g_{North-Asia}, g_{South-Asia}, g_{Sout\ East\ Asia} \qquad (5.9)$$

And

$$\{g_{East\ Asia}, g_{North-Asia}, g_{South-Asia}, g_{Sout\ East\ Asia}\} \subseteq \mathrm{d(Asia\text{-}Oceania)} \qquad (5.10)$$

Thus, the final feature set (FS) can be written as equation (5.11) given below:

$$FS = \{g_{East\ Asia} \cup g_{North-Asia} \cup g_{South-Asia} \cup g_{Sout\ East\ Asia}\} \qquad (5.11)$$

Where FS represents the feature set we have extracted via DSW. This FS has got the features that we intend to use in this research. However, from port statistic, the total import demand can be calculated using the equation (5.12).

$$D^t(I) = C_F(I) + C_E(I) \qquad (5.12)$$

Where $D^t(I)$ represents total import shipment demand, $C_F(I)$ and $C_E(I)$ represents a total of full and empty containers for imports. In order to calculate the total import shipment demand for the Asia Oceania trade lane, total demand from all the participating regional ports in the trade lane (i.e., East Asia, North Asia, South Asia and South-East Asia) are added. Subtracting the sum of total demand of all the participating ports from the total import shipment demand for all the trade lanes ($D^t(I)$) gives the total import shipment demand for the Asia Oceania Trade lane. This can be expressed in equation (5.14).

$$D^t_{Asia-Oceania}(I) = D^t(I) - \mathrm{FS} \qquad (5.13)$$

$$\begin{aligned} &D^t_{Asia-Oceania}(I) \\ &\quad = D^t(I) - \{D^t_{East\ Asia}(I) + D^t_{North\ Asia}(I) + D^t_{South\ Asia}(I) \\ &\quad + D^t_{South\ East\ Asia}(I)\} \end{aligned} \qquad (5.14)$$

The trade data set acquired so far is the combination of full and empty containers coming in and going out from the Australian port over the Asia-

Oceania trade lane. Full containers are the number of containers (measured in TEUs) that are coming in and out filled with goods. In contrast, the empty containers are the ones which are being emptied on the way before entering into Australian water or are moving back without goods filled in them. However, in the total shipment demand, both the filled and empty containers are equally important. The feature in the dataset is shown in Table 5.1. In table 5.1, the date refers to the first day of the week and provides insight for the same whole week. Hence, for both imports and exports, the dataset contains the total number of filled incoming (inbound) and outgoing (outbound) containers as well as empty containers.

Table 5-1 List of Demand Dataset Features [3]

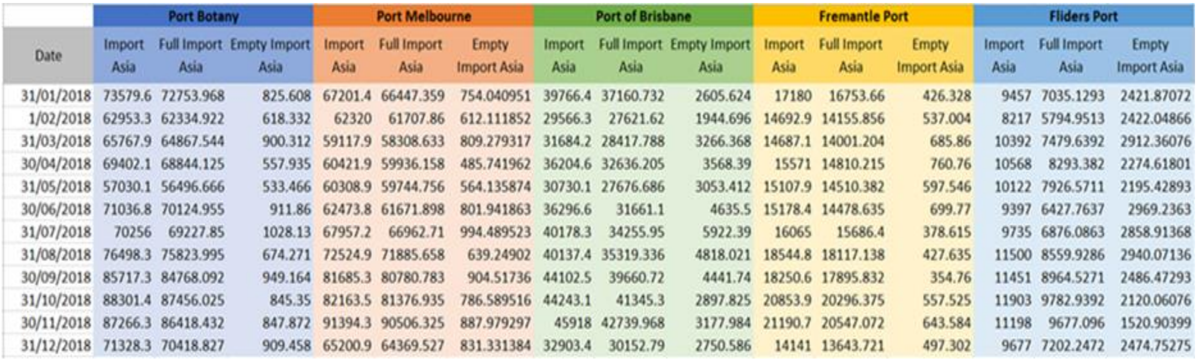
Region	Feature Name	Description.
Asia Oceania	Imports	Number of inbound containers
	Full	Total Inbound container (filled)
	Empty	Total Inbound container(Empty)
	Date	Starting day of week
	Exports	Number of outbound containers
	Full	Total outbound container (filled)
	Empty	Total outbound container(Empty)
	Date	Starting day of week

In equation (5.14), $D_{East\ Asia}^{t}(I)$ presents the total shipment imports demand of East Asia. $D_{North\ Asia}^{t}(I)$, presents the total shipment imports demand of North Asia, $D_{South\ Asia}^{t}(I)$ presents the total shipment imports demand of South Asia and $D_{South\ East\ Asia}^{t}(I)$ presents total shipment imports demand of South-East Asia. This segregates the import shipment demand for the Asia Oceania trade lane. However, for the other three Australian Ports (i.e., Port of Brisbane, Port of Melbourne and Fremantle Port), only total shipment demand is provided in their trade summaries. No division of shipment demand based on trade lane is available. To do so, we have calculated the ratio for Asia-Oceania trade lane shipment demand from Port Botany. Although we do have Flinders Port with the same statistics, Port Botany is a big port in comparison to Flinders Port and the ratio of import shipment demand for Asia Oceania trade lane can represent the remainder of the ports [11]. This ratio for Asia Oceania imports shipment demand can be calculated using equation (5.15) as shown below[3].

$$R^{t}_{Asia-Oceania}(I) = {D^{t}_{Asia-Oceania}(I)} \Big/ {D^{t}(I)} \tag{5.15}$$

Where $R^{t}_{Asia-Oceania}(I)$ represents calculated ratio discussed in the above paragraph. Thus, import shipment demand specifically for Asia Oceania trade lane for Port of Brisbane, Port of Melbourne and Fremantle Port can be calculated using equation (5.16). This shipment demand is in a monthly view as shown in Fig 5.9.

$$D^{t}_{Asia-Oceania}(I) = R_{Asia-Oceania}(I) * D^{t}(I) \tag{5.16}$$



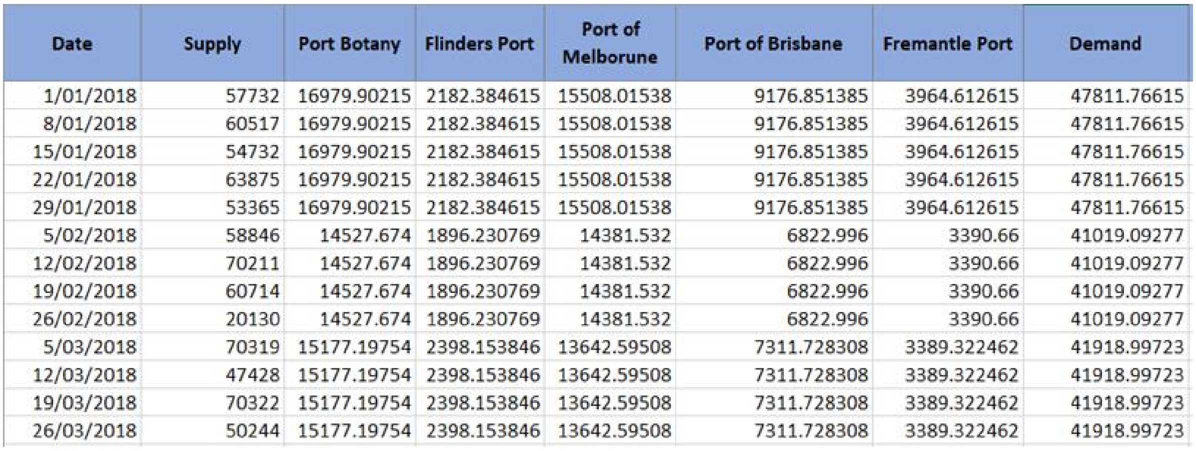
	Port Botany			Port Melbourne			Port of Brisbane			Fremantle Port			Fliders Port		
Date	Import Asia	Full Import Asia	Empty Import Asia	Import Asia	Full Import Asia	Empty Import Asia	Import Asia	Full Import Asia	Empty Import Asia	Import Asia	Full Import Asia	Empty Import Asia	Import Asia	Full Import Asia	Empty Import Asia
31/01/2018	73579.6	72753.968	825.608	67201.4	66447.359	754.040951	39766.4	37160.732	2605.624	17180	16753.66	426.328	9457	7035.1293	2421.87072
1/02/2018	62953.3	62334.922	618.332	62320	61707.86	612.111852	29566.3	27621.62	1944.696	14692.9	14155.856	537.004	8217	5794.9513	2422.04866
31/03/2018	65767.9	64867.544	900.312	59117.9	58308.633	809.279317	31684.2	28417.788	3266.368	14687.1	14001.204	685.86	10392	7479.6392	2912.36076
30/04/2018	69402.1	68844.125	557.935	60421.9	59936.158	485.741962	36204.6	32636.205	3568.39	15571	14810.215	760.76	10568	8293.382	2274.61801
31/05/2018	57030.1	56496.666	533.466	60308.9	59744.756	564.135874	30730.1	27676.686	3053.412	15107.9	14510.382	597.546	10122	7926.5711	2195.42893
30/06/2018	71036.8	70124.955	911.86	62473.8	61671.898	801.941863	36296.6	31661.1	4635.5	15178.4	14478.635	699.77	9397	6427.7637	2969.2363
31/07/2018	70256	69227.85	1028.13	67957.2	66962.71	994.489523	40178.3	34255.95	5922.39	16065	15686.4	378.615	9735	6876.0863	2858.91368
31/08/2018	76498.3	75823.995	674.271	72524.9	71885.658	639.24902	40137.4	35319.336	4818.021	18544.8	18117.138	427.635	11500	8559.9286	2940.07136
30/09/2018	85717.3	84768.092	949.164	81685.3	80780.783	904.51736	44102.5	39660.72	4441.74	18250.6	17895.832	354.76	11451	8964.5271	2486.47293
31/10/2018	88301.4	87456.025	845.35	82163.5	81376.935	786.589516	44243.1	41345.3	2897.825	20853.9	20296.375	557.525	11903	9782.9392	2120.06076
30/11/2018	87266.3	86418.432	847.872	91394.3	90506.325	887.979297	45918	42739.968	3177.984	21190.7	20547.072	643.584	11198	9677.096	1520.90399
31/12/2018	71328.3	70418.827	909.458	65200.9	64369.527	831.331384	32903.4	30152.79	2750.586	14141	13643.721	497.302	9677	7202.2472	2474.75275

Fig. 5-9 Monthly demand for Asia-Oceania trade lane in the Australian shipping industry

Since the dataset for shipment price and capacity are available in a weekly view, this shipment demand has to be converted into a weekly view. To do so, equations (5.17) is used.

$$D^t_{weekly}(I) = {D^t_{Asia-Oceania}(I) * 12}/_{52} \qquad (5.17)$$

Once the weekly shipment demand is calculated for each port, they are summed up to get total shipment demand for the Asia Oceania trade for Australia (see Fig. 5.10).



Date	Supply	Port Botany	Flinders Port	Port of Melborune	Port of Brisbane	Fremantle Port	Demand
1/01/2018	57732	16979.90215	2182.384615	15508.01538	9176.851385	3964.612615	47811.76615
8/01/2018	60517	16979.90215	2182.384615	15508.01538	9176.851385	3964.612615	47811.76615
15/01/2018	54732	16979.90215	2182.384615	15508.01538	9176.851385	3964.612615	47811.76615
22/01/2018	63875	16979.90215	2182.384615	15508.01538	9176.851385	3964.612615	47811.76615
29/01/2018	53365	16979.90215	2182.384615	15508.01538	9176.851385	3964.612615	47811.76615
5/02/2018	58846	14527.674	1896.230769	14381.532	6822.996	3390.66	41019.09277
12/02/2018	70211	14527.674	1896.230769	14381.532	6822.996	3390.66	41019.09277
19/02/2018	60714	14527.674	1896.230769	14381.532	6822.996	3390.66	41019.09277
26/02/2018	20130	14527.674	1896.230769	14381.532	6822.996	3390.66	41019.09277
5/03/2018	70319	15177.19754	2398.153846	13642.59508	7311.728308	3389.322462	41918.99723
12/03/2018	47428	15177.19754	2398.153846	13642.59508	7311.728308	3389.322462	41918.99723
19/03/2018	70322	15177.19754	2398.153846	13642.59508	7311.728308	3389.322462	41918.99723
26/03/2018	50244	15177.19754	2398.153846	13642.59508	7311.728308	3389.322462	41918.99723

Fig. 5-10 Weekly demand for Asia-Oceania trade lane in the Australian shipping industry [3]

5.5 Data Pre-processing and Cleansing:

Data pre-processing and cleansing are the data mining techniques that transforms raw data into meaningful insights[14]. Data directly taken from the sources is likely to have numerous inconsistencies. The steps involved in data pre-processing is very much dependent over the data set and its usage in any data driven application [16]. In this research the steps involved in data pre-processing and cleansing are shown in figure 5.11. The first step is to select horizon from vast shipping dataset followed by trade lane selection. Finally, the import/export data and trade lane specific data values are segregated using DSW as explained in section 5.3. Once the feature selection is performed, it is followed by missing values removal, the final dataset is analyzed making useful insights. Finally, there is a need to understand the dataset. To do so, data distribution is an excellent tool to analysis what possible values are present in it. In addition to understand data distribution, study of the regression plots is very useful to understand which of the variables have a positive or negative relation with each other.

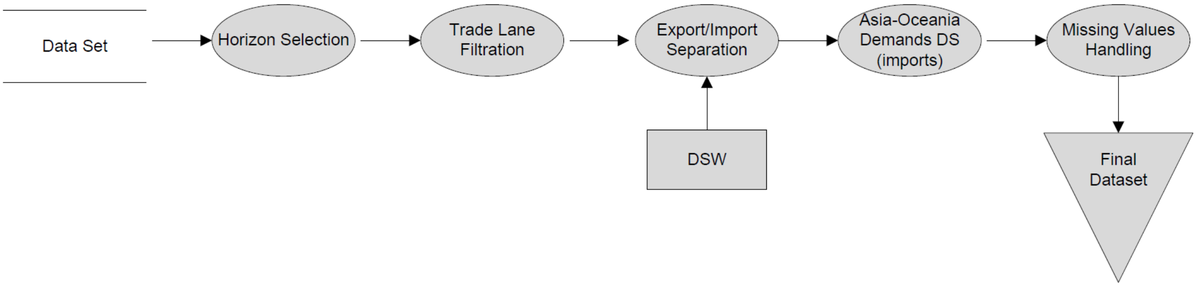


Fig. 5-11 Data Cleansing Process

5.5.1 Horizon Selection:

The first step is horizon selection. This is done by simple filtration using the expertise from domain expert. Time horizon of three years of weekly data is selected for all the data sets i.e. shipment demand dataset, shipment available capacity dataset and shipment pricing dataset.

5.5.2 Missing values

From the last few decades, the problem of missing values in the dataset is quite a serious issue. Not all the data values are logged into the sourced dataset [17]. There are varied techniques to handle the missing values in the dataset [18]. For demand dataset, the missing values are handled using averaging techniques [19]. The missing demand values are filled with the calculated average demand from the previous and coming year's similar timestamps. In capacity dataset, there are no missing values in the available shipping capacity (supply) dataset.

However, the missing values in the pricing data are handled using forward fill methodology and those of shipment demand data are filled by using an average from the previous or coming year at the similar timestamps (whichever is applicable)[20]. Figure 5.12 shows the container shipment demand (shown on the x-axis of figure 5.12) for the Asia-Oceania trade lane (imports only) for the year 2016 through 2018 (presented on the y-axis of figure 5.12).

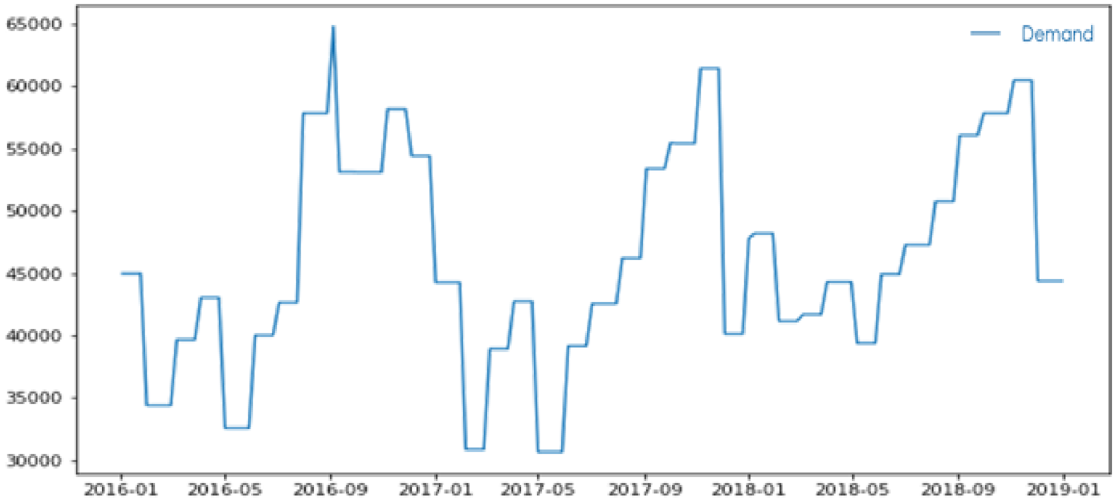
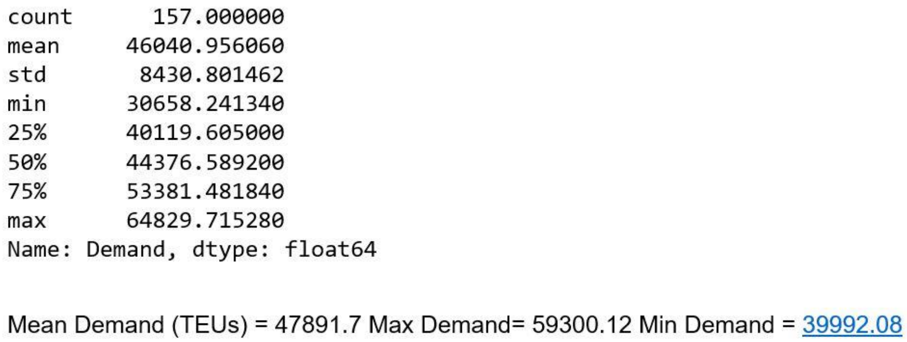


Fig. 5-12 Shipment demand dataset visual

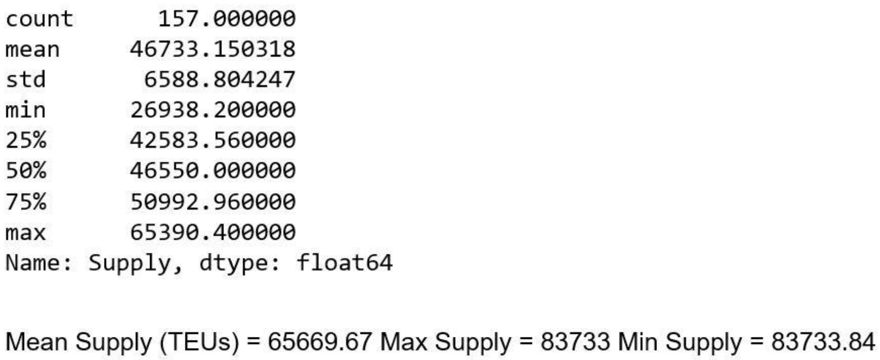
The individual data description of all the features in the dataset can be seen in figure 5.13 (a), (b) and (c) respectively.



```
count       157.000000
mean      46040.956060
std        8430.801462
min       30658.241340
25%       40119.605000
50%       44376.589200
75%       53381.481840
max       64829.715280
Name: Demand, dtype: float64
```

Mean Demand (TEUs) = 47891.7 Max Demand= 59300.12 Min Demand = 39992.08

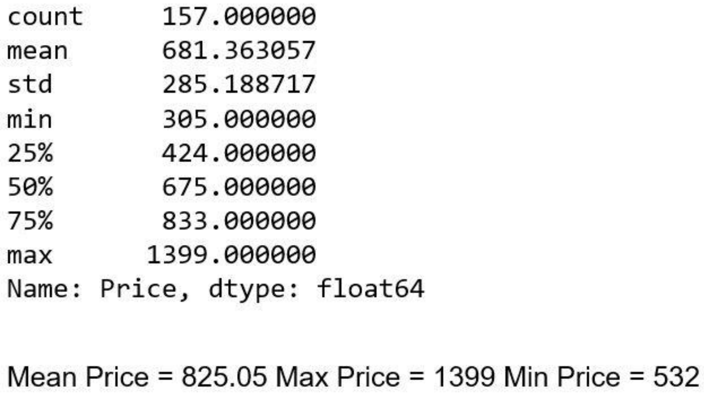
(a)



```
count      157.000000
mean     46733.150318
std       6588.804247
min      26938.200000
25%      42583.560000
50%      46550.000000
75%      50992.960000
max      65390.400000
Name: Supply, dtype: float64
```

Mean Supply (TEUs) = 65669.67 Max Supply = 83733 Min Supply = 83733.84

(b)



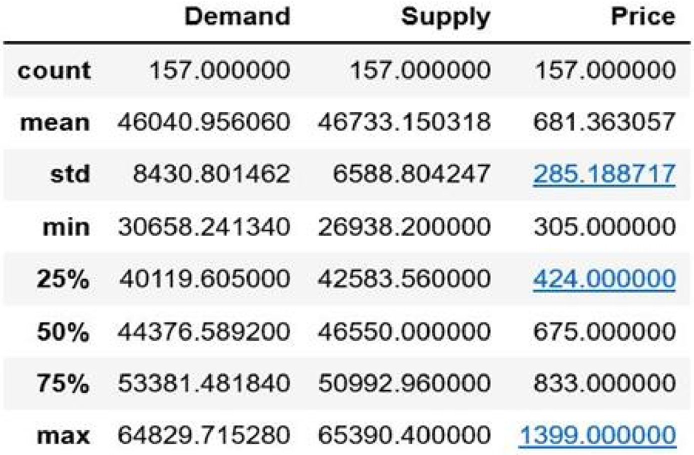
```
count      157.000000
mean       681.363057
std        285.188717
min        305.000000
25%        424.000000
50%        675.000000
75%        833.000000
max       1399.000000
Name: Price, dtype: float64
```

Mean Price = 825.05 Max Price = 1399 Min Price = 532

(c)

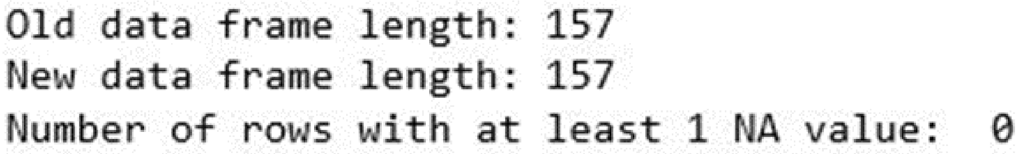
Fig. 5-13 Data description (a) demand data (b) available shipping capacity (supply) data (c) price data

The data description of the final dataset is shown in Fig 5.14. Missing values are handled by filling with the calculated average demand from the previous and coming year's similar timestamps [21]. Figure 5.15 shows the missing values existence status (if any) in the dataset.



	Demand	Supply	Price
count	157.000000	157.000000	157.000000
mean	46040.956060	46733.150318	681.363057
std	8430.801462	6588.804247	285.188717
min	30658.241340	26938.200000	305.000000
25%	40119.605000	42583.560000	424.000000
50%	44376.589200	46550.000000	675.000000
75%	53381.481840	50992.960000	833.000000
max	64829.715280	65390.400000	1399.000000

Fig. 5-14 Snapshot of the Australian shipping dataset for Asia-Oceania trade lane

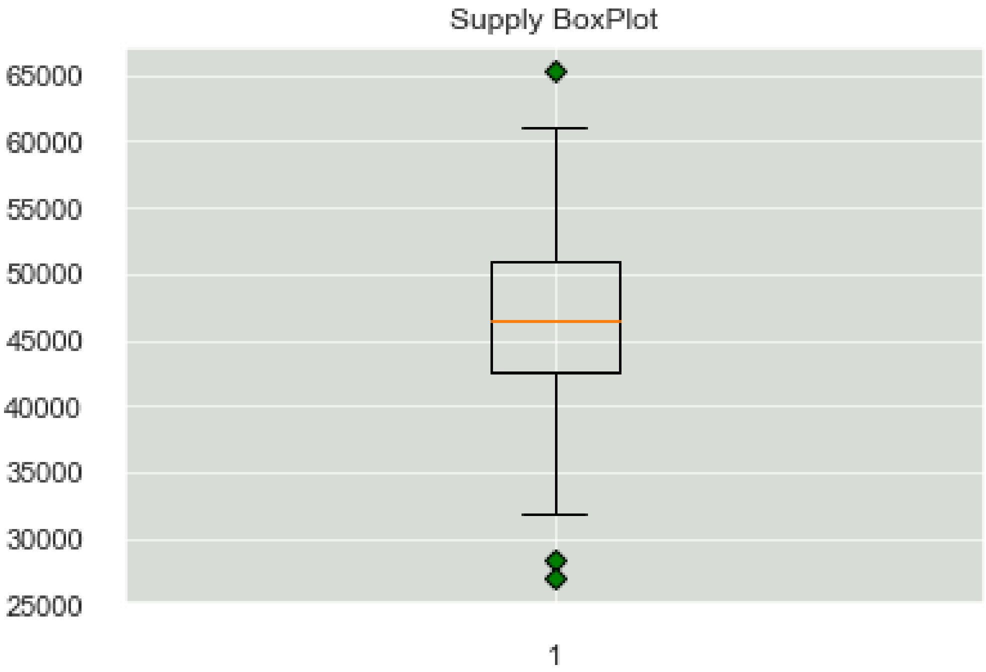


```
Old data frame length: 157
New data frame length: 157
Number of rows with at least 1 NA value:  0
```

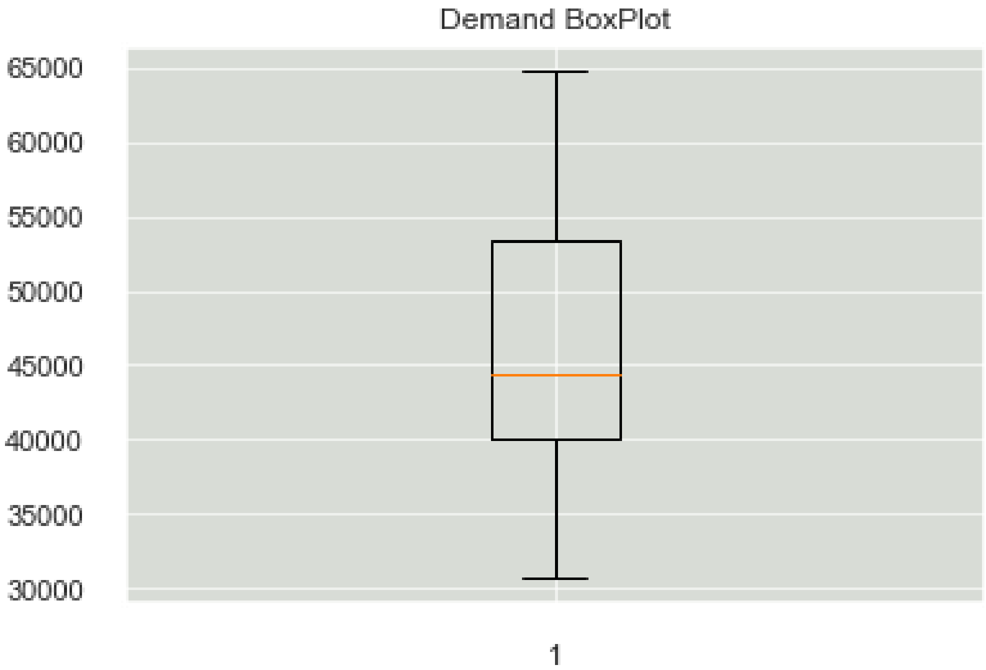
Fig. 5-15 Missing Value Checking

5.5.3 Outlier Detection and removal:

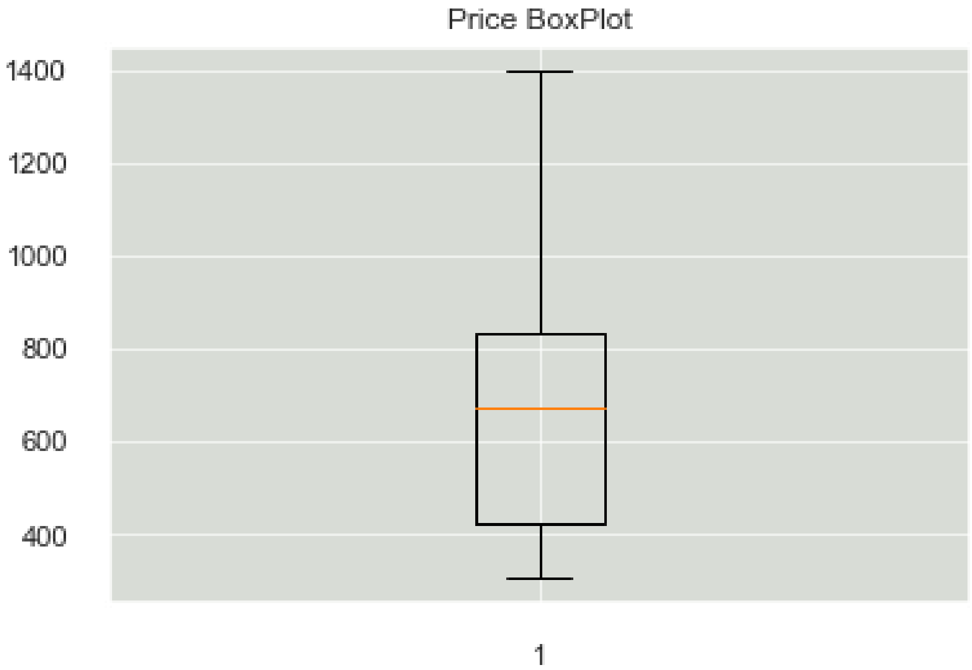
Outliers may be treated as rare data points that fall above or below normal data point's density/region. In other words, outlier detection may be treated as searching the data points with high similarities or densities and separating them from the data points which fall away from these normal point [22]. The presence of outliers greatly affect the performance of the data driven models adversely [23]. Thus, the next step is to remove any outliers from the dataset. To analyze the outliers, we have used boxplots [24]. Figure 5.16 presents the box plots to determine any possible outliers. From figure 5.16 (a), it is evident that there is exist few outliers in the available capacity (supply) variable. However, for demand and price, there exist no such outlier.



.(a)


Demand BoxPlot
65000
60000
55000
50000
45000
40000
35000
30000
1

(b)


Price BoxPlot
1400
1200
1000
800
600
400
1

(c)

Fig. 5-16 Box plot analysis for dataset variables (a) Capacity (supply) (b) Demand (c) Price

5.5.4 Data Distribution and Regression Plot:

The visual representation of dataset after imputing missing values and removing outliers provides useful insights for further data analytics [25]. For this purpose, the first step is to present the data distribution. Along with the data distribution, the regression plot is also shown which clarifies the relationship between the variables [2] .The Figure 5.17 shows the data distribution and regression analyses between shipment demand, and available shipping capacity with respect to the prices being quoted (named demand and supply in figure 5.17 respectively). The pair wise regression metric is presented in figure 5.18 (Also called scatter matrix).

As seen from the regression analysis and density distribution, it can be concluded that demand, price and shipment capacity do not have any definite connection between them. These do not have any positive correlation among each other. The hypothesis presented earlier in this thesis can further be augmented by looking at the regression line between the said variables is shown in figure 5.19. Additionally, to augment the presented hypothesis, correlation analysis is also performed over the dataset [26, 27]. The quantified correlation heat map is generated and is shown in figure 5.20.

Once the relationships between the variable are determined, we are ready to start with data analytics and use it for ML algorithm. The next step is to divide the dataset into training and test data.

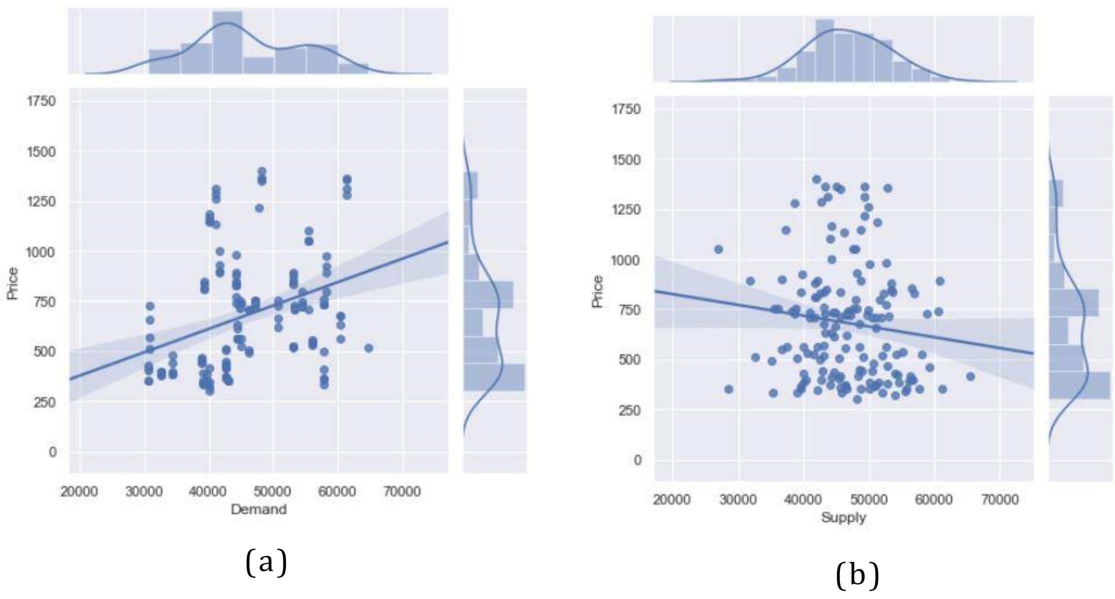


Fig. 5-17 Density distribution and regression plots (a) Shipment Demand and prices (b) Shipment available capacity (supply) and prices

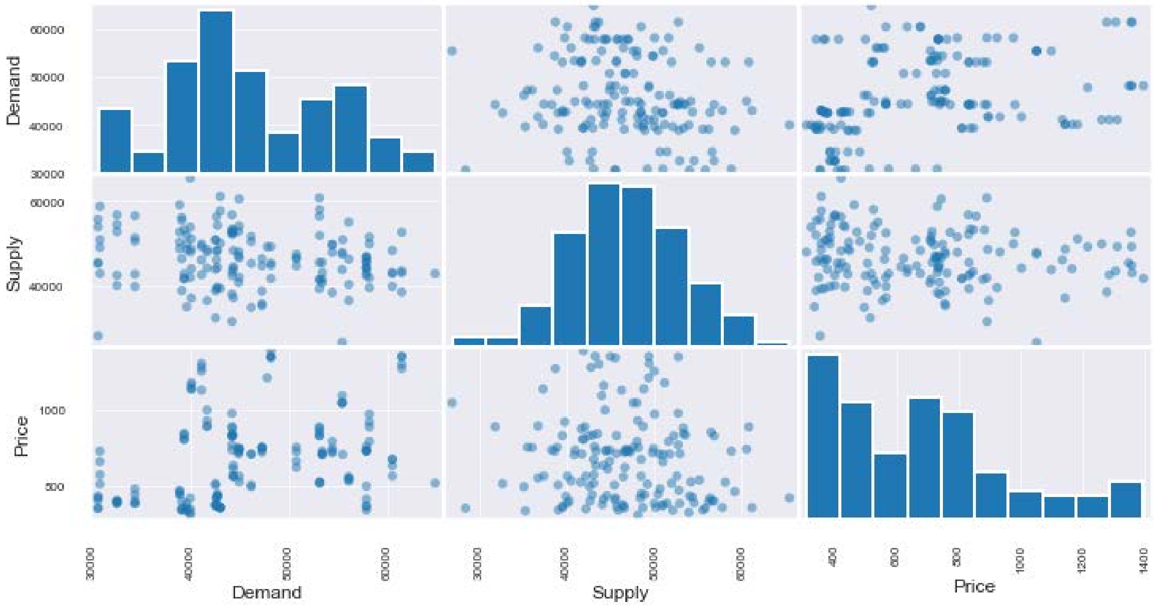


Fig. 5-18 Scatter matrix of dataset

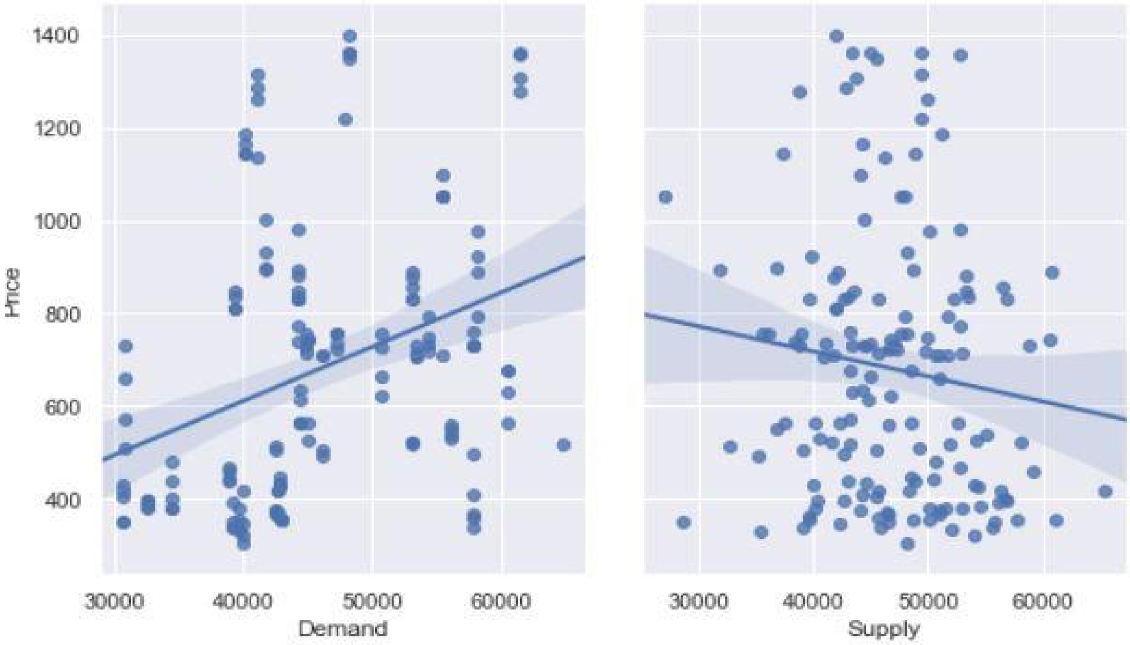


Fig. 5-19 Regression plot between Demand, capacity (supply) and prices.

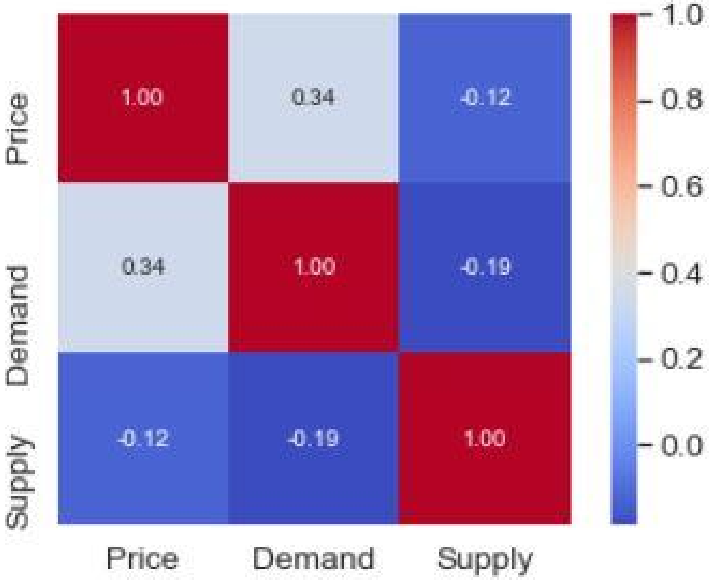


Fig. 5-20 Correlation heat map between demand, shipping capacity (supply) and prices.

5.5.5 Test Train Split:

In order to segregate the data set into training and test data, is divided into two parts. Train and test dataset (see figure 5.21). 70% of the data is used for training the model, the remaining 30% of data is used for testing the model. In time series models, usual test train split does not work as the values are dependent on time [28]. Performing random partitioning can cause misleading results. Hence, we have selected a cut-off date that corresponds to approximately 70% of dataset i.e. , Feb 2018 (see vertical red line in figure 5.21) for training data in order to capture enough seasonality and trends of the time series under observation and we have use rest of the data as test data.

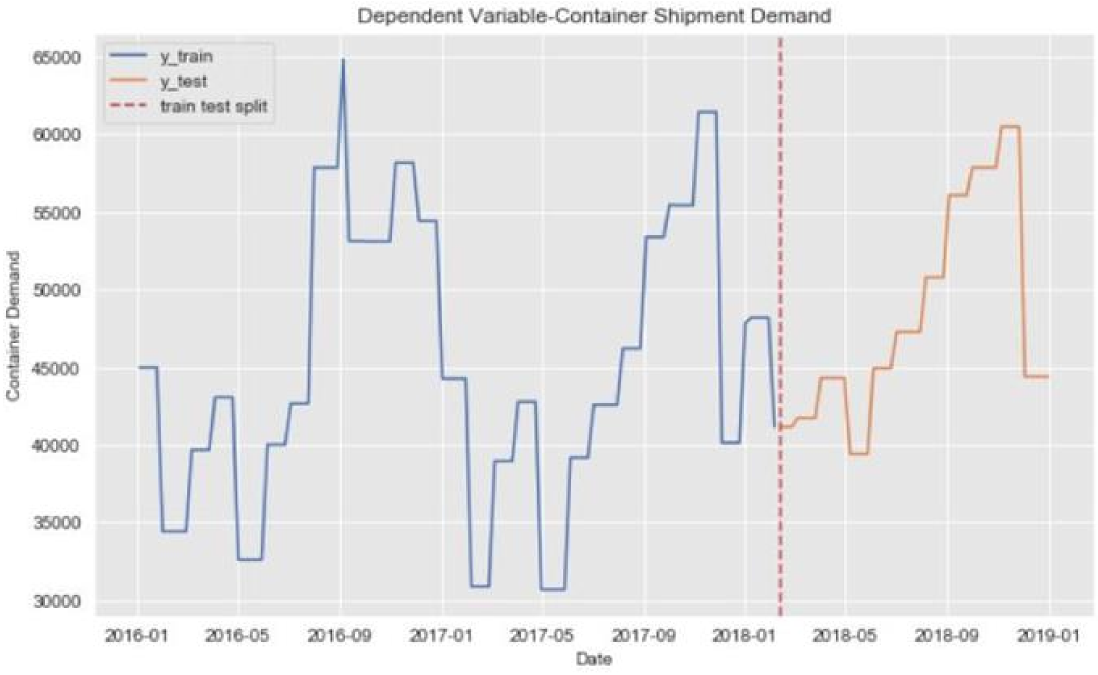


Fig. 5-21 Test train split of demand dataset (70-30)

5.6 Summary:

This chapter explains the data sourcing process for creating shipment demand dataset. Once the dataset is collected, it is cleansed to discard unusable features followed by data pre-processing. Finally, the dataset is segregated into train and test data to use it for the data driven model.

In the next chapter we discuss the optimal shipment price calculation in detail. This chapter explains how to determine optimal price that can provide business to the shipping industry's stakeholders.

5.7 Reference:

1. www.Sea-Intelligence.com. Sea-intelligence

2. Brook, R.J. and G.C. Arnold, Applied regression analysis and experimental design. 2018: CRC Press.

3. Ubaid, A., F. Hussain, and J. Charles, Modeling Shipment Spot Pricing in the Australian Container Shipping Industry: Case of ASIA-OCEANIA trade lane. Knowledge-Based Systems, 2020. 210: p. 106483.

4. Index, S.F. www.en.sse.net.cn/indices/scfinew.jsp. [cited 3 May.

5. Mizzen Group Pty Ltd. Available from: https://www.mizzengroup.com/.

6. www.nswports.com.au/resources/trade-results/. Port Botany, NSW, Australia. [cited 2019 27 May].

7. www.portofmelbourne.com/about-us/trade-statistics/monthly-trade-reports/. Port of Melbourne, VIC, Australia. 27 May 2019].

8. Fremantleport ,WA, Australia. [cited 2019 27 May]; Available from: www.fremantleports.com.au/trade-business/container-traffic-reports

9. www.flindersports.com.au/ports-facilities/port-statistics/. Flinders Port, SA, Australia. [cited 2019 27 May].

10. www.portbris.com.au/Operations-and-Trade/Trade-Development/. Port of Brisbane, QLD,Australia. [cited 2019 27 May].

11. Dong, G. and H. Liu, Feature engineering for machine learning and data analytics. 2018: CRC Press.

12. Shang, C. and F. You, Data analytics and machine learning for smart process manufacturing: recent advances and perspectives in the big data era. Engineering, 2019. 5(6): p. 1010-1016.

13. Ubaid, A., F. Dong, and F.K. Hussain. Framework for Feature Selection in Health Assessment Systems. in Advanced Information Networking and Applications. 2020. Cham: Springer International Publishing.

14. Obaid, H.S., S.A. Dheyab, and S.S. Sabry. The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning. in 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON). 2019. IEEE.

15. Salvador García, J.L., Francisco Herrera, Data Preprocessing in Data Mining. Vol. 72. Springer, Cham.

16.Ramírez-Gallego, S., et al., A survey on data preprocessing for data stream mining: Current status and future directions. Neurocomputing, 2017. 239: p. 39-57.

17.Enders, C.K., Multiple imputation as a flexible tool for missing data handling in clinical research. Behaviour Research and Therapy, 2017. 98: p. 4-18.

18.Wu, C. and M.E. Thompson, Methods for Handling Missing Data, in Sampling Theory and Practice. 2020, Springer. p. 193-221.

19.Bertsimas, D., C. Pawlowski, and Y.D. Zhuo, From predictive methods to missing data imputation: an optimization approach. The Journal of Machine Learning Research, 2017. 18(1): p. 7133-7171.

20.www.pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html.
Pandas Official Website. [cited 2019 25 July].

21. [cited 2020 18 May]; Missing value handeling]. Available from: https://scikit-learn.org/stable/modules/impute.html.

22.Zhu, J., et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data. Annual Reviews in Control, 2018. 46: p. 107-133.

23.Nyitrai, T. and M. Virág, The effects of handling outliers on the performance of bankruptcy prediction models. Socio-Economic Planning Sciences, 2019. 67: p. 34-42.

24.Thirumalai, C., M. Vignesh, and R. Balaji. Data analysis using box and whisker plot for lung cancer. in 2017 Innovations in Power and Advanced Computing Technologies (i-PACT). 2017.

25.Kim, Y. and J. Heer. Assessing effects of task and data distribution on the effectiveness of visual encodings. in Computer Graphics Forum. 2018. Wiley Online Library.

26.Makowski, D., et al., Methods and algorithms for correlation analysis in R. Journal of Open Source Software, 2020. 5(51): p. 2306.

27.Yang, X., et al., A survey on canonical correlation analysis. IEEE Transactions on Knowledge and Data Engineering, 2019.

28. Timse Series Split. 2020 [cited 2020 25-08]; Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html.

6 Opti Price-An Optimal Shipping Price Calculation Model

6.1 Introduction:

The shipping industry is fairly volatile pertaining to shipment pricing. To handle this price volatility, two types of pricing strategies are employed in the shipping sector, contract market pricing and spot pricing. The contract market offers a fixed shipment price for a known cargo task over a set period, with secured booking space in periods of high demand. The spot market has a fluctuating shipment price, where one can benefit from lower prices than contract rate shipment prices in the low season but face escalating shipment prices and less certainty of being able to secure a booking on a particular vessel in peak season, as space is reserved for contracted customers. However, both the pricing strategies currently followed have no relationship between current shipment demand and available shipping capacity and shipment prices are quoted based on predefined price lists.

In this chapter, we address this research gap of optimal spot shipment price calculation based on current shipment demand and available shipping capacity. To do so, we have developed an optimal pricing model named ‘Opti Price’ that utilizes historical data to calculate spot pricing for container shipments. The proposed model is capable of calculating shipment spot prices based on shipment demand and capacity. Data from various sources was gathered to generate a shipping data set for three years (i.e., from 2016 until 2018). Regression and correlation analysis are used to quantify research outcomes. Results have shown that the proposed model significantly increases the correlation between shipment price and shipment demand from 0.33 to 0.88 and available capacity from -0.12 to 0.35 respectively.

This chapter is organized as follows. In Section 6.2, we describe the research settings. In section 6.3, the relationship modeling between key variables are explained. Section 6.4 explains the mathematic behind the model followed by results and evaluation in section 6.5.The summary of chapter is presented in section 6.6.

6.2 Research Settings:

In order to find an optimal spot shipment price based on current shipment demand and available shipping capacity for a cargo container, we consider the price-setting process as explained in Section 1.2 (also see Fig 1.1 for further reference).

In order to do so, we have limited the scope of this research for the Asia Oceania trade lane in the Australian shipping industry. Additionally, we do not consider any competition between shipping lines and carriers. Moreover, the shipping prices must lie between the maximum and minimum threshold already set by the Australian shipping industry carriers. This is because clients are not willing to pay shipment prices above a particular value [1]. Hence, we aim to develop a pricing model that will keep the minimum and maximum thresholds in the loop and will provide a shipment price that can maximize the profit within those thresholds. The first step is data collection and data cleansing followed by relationship modeling of the key variables (i.e. , shipment demand, capacity, and price). Finally, the design of the optimal spot pricing model is implemented.

In the next section, we describe the relationship modeling between the shipment price, current shipment demand and available shipping capacity. We also present the change in the relationship between these vital variables after applying the Opti Price model over the dataset.

6.3 Relationship Modeling:

In order to model the relationship between shipment demand, available shipping capacity (supply) and pricing in the Australian shipping industry for the Asia-Oceania trade lane, statistical analysis is performed. The results from the preliminary exploratory data analysis (EDA) depict a 'disconnect between suppliers, shipment demand, and pricing'. There exists no linear relationship between available shipping capacity (supply), shipment demand, and pricing [2]. This is further augmented using density plots (see Fig 6.1 (a) and 6.1(b)). From Fig 6.1 (a), it is evident that the relation between shipment price and shipment demand is not obvious. This means that an increase in shipment demand does not necessarily increase the shipment price. In addition to this, the relationship between available shipping capacity (supply) and shipment price also appears to be non-binding. The density distribution between shipment demand and shipment price is bimodal showing that firstly, shipment prices increase with shipment demand but do not follow the same trend and start falling with an increase in shipment demand. On the other hand, the density distribution between available shipping capacity (supply) and shipment price is unimodal and a bit left-skewed. As available shipping capacity (supply) increases, shipment prices do not seem to have any dependency on it. Moreover, the left skewness is not very prominent. It can be concluded that shipping prices are more jumbled in the center of the distribution showing no evident positive relationship between the variables under observation.

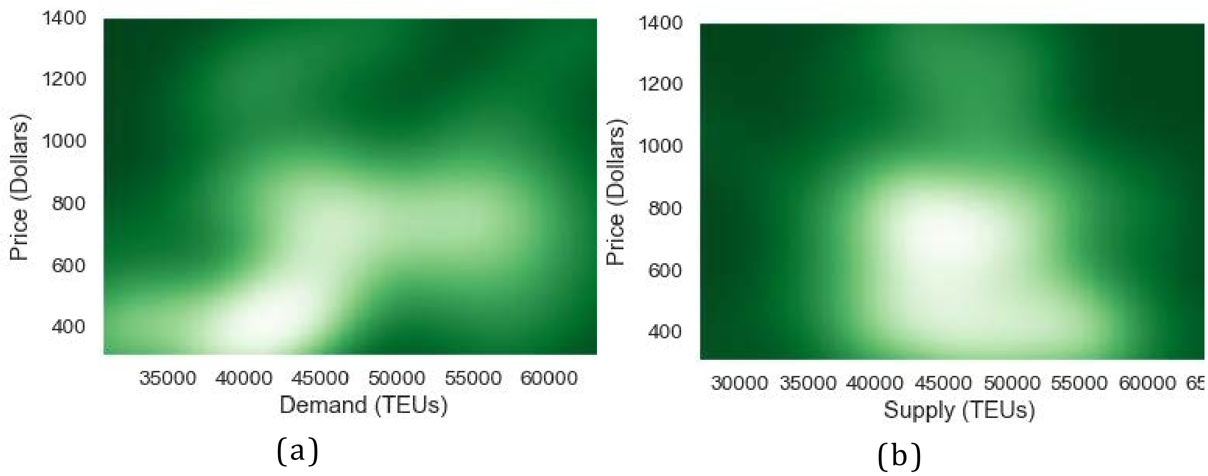


Fig. 6-1 Density Plot (a) Demand and Price (b) Available shipping capacity (supply) and Price

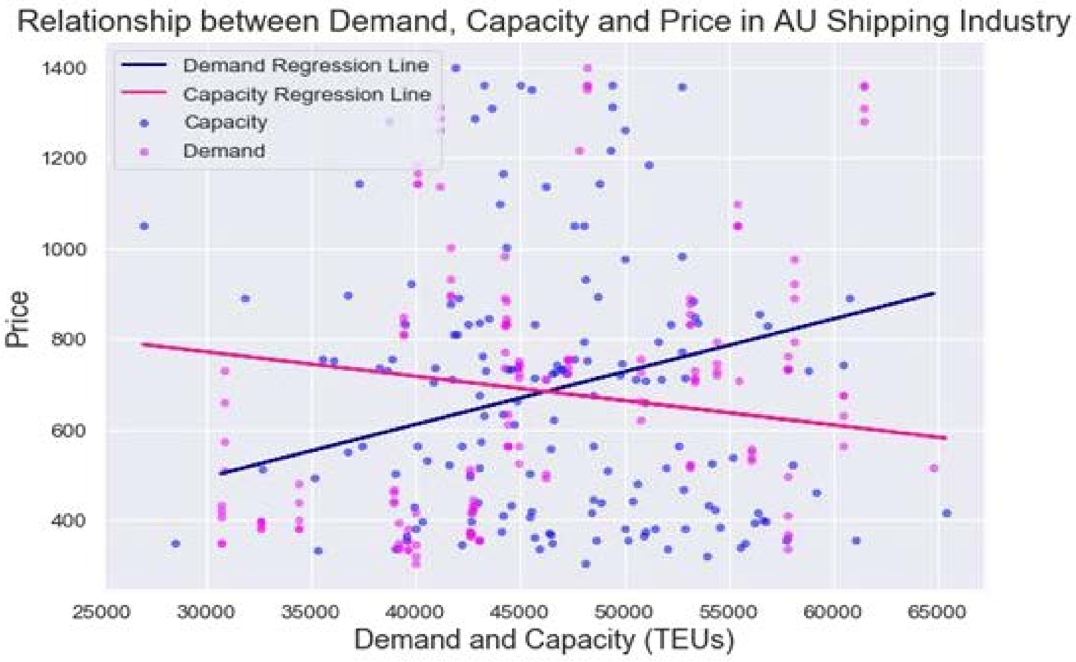


Fig. 6-2 Scatter plot and Regression analysis Price vs Demand and Available shipping capacity (supply)

To further the analysis, we have performed a regression analysis on the dataset (see Fig 6.2). From Fig 6.2, it can be seen that there exists a correlation between shipment demand and shipment price, but the relationship is not obvious. However, there is no relationship between capacity and shipment price. Thus, it can be shown from the above-presented analysis that shipment prices are not steered by cargo shipment demand and available shipping capacity in the Australian shipping industry. An increase in shipment demand does not necessarily increase the cargo pricing. Moreover, there is no relationship between shipment demand and capacity. However, capacity and shipment prices are bound in different types of relationships. Thus, from this Exploratory Data Analysis (EDA), it can be concluded that the shipment demand, available shipment capacity, and cargo shipment price setting are not interrelated in the container shipping industry. The results from EDA depict the natural consequence of an uninformed market with material opportunity cost to shipping lines and their customers, which is what our research will address. Thus, with this study, visibility into the current operations of the container shipping available shipping capacity (supply) chain is provided and the long-held objective for industry stakeholders is proven statistically. In order to solve this long-standing issue, we have designed an optimal shipment price calculation formula using historic data.

In the next section, we explain the details of mathematics and logic behind the Opti Price.

6.4 Optimal Price Calculation:

In order to calculate the opportunity cost for shipping companies based on shipment demand and available shipment capacity, we analyzed the historical data. There are three predefined propositions for pricing decisions.

Proposition 1

Shipment demand is greater than the available shipping capacity (supply). In this case, ideally, the increase in shipment demand will cause an increase in shipment price, provided the shipping capacity remains the same. Hence, it can be written as equation (6.1) and (6.2). Thus, cargo shipment demand can be written as equation (6.3).

$$D > S \tag{6.1}$$

$$D \propto P \tag{6.2}$$

$$D = \frac{1}{m_{max}} P \tag{6.3}$$

In equation (3), m is a constant of proportionality. The values of m vary from the maximum shipment price limit per TEU to minimum shipment price per TEU and are derived from historic data. The constant m can be calculated using equation (6.4). The value of m_{max} and m_{min} is derived from historic data (see Equation (6,7), (6.8) and (6.9)).

$$m_{(\min)} \leq m \leq m_{(\max)} \tag{6.4}$$

Proposition 2

Shipment demand is less than available capacity. Then the increasing shipment demand should not result in increasing the price to the maximum level. However, average pricing must be quoted so carriers avoid quoting shipment prices which are below average. This would help them earn reasonable revenue to keep their business profitable. This equation (6.3) can be written as

$$\mathrm{D} \leq S \tag{6.5}$$

$$D = \frac{1}{m_{av}} P \tag{6.6}$$

Calculating Constant of proportionality

The opportunity cost calculation is highly correlated with the constant of proportionality used in equations (7) and (8). Hence, it is very important to calculate it precisely. The value of m is driven by historic shipment demand and available shipping capacity (supply) data. Equations (6.7), (6.8) and (6.9) are used in this research to calculate the respective values of m_{max}, m_{avg} and m_{min}.

$$\mathrm{m_{max}} = \frac{\mathrm{P_{max}}}{\mathrm{Demand(P_{max})}} \tag{6.7}$$

$$m_{avg} = \frac{P_{avg}}{Demand(P_{avg})} \tag{6.8}$$

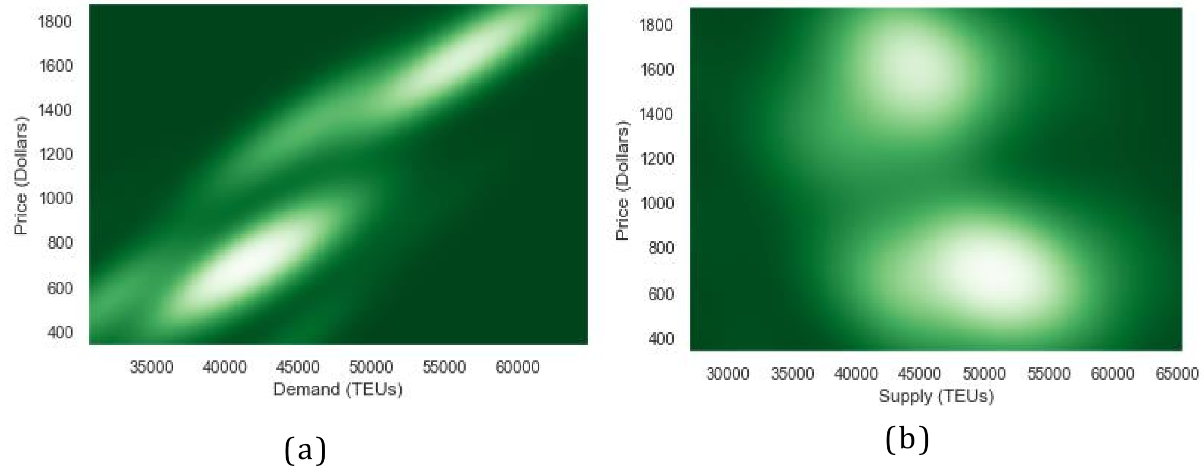
$$m_{min} = \frac{P_{min}}{Demand(P_{min})} \tag{6.9}$$

where, P_{max} and P_{min} is the maximum and minimum threshold of the shipment price the shipping companies can quote to the customers. P_{avg} , is the average shipment price quoted and is calculated from the pricing presented in the dataset under observation. These shipment prices are fixed and spot pricing must lie between these price limits. Thus, the opportunity cost model must provide the cost between these set thresholds. Thus, the formal model for spot pricing opportunity cost is given in equation (6.10) below.

$$p(D) = \begin{cases} Dm_{max}, & D > S \\ Dm_{avg}, & D \leq S \end{cases} \tag{6.10}$$

6.5 Results and Evaluations:

In this section, we evaluate the Opti Price proposed in this chapter. Using equation (15) for optimal spot shipment price calculation, we have calculated shipment prices based on shipment demand and available shipping capacity. In order to determine the relationship between the variables, data analytics is performed (similar to Section 4.2). From the density plots, it can be seen that the correlation between the demand and shipment price is improved and has become more positive in comparison to previous values (see Fig 6.3 (a) in comparison with Fig 6.1(a)). The same is true for capacity and shipment price (see Figure 6.3 (b) in comparison with Fig 6.1(b)). Moreover, from the regression analysis (see Fig 6.4), it can be seen that the regression line between shipment demand and shipment price becomes positive in comparison to Fig 6.2 (according to current pricing method). However, the regression line between available shipping capacity and shipment price remains negative but seclusion is achieved.



(a)

(b)

Fig. 6-3 Density Plot (a) Demand and Price (b) Available shipping capacity (supply) and Price

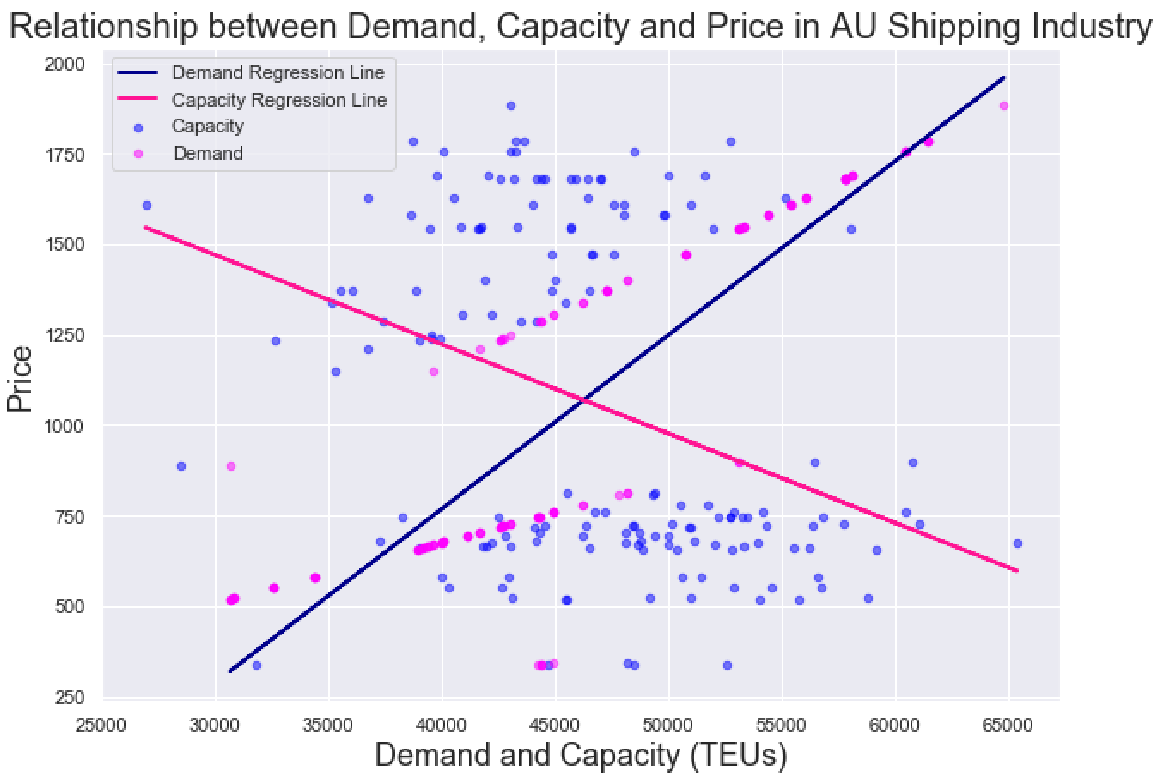


Fig. 6-4 Scatter plot and Regression analysis Optimal Price vs Demand and Available shipping capacity (supply)

The graph below (see Fig.6.5) shows the optimal spot pricing and current pricing with respect to cargo container shipment demand. It is evident that the designed model is able to calculate spot shipment prices based on cargo shipment demand and can help carriers increase their businesses. However, the current pricing is quite low and does no good for their businesses. To further the results concluded from the aforementioned graphs, we also performed correlation analysis on both datasets.

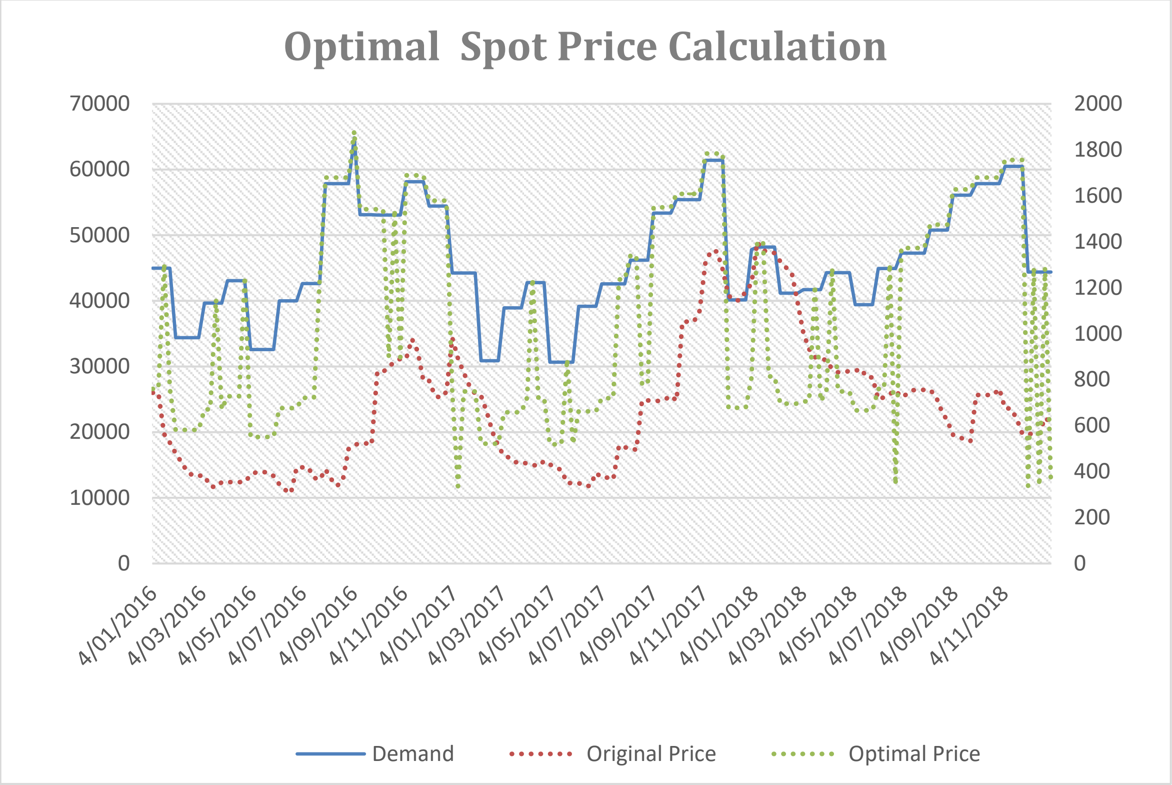


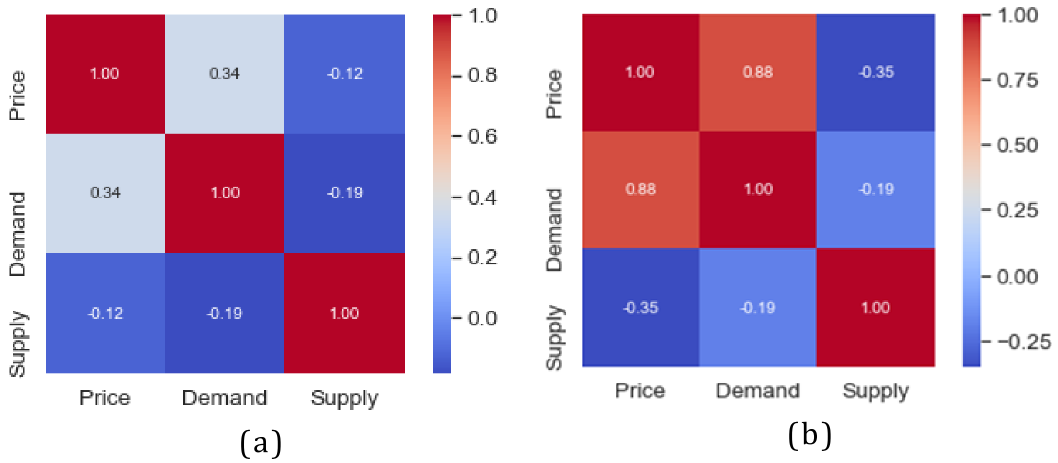
Fig. 6-5 Shipment optimal pricing (a) Traditional price calculation (b) Proposed optimal price calculation.

In Fig 6.6, we quantify our correlation results. In traditional pricing, there is no relationship between shipment demand, available shipping capacity (shown as supply), and spot pricing. The correlation between shipment demand and shipment price is 0.34, showing less correlation between the two variables. Furthermore, the correlation between capacity and pricing is -0.12, depicting the negative relationship between capacity and shipment price. On the other hand, the designed optimal pricing formula designed to calculate shipment spot pricing based on shipment demand and available capacity increased the shipment demand-shipment price correlation up to 0.88 and that of capacity and shipment price is improved to -0.35 respectively. Thus, the proposed model is capable of generating a positive relationship between variables that exist inherently in datasets, but which have been unused due to a lack of visibility in the shipping operation.

Hence, the proposed algorithmic method to calculate shipment spot pricing based on current market demand and available shipment capacity has proved to be optimal, based on four statistical matrices:

a) Density plots to explore relationship between spot prices, demand and capacity (see Fig 6.1 and 6.3);
b) Scatter plots and regression analysis (see Fig 6.2 and 6.4) ;
c) Line graph visualization for comparison of current spot prices and optimal spot prices with respect to shipment current demand (see Fig 6.5);
d) In the end we have quantified our research with a correlation heat map (see Fig 6.6). Results from all of the metrics indicate that the proposed model is optimal for calculating shipment spot prices based on current shipment demand and available capacity.

In the next section, we conclude this chapter.



(a)

(b)

Fig. 6-6 Correlation heat maps (a) current spot pricing (b) optimal spot pricing

6.6 Summary:

The shipping industry is fairly volatile pertaining to shipment pricing. To handle this volatility, two types of pricing strategies are employed in the shipping sector, contract market pricing and spot pricing. The contract market offers a fixed shipment price for a known cargo task over a set period, with secured booking space in periods of high demand. The spot market has a fluctuating shipment price, where one can benefit from lower prices than contract rate shipment prices in the low season but face escalating shipment prices and less certainty of being able to secure a booking on a particular vessel in peak season, as space is reserved for contracted customers. However, both the pricing strategies followed have no relationship between current shipment demand and available shipping capacity and shipment prices are quoted based on predefined price lists (hard copy). In this chapter, we addressed the issue of optimal spot shipment price calculation based on current shipment demand and available shipping capacity. To do so, we have developed a model that utilizes historical data to calculate spot pricing for container shipments. The proposed model is capable of calculating shipment spot prices based on shipment demand and capacity. Data from various sources was gathered to generate a shipping data set for three years (i.e., from 2016 until 2018). Regression and correlation analysis are used to quantify research outcomes. Results have shown that the proposed model significantly increases the correlation between shipment price and shipment demand from 0.33 to 0.88 and available capacity from -0.12 to 0.35 respectively.

6.7 References:

1.Time Series Split. 2020 [cited 2020 25-08]; Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html.

7 Container shipment Demand Forecasting for the Australian shipping industry

7.1 Introduction:

Demand forecasting plays a pivotal role in making informed business decisions by predicting future sales using historical data. Traditionally, demand forecasting has been widely used in the management of production, staffing, and warehousing for sales and marketing data.[1] However, the use of demand forecasting has received little attention from the container shipping industry. Improved visibility into the demand for container shipment has been a long-held objective of industry stakeholders [2]. This thesis addresses this shortcoming of predicting both short-term and long-term shipment demand for the Australian container shipping industry.

In this study, we compare three forecasting models, namely seasonal auto-regressive integrated moving average (SARIMA), Seasonal Holt-Winters' and Facebook Prophet, to find the optimal model for short-term and long-term import demand forecasting in the Australian shipping industry. Demand data from three years, i.e., 2016- 2018, is used for the Asia - Oceania trade lane. Mean absolute percentage error (MAPE), and root mean squared error (RMSE) is observed for model evaluation. The experiment results show that Prophet outperforms other models in its comparison by achieving lower RMSE and MAPE. This chapter is organized as follows; research design explaining research settings, data sourcing and cleansing, missing value handling, and test-train split is explained in section 7.2. In section 7.3, the selected time series models are explained and their application on shipment demand dataset is discussed. Section 7.4, explains the experimental results.

7.2 Research Design:

Although researchers are actively working for forecasting demand for various domains such as sales and electricity from the last few decades, the problem of demand forecasting for container shipment has still not been addressed. To our knowledge, there are no prominent studies that can assist in forecasting container shipment demand. To address the gap, as mentioned earlier we have performed a comparative study between three state-of-the-art time series models, namely SARIMA and Facebook Prophet, to forecast container shipment demand for both short and long-term in the Australian container shipping industry. Three years of historical demand data, i.e. 2016-2018 are collected from five international Australian ports for the Asia Oceania trade lane [3] [4] [5] [6] [7]. Root mean squared error (RMSE) and mean average percentage error (MAPE) are used as evaluation metrics to evaluate the different models and select the optimal one.

7.2.1 Data Sourcing and Cleansing:

This section explains the data sourcing and cleansing methods to make the data suitable for short-term demand forecasting. To our knowledge, there exist no data set that can provide insights into the shipment demand in the Australian Shipping industry. We started by collecting real-time shipment demand datasets from five of the international ports to form a consolidated demand dataset. These ports include Port of Melbourne [3], Fremantle Port [4], Flinders Port [6], Port of Brisbane [5] and Port Botany[7] as shown in Figure 7.1. The detailed methodology for shipment demand data can be seen from [8]

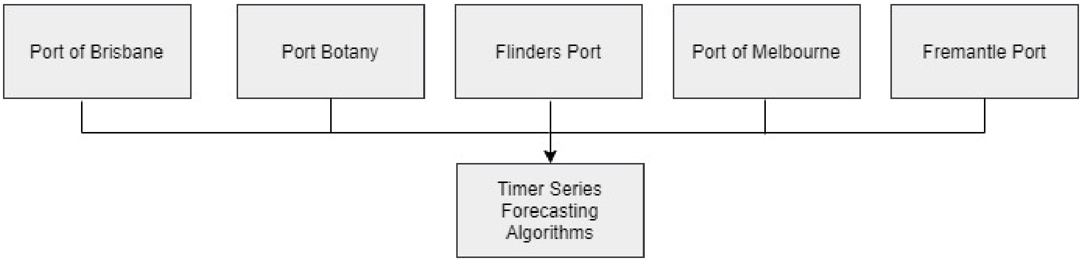


Fig. 7-1 Data Sources for Shipment Demand Dataset

Figure 7.2 explains the data cleansing process of the sources data to get a trade lane specific dataset ready to be used in machine learning algorithms. The initial step of data cleansing is to select the time horizon. Historic data from 2016-2018 is extracted. This selected dataset consists of data from all the trade lanes operating from Australia (import and export). Since in this research study, we aim to target only Asia-Oceania trade, trade lane-specific data to and from Asia-Oceania trade lane is segregated.

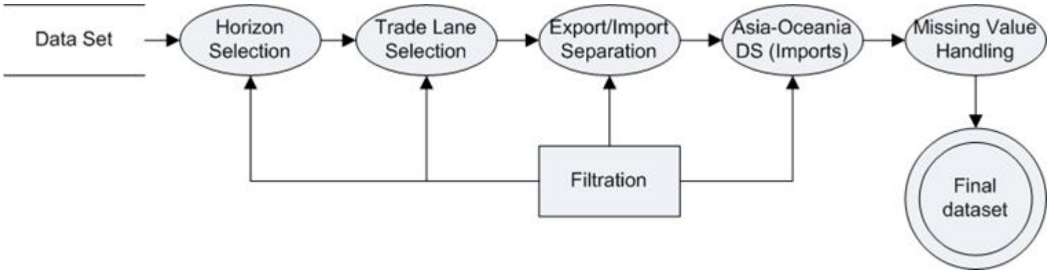


Fig. 7-2 Overview if the Data Cleansing Process

The trade lane data set acquired so far is the combination of full and empty containers coming in and going out from the Australian ports over the Asia-Oceania trade lane. Full containers are the number of containers (measure in TEUs) that are coming in and out filled with goods. In contrast, the empty containers are the ones which are being emptied on the way before entering into Australian waters or are moving back without goods filled in them. However, in the total shipment demand, both the filled and empty containers are equally important. The feature in the dataset is shown in

Table 7.1. In the table 'date' refers to the first day of the week and provides insight for the same whole week. Hence, for both imports and exports, the dataset contains the total number of filled incoming (inbound) and outgoing (outbound) containers as well as empty containers.

Table 7-1 List of Demand Dataset Features

Region	Feature Name	Description.
Asia Oceania	Imports	Number of inbound containers
	Full	Total Inbound container (filled)
	Empty	Total Inbound container(Empty)
	Date	Starting day of week
	Exports	Number of outbound containers
	Full	Total outbound container (filled)
	Empty	Total outbound container(Empty)
	Date	Starting day of week

Since the scope of this study is demand forecasting for Asia Oceania trade lane imports only, we have filtered import and export in the next step. Table 7.2 shows the dataset features for Asia Oceania trade lane imports.

Table 7-2 Demand Dataset Feature for Asia Oceania Imports

Region	Feature Name	Description.
ASIA-Oceania (Imports)	Full	Total Inbound container (Filled)
	Empty	Total Inbound container (Empty)
	Date	Starting day of week

Total shipment import demand can be calculated by adding total incoming empty containers ($Empty_{Containers}$) and total incoming full containers ($Empty_{Containers}$). This can be expressed as in equation (7.1) is used.

$$Import\ Demand_{Total} = Empty_{Containers} + Full_{Containers} \tag{7.1}$$

7.2.2 Missing Value Handling:

Missing values are handled by filling with the calculated average demand from the previous and coming year's similar timestamps [9]. Figure 7.3 shows the container shipment demand (shown on the X-axis of Figure 8.3) for the Asia-Oceania trade lane (imports only) for the year 2016 through 2018 (presented on the Y-axis of Figure 8.3) and Figure 7.4 shows the data description of demand.

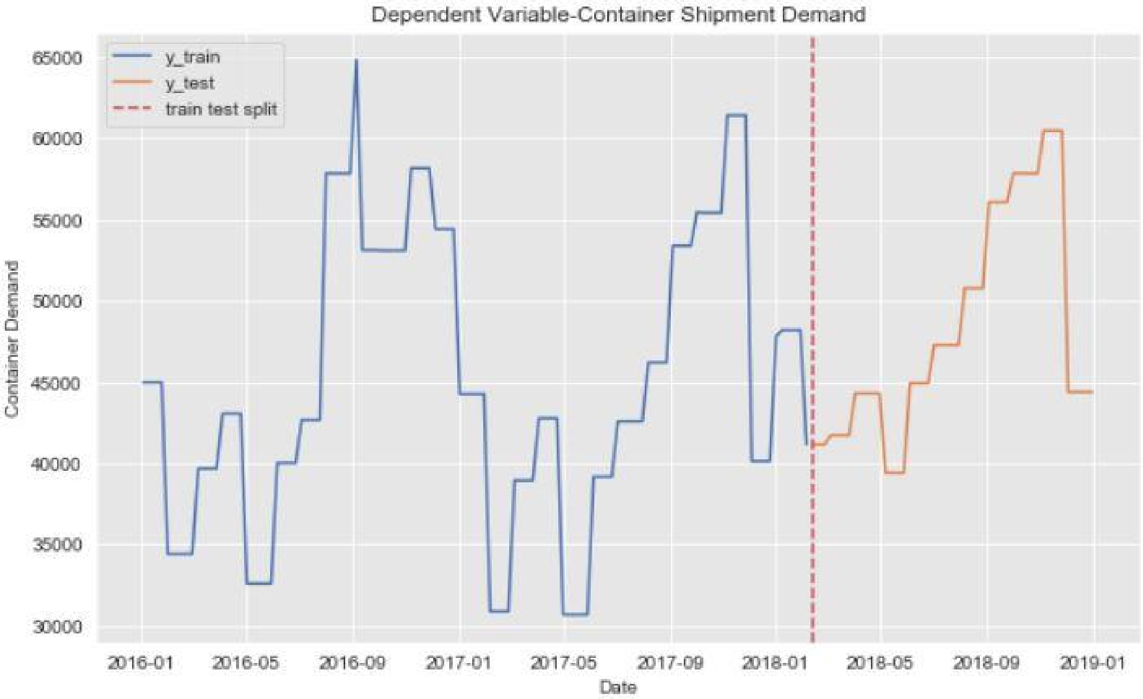
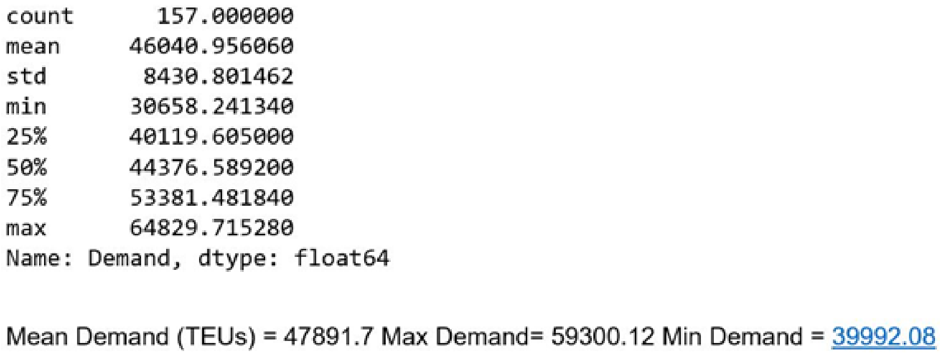


Fig. 7-3 Shipment Demand Dataset Visual



```
count      157.000000
mean     46040.956060
std       8430.801462
min      30658.241340
25%      40119.605000
50%      44376.589200
75%      53381.481840
max      64829.715280
Name: Demand, dtype: float64
```

Mean Demand (TEUs) = 47891.7 Max Demand= 59300.12 Min Demand = 39992.08

Fig. 7-4 Data description of demand dataset [8]

7.2.3 Test Train Split:

Once the missing values are filled, the data set is divided into two part. Train and test dataset (see figure 8.5). 70% of the data is used for training the model, the remaining 30% of the data is used for testing the model. In time series models, usual test train split does not work as the values are dependent on time [10]. Performing random partitioning can cause misleading results. Hence, we have selected a cut-off date that corresponds to approximately 70% of the dataset i.e., Feb 2018 (see vertical red line in figure 7.5) for training data in order to capture enough seasonality and trends of the time series under observation and have used the rest of the data as test data.

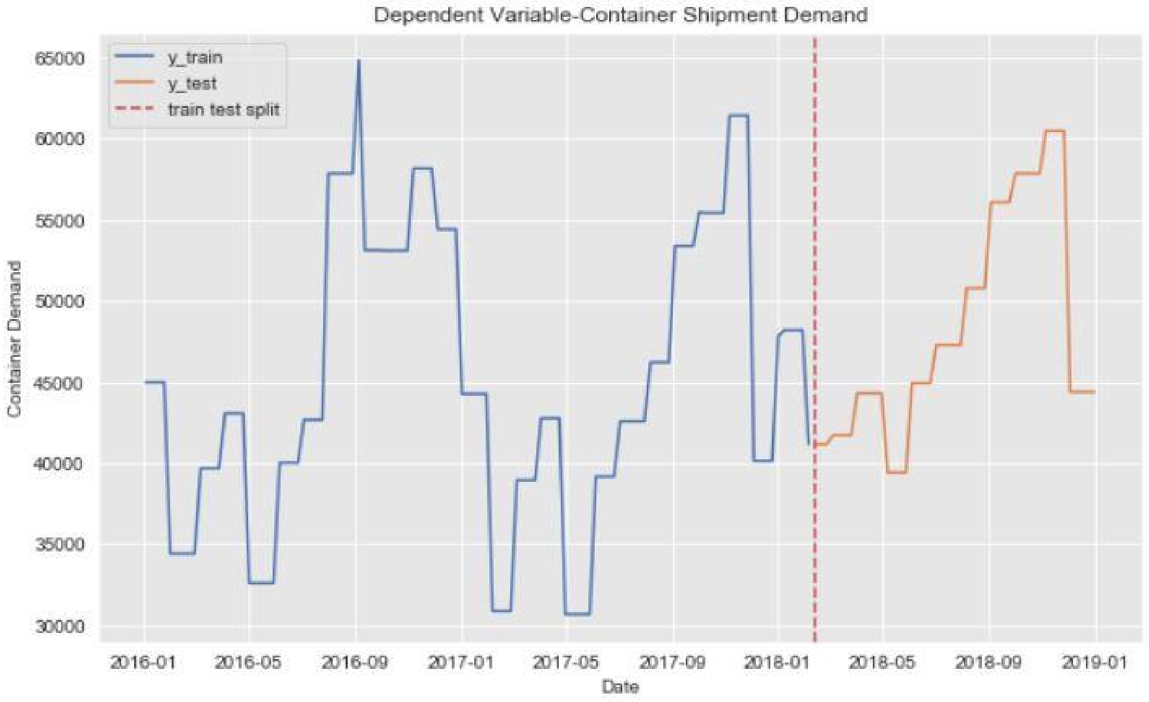


Fig. 7-5 Test train split of demand dataset (70-30)

7.3 Forecasting Models:

Three state-of-the-art time series models are selected to forecast shipment demand. These include SARIMA, Seasonal Holt-Winters' and Facebook Prophet. All of these models are capable of time series analysis using the seasonality present in the historical data.

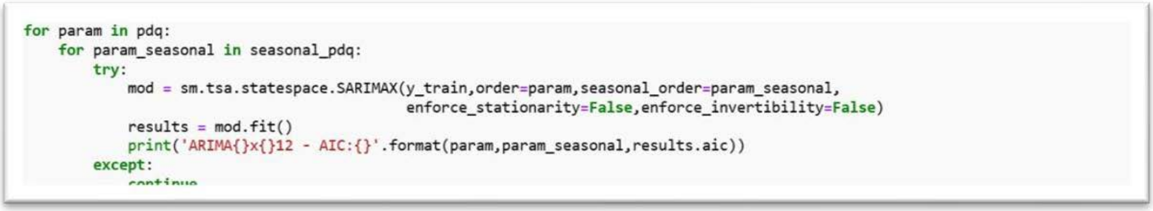
7.3.1 SEASONAL AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (SARIMA):

SARIMA is an extension of the autoregressive integrated moving average (ARIMA) [11], with an additional capability to handle seasonality in time series. Hence, we use SARIMA to solve our research problem as it adds three more parameters than ARIMA to cater for the seasonality in time series. Mathematically,

$$SARIMA = (p, d, q)(P, D, Q)^m \tag{7.2}$$

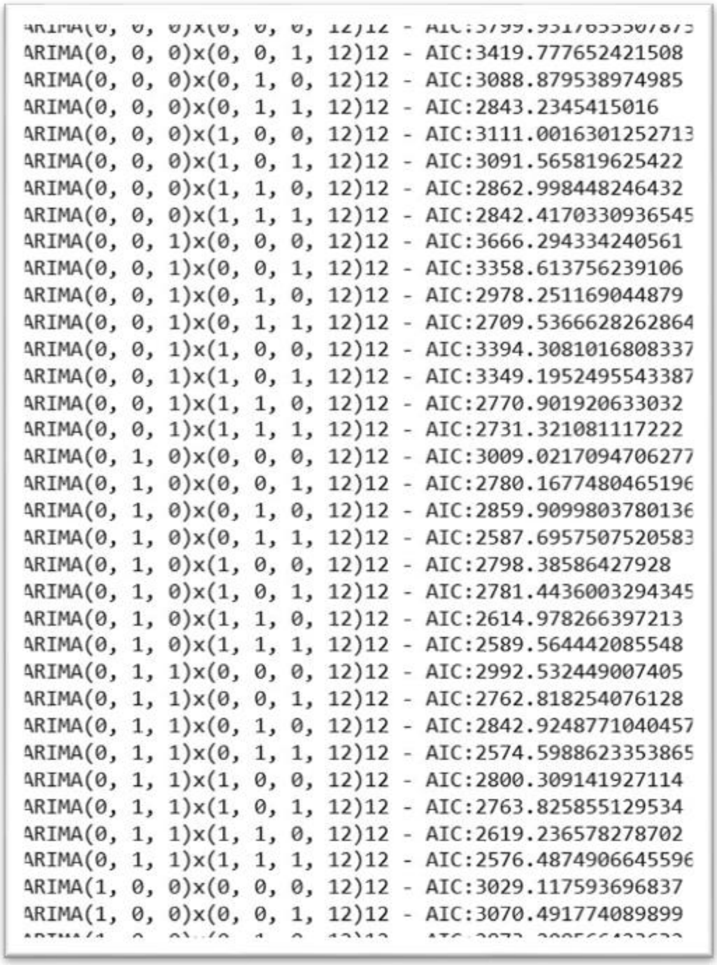
were m is the seasonality pattern, p is the number of lag observations extracted through the partial autocorrelation (PACF) plot. PACF may be considered as the partial correlation between the series, and its lag values which is difficult to explain by realizing their mutual correlation. d is the number of time difference that has to be calculated for making the time series stationary and q is the size of the moving window set by ACF value. The series with no trend has $d = 1$. However, if the series has a trend $d \geq 1$, P is a seasonal autoregressive order, and D is the seasonal difference order. If the series has a stable seasonal trend, then $D = 1$. If the seasonal pattern is unstable, then $D = 0$. Q is the seasonal moving average. Both P and Q are set by the ACF plots [11].

To find the optimal parameter, we use a grid search to determine the value for both (p, d, q) and (P, D, Q). Figure 7.6 shows the SARIMA parameters used in this research. To determine $(p, d, q)(P, D, Q)^m$ values, we have used the grid search method (see figure 7.6(a)). The grid search results are shown in figure 7.6 (b). Final parameters used for model fitting are shown in figure 7.6(c).



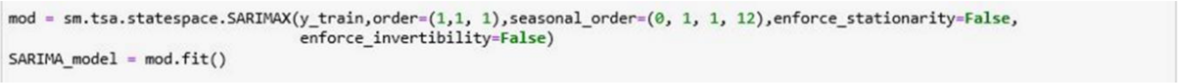
```
for param in pdq:
    for param_seasonal in seasonal_pdq:
        try:
            mod = sm.tsa.statespace.SARIMAX(y_train,order=param,seasonal_order=param_seasonal,
                                            enforce_stationarity=False,enforce_invertibility=False)
            results = mod.fit()
            print('ARIMA{}x{}12 - AIC:{}'.format(param,param_seasonal,results.aic))
        except:
            continue
```

(a)



```
ARIMA(0, 0, 0)x(0, 0, 0, 12)12 - AIC:3799.9317655507873
ARIMA(0, 0, 0)x(0, 0, 1, 12)12 - AIC:3419.777652421508
ARIMA(0, 0, 0)x(0, 1, 0, 12)12 - AIC:3088.879538974985
ARIMA(0, 0, 0)x(0, 1, 1, 12)12 - AIC:2843.2345415016
ARIMA(0, 0, 0)x(1, 0, 0, 12)12 - AIC:3111.0016301252713
ARIMA(0, 0, 0)x(1, 0, 1, 12)12 - AIC:3091.565819625422
ARIMA(0, 0, 0)x(1, 1, 0, 12)12 - AIC:2862.998448246432
ARIMA(0, 0, 0)x(1, 1, 1, 12)12 - AIC:2842.4170330936545
ARIMA(0, 0, 1)x(0, 0, 0, 12)12 - AIC:3666.294334240561
ARIMA(0, 0, 1)x(0, 0, 1, 12)12 - AIC:3358.613756239106
ARIMA(0, 0, 1)x(0, 1, 0, 12)12 - AIC:2978.251169044879
ARIMA(0, 0, 1)x(0, 1, 1, 12)12 - AIC:2709.5366628262864
ARIMA(0, 0, 1)x(1, 0, 0, 12)12 - AIC:3394.3081016808337
ARIMA(0, 0, 1)x(1, 0, 1, 12)12 - AIC:3349.1952495543387
ARIMA(0, 0, 1)x(1, 1, 0, 12)12 - AIC:2770.901920633032
ARIMA(0, 0, 1)x(1, 1, 1, 12)12 - AIC:2731.321081117222
ARIMA(0, 1, 0)x(0, 0, 0, 12)12 - AIC:3009.0217094706277
ARIMA(0, 1, 0)x(0, 0, 1, 12)12 - AIC:2780.1677480465196
ARIMA(0, 1, 0)x(0, 1, 0, 12)12 - AIC:2859.9099803780136
ARIMA(0, 1, 0)x(0, 1, 1, 12)12 - AIC:2587.6957507520583
ARIMA(0, 1, 0)x(1, 0, 0, 12)12 - AIC:2798.38586427928
ARIMA(0, 1, 0)x(1, 0, 1, 12)12 - AIC:2781.4436003294345
ARIMA(0, 1, 0)x(1, 1, 0, 12)12 - AIC:2614.978266397213
ARIMA(0, 1, 0)x(1, 1, 1, 12)12 - AIC:2589.564442085548
ARIMA(0, 1, 1)x(0, 0, 0, 12)12 - AIC:2992.532449007405
ARIMA(0, 1, 1)x(0, 0, 1, 12)12 - AIC:2762.818254076128
ARIMA(0, 1, 1)x(0, 1, 0, 12)12 - AIC:2842.9248771040457
ARIMA(0, 1, 1)x(0, 1, 1, 12)12 - AIC:2574.5988623353865
ARIMA(0, 1, 1)x(1, 0, 0, 12)12 - AIC:2800.309141927114
ARIMA(0, 1, 1)x(1, 0, 1, 12)12 - AIC:2763.825855129534
ARIMA(0, 1, 1)x(1, 1, 0, 12)12 - AIC:2619.236578278702
ARIMA(0, 1, 1)x(1, 1, 1, 12)12 - AIC:2576.4874906645596
ARIMA(1, 0, 0)x(0, 0, 0, 12)12 - AIC:3029.117593696837
ARIMA(1, 0, 0)x(0, 0, 1, 12)12 - AIC:3070.491774089899
```

(b)



```
mod = sm.tsa.statespace.SARIMAX(y_train,order=(1,1, 1),seasonal_order=(0, 1, 1, 12),enforce_stationarity=False,
                                enforce_invertibility=False)
SARIMA_model = mod.fit()
```

(c)

Fig. 7-6 SARIMA parameters used in research (a) Grid search (b) Grid search results snapshot (c) Fitted Model

7.3.2 SEASONAL HOLT WINTERS' METHOD:

Holt-Winters' methods are suitable for data with trends and seasonality [12]. Similar to SARIMA, it has two variants, Additive and multiplicative. Mathematically it can be written as equation (7.3), where, L_t is the level equation which depicts the weighted averaged among seasonal and non-seasonal forecasts. $\mathrm{kb_t}$, represents the trend of the data while $\mathrm{S_{t+k-s}}$ represents the seasonal patterns and can be calculated using equation (7.4), (7.5), and (7.6) respectively. The co-efficient α, β, and γ are smoothing factors and their values lie between 0 and 1.

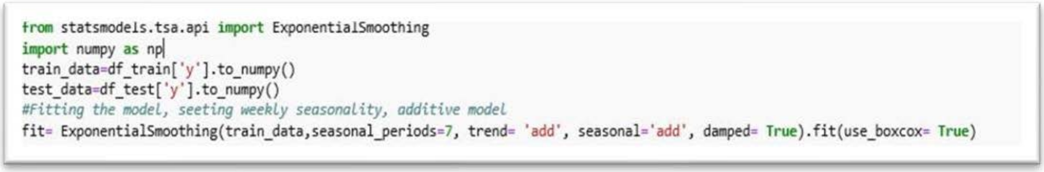
$$\mathrm{F_{t+k}} = \mathrm{L_t} + \mathrm{kb_t} + \mathrm{S_{t+k-s}} \tag{7.3}$$

$$L_t = \alpha(y_t - S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1}) \tag{7.4}$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \tag{7.5}$$

$$S_t = \gamma(y_t - L_t) + (1-\gamma)S_{t-s} \tag{7.6}$$

The parameters used to fit Holt’s Winter’s Seasonal methods are shown in Figure 7.7 below.



```
from statsmodels.tsa.api import ExponentialSmoothing
import numpy as np
train_data=df_train['y'].to_numpy()
test_data=df_test['y'].to_numpy()
#Fitting the model, seeting weekly seasonality, additive model
fit= ExponentialSmoothing(train_data,seasonal_periods=7, trend= 'add', seasonal='add', damped= True).fit(use_boxcox= True)
```

Fig. 7-7 Parameters for Holts’ Winter Model Training

7.3.3 FACEBOOK PROPHET:

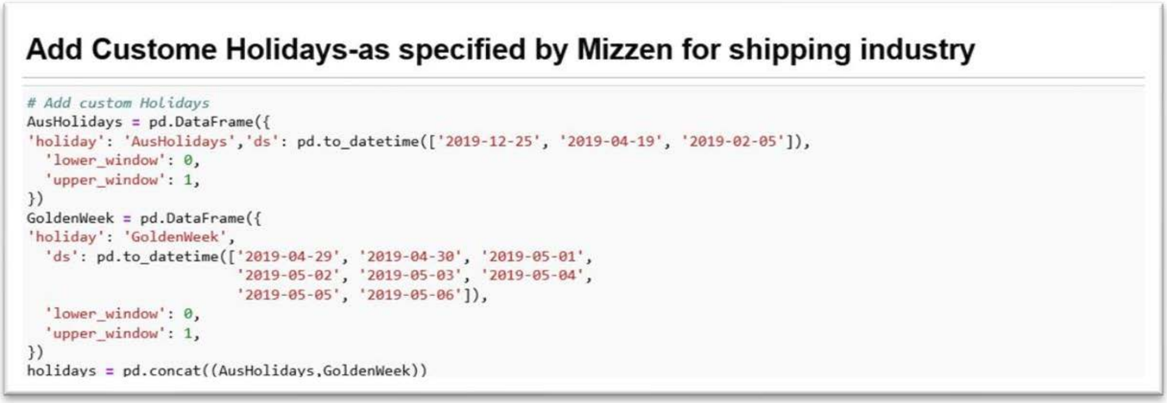
The Prophet is an open-source time series forecasting model [13]. It is designed to have in-built parameters that can be adjusted without going deep into the model's implementation details. At the core of the model, a decomposable time series model is running. These include trend, seasonality, and holidays. Mathematically, it can be written as

$$y(t) = g(t) + s(t) + h(t) + \in(t) \tag{7.7}$$

In equation (7.7),

- $g(t)$ is a piece-wise linear or logistic growth curve
- $s(t)$ is periodic change (e.g., weekly/yearly seasonality)
- $h(t)$ is user-provided holiday effect, and
- $\in (t)$ is error term accounting for unusual changes.

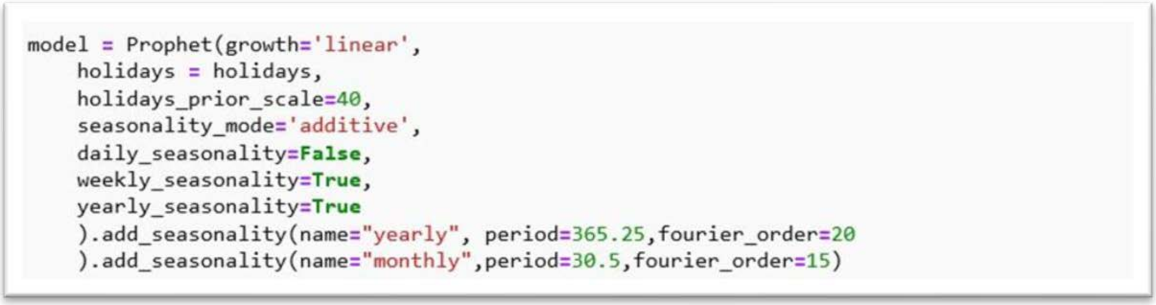
The role of the domain expert is significant in every phase of modelling. The domain expert can tweak the Fourier order and identify whether the details present in the data points are noise or a trend. Since Facebook Prophet treats forecasting as a curve-fitting problem, it is inherently robust to outlier and missing data. Custom defined holidays can also be used while fitting this model. This capability has not been provided by any of the existing algorithms yet. Hence, we define custom holidays as desired by our industry partner [14] (see Figure 7.8 (a)). We have used the same in our research for model fitting (see Figure 7.8 (b)).



Add Custome Holidays-as specified by Mizzen for shipping industry

```
# Add custom Holidays
AusHolidays = pd.DataFrame({
'holiday': 'AusHolidays','ds': pd.to_datetime(['2019-12-25', '2019-04-19', '2019-02-05']),
  'lower_window': 0,
  'upper_window': 1,
})
GoldenWeek = pd.DataFrame({
'holiday': 'GoldenWeek',
  'ds': pd.to_datetime(['2019-04-29', '2019-04-30', '2019-05-01',
                        '2019-05-02', '2019-05-03', '2019-05-04',
                        '2019-05-05', '2019-05-06']),
  'lower_window': 0,
  'upper_window': 1,
})
holidays = pd.concat((AusHolidays,GoldenWeek))
```

(a)



```
model = Prophet(growth='linear',
    holidays = holidays,
    holidays_prior_scale=40,
    seasonality_mode='additive',
    daily_seasonality=False,
    weekly_seasonality=True,
    yearly_seasonality=True
    ).add_seasonality(name="yearly", period=365.25,fourier_order=20
    ).add_seasonality(name="monthly",period=30.5,fourier_order=15)
```

(b)

Fig. 7-8 FACEBOOK PROPHET Parameters used in this research (a) custom defined holidays (b) model definition for container shipment demand forecasting.

7.4 Results:

In this section, we explain the results that are achieved by applying the selected models over the shipping datasets. The primary stage is to perform time series decomposition. Figure 7.9 shows the decomposition of demand data time series into different components, namely observed (green curve), trend (see blue curve) and cycle also known as seasonality (see red curve). From figure 7.9, it is evident that the shipment demand dataset is non-stationary.

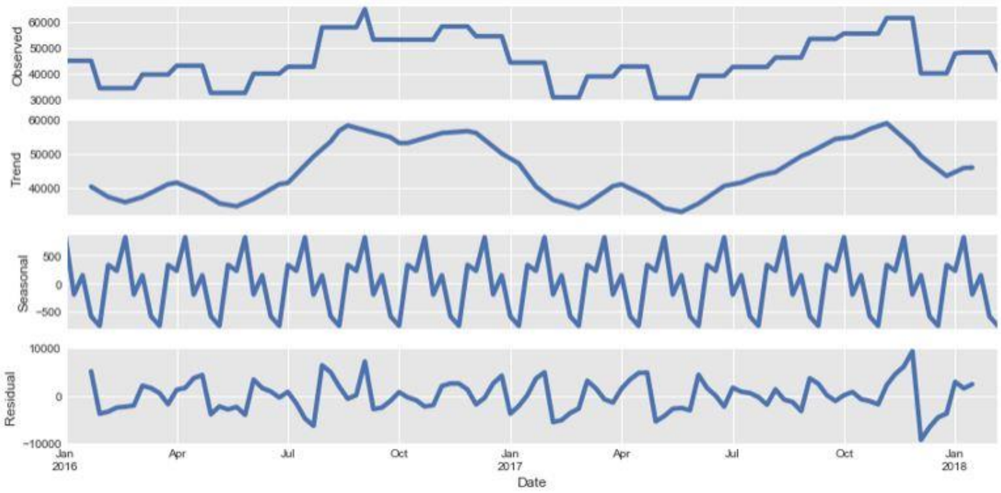
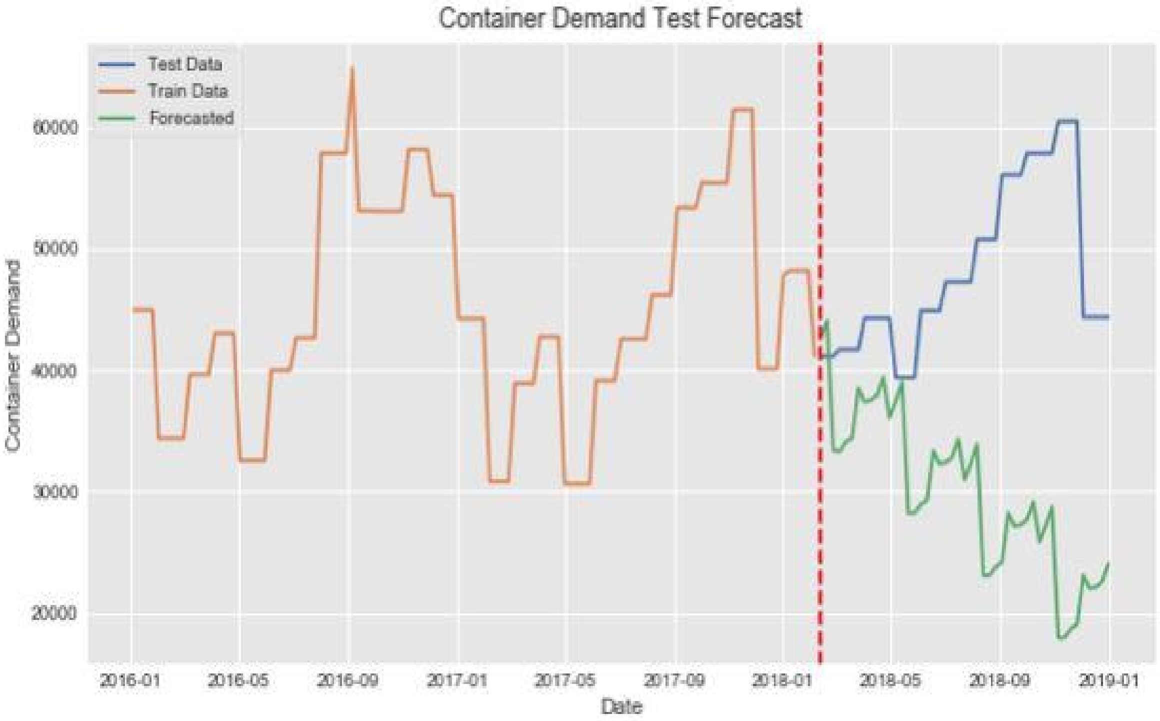
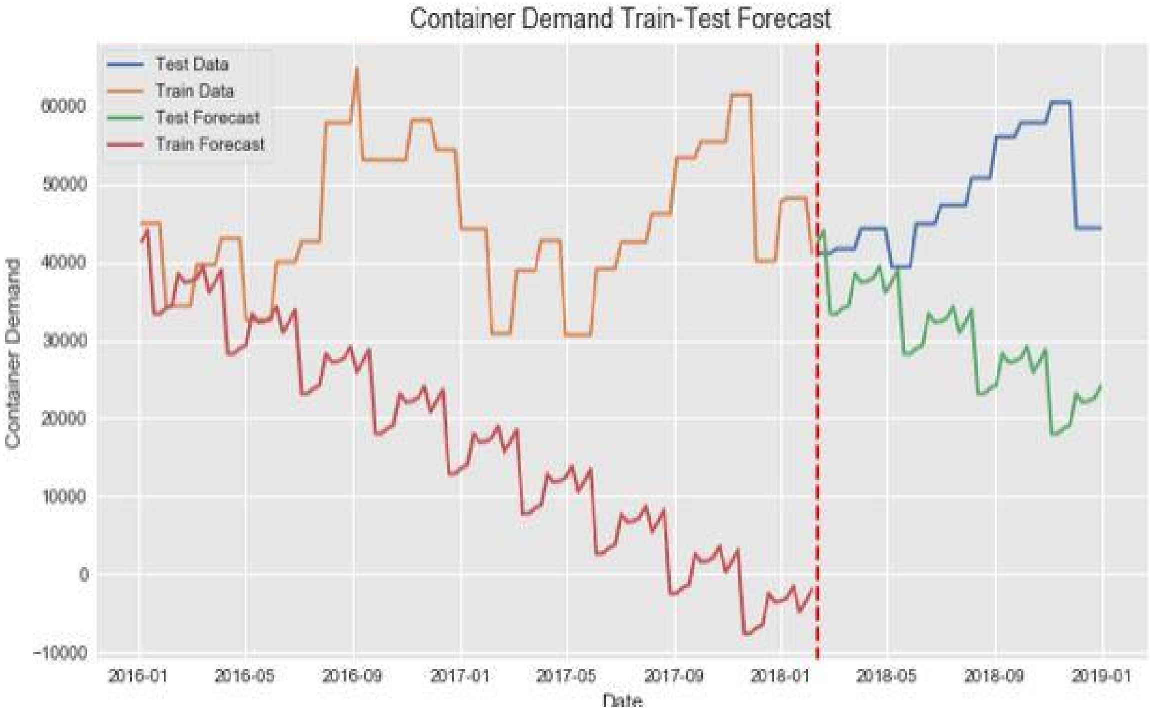


Fig. 7-9 Shipment Demand Dataset Decomposition

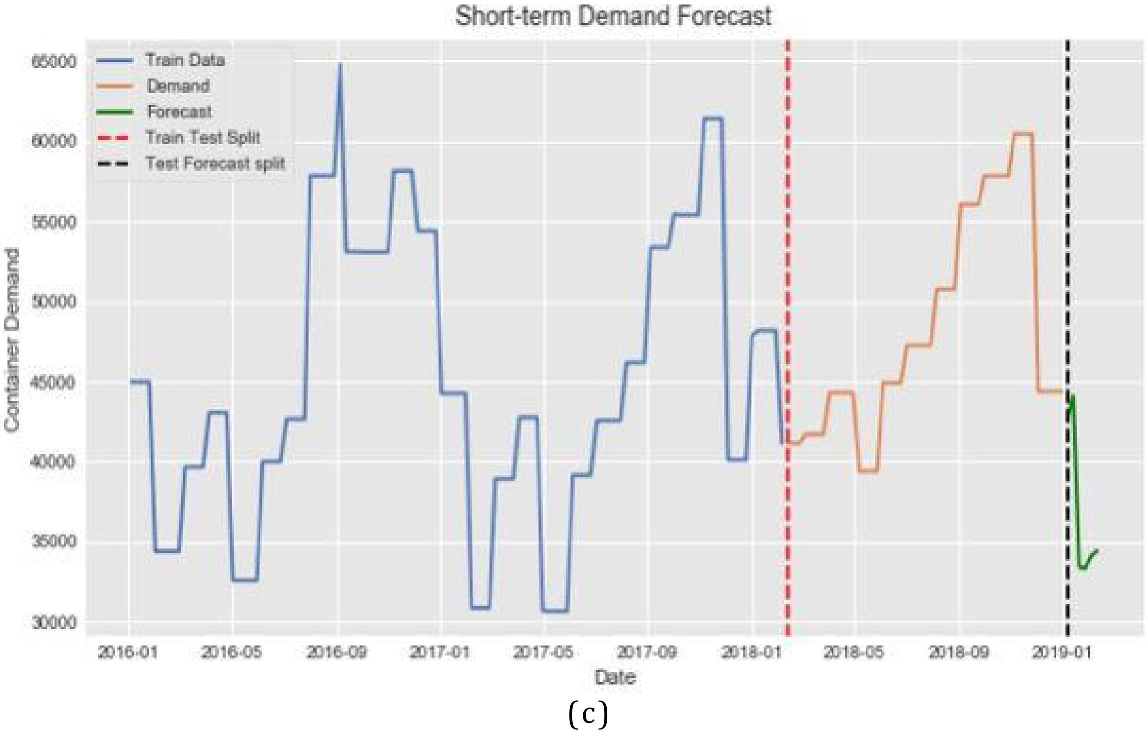
In the very first step, SARIMA is trained over the training data using parameters shown in figure 7.6 (c). Once the model is trained, the trained model is tested over the test data. As explained earlier, the configuration of ARIMA requires finding the optimal $(p,d,q)(P,D,Q)^m$ combination which has been achieved by performing grid search in our research (see figure 7.6(b)). The minimum AIC value achieved is used to fit the model. For our data set $(1,1,1)(0,1,1)^{12}$ provide the minimum AIC value, and is therefore used in this research. Once the model is tested on test data, demand forecasts are made for both short-term i.e., 6 weeks and long-term i.e. 52 weeks. Figure 7.10 (a) shows the test forecast provided by SARIMA. Train and test forecast can be seen in figure 7.10 (b). Short-term and long-term forecasts are shown in figure 7.10 (c) and (d) respectively.



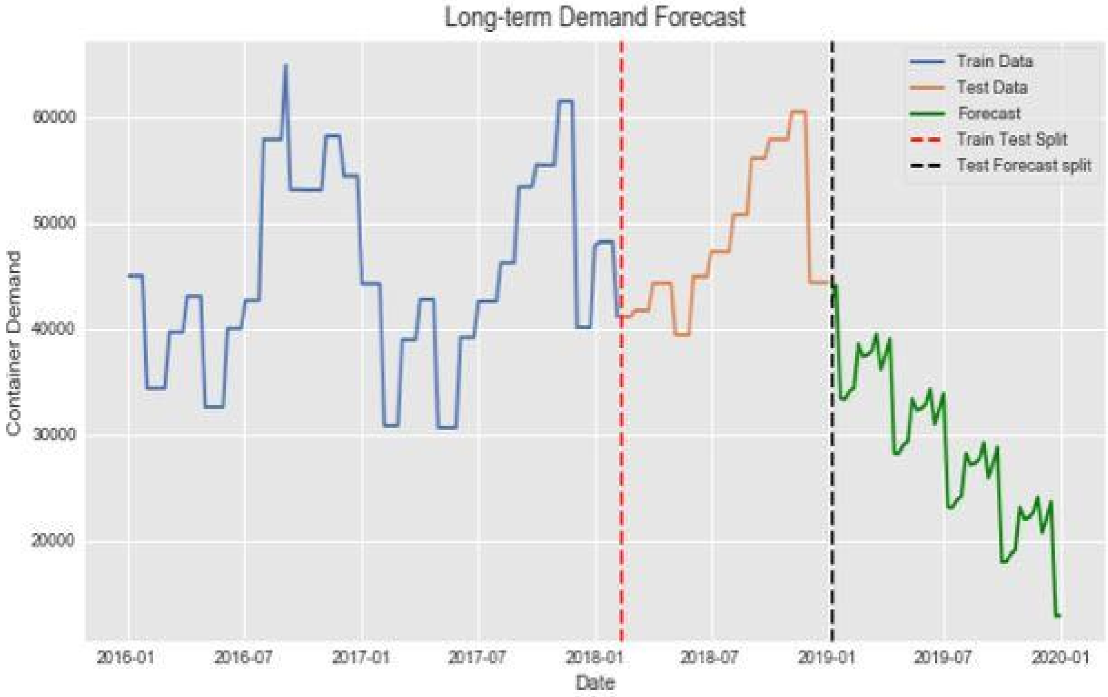
(a)


Container Demand Train-Test Forecast
Test Data
Train Data
Test Forecast
Train Forecast
Container Demand
60000
50000
40000
30000
20000
10000
0
-10000
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
2018-09
2019-01
Date

(b)


Short-term Demand Forecast
Train Data
Demand
Forecast
Train Test Split
Test Forecast split
Container Demand
65000
60000
55000
50000
45000
40000
35000
30000
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
2018-09
2019-01
Date

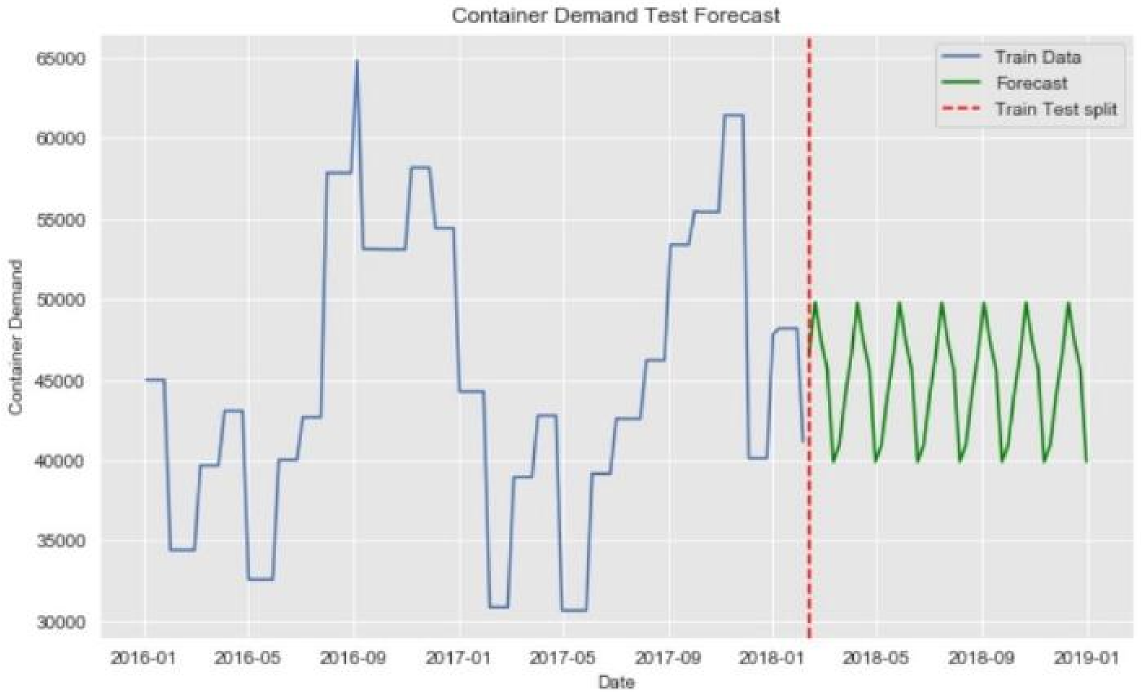
(c)



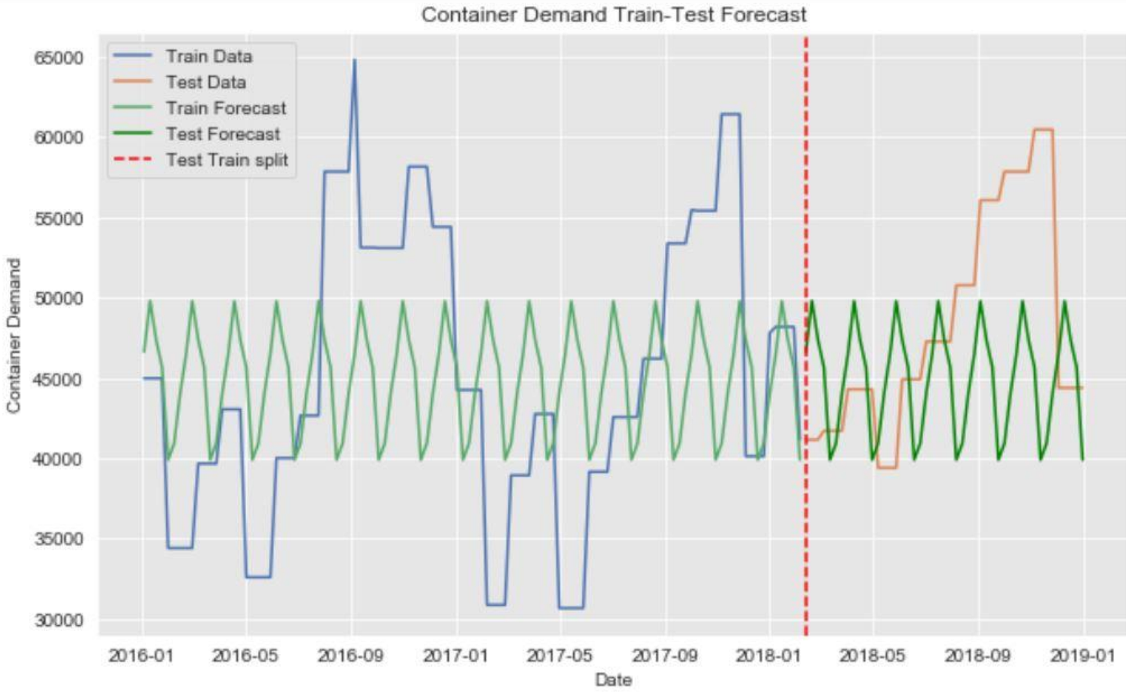
(d)

Fig. 7-10 Demand Forecast by SARIMA (a) Test forecast (b)Train test forecast (c) Short-term forecast (d) Long-term forecast

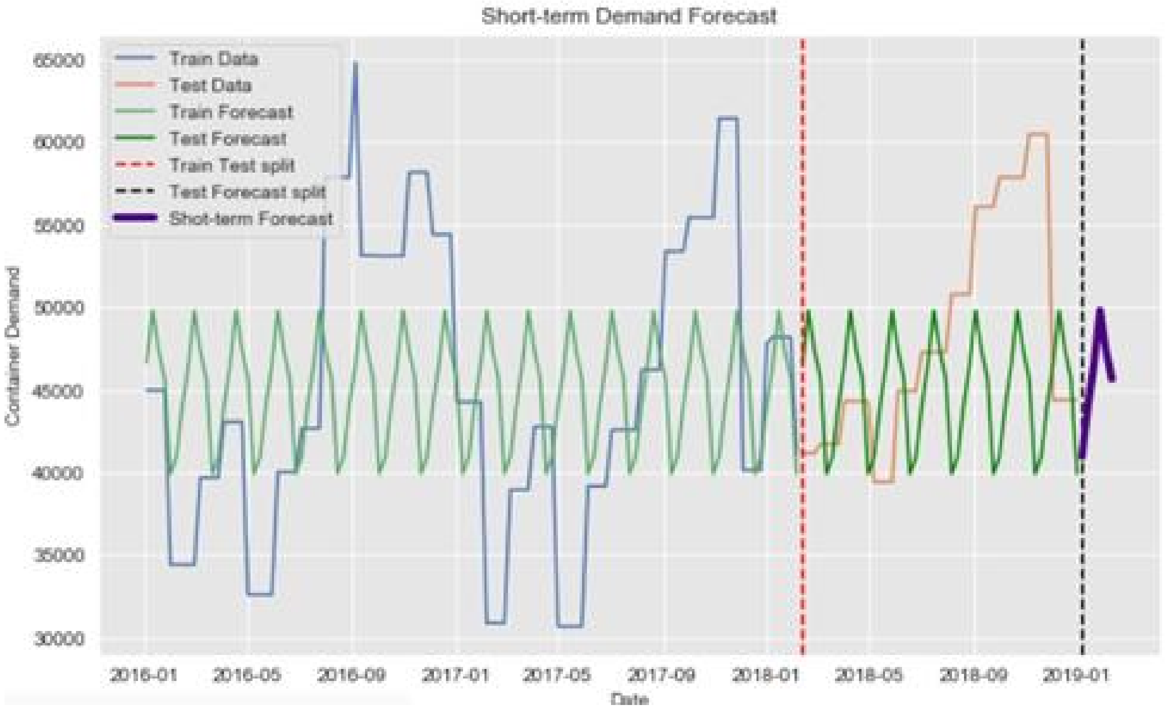
In the second step, the Seasonal Holt-Winters' model is trained. From the decomposition of time series, it is evident that the dataset is seasonal in nature and is suitable for additive seasonality. The parameters are shown in figure 7.7. Test forecast provided by Seasonal Holt-Winters' model is shown in figure 7.11(a). The forecast over both train and test data can be seen in figure 7.11(b). The short-term i.e., 6 weeks and long-term i.e., 52 weeks demand forecasted are shown in Figure 7.11(c) and (d) respectively.


Container Demand Test Forecast
Train Data
Forecast
Train Test split
Container Demand
65000
60000
55000
50000
45000
40000
35000
30000
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
2018-09
2019-01
Date

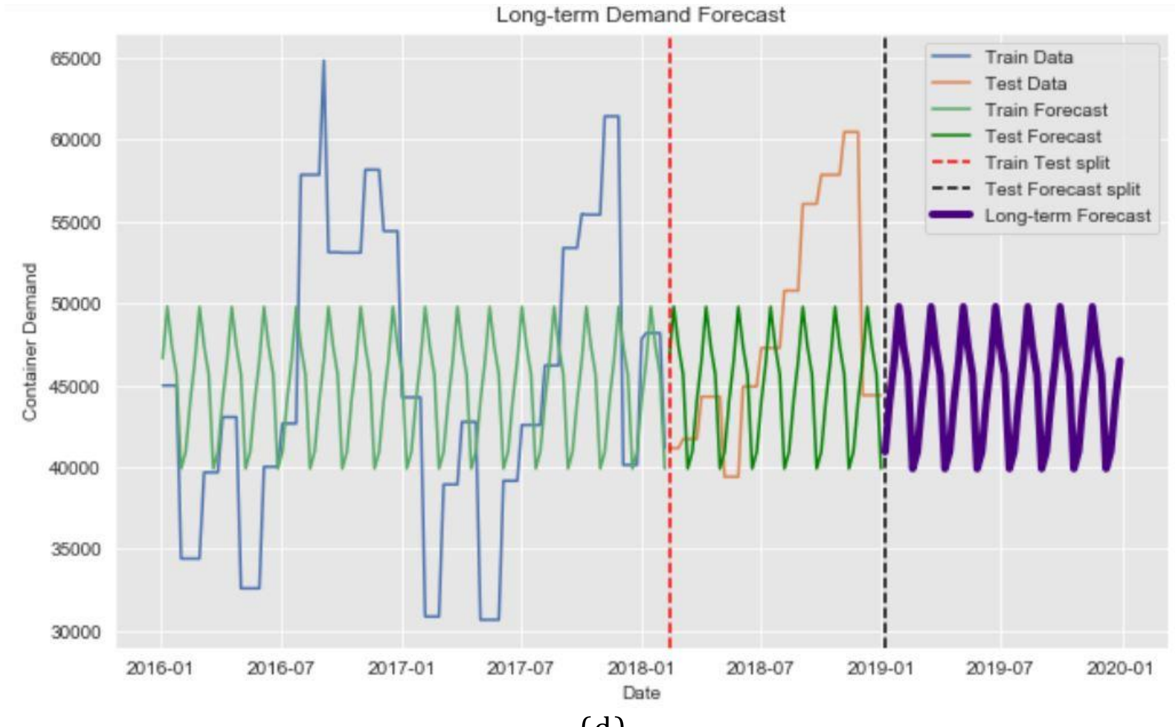
(a)


Container Demand Train-Test Forecast
Train Data
Test Data
Train Forecast
Test Forecast
Test Train split
Container Demand
65000
60000
55000
50000
45000
40000
35000
30000
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
2018-09
2019-01
Date

(b)


Short-term Demand Forecast
Train Data
Test Data
Train Forecast
Test Forecast
Train Test split
Test Forecast split
Shot-term Forecast
Container Demand
65000
60000
55000
50000
45000
40000
35000
30000
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
2018-09
2019-01
Date

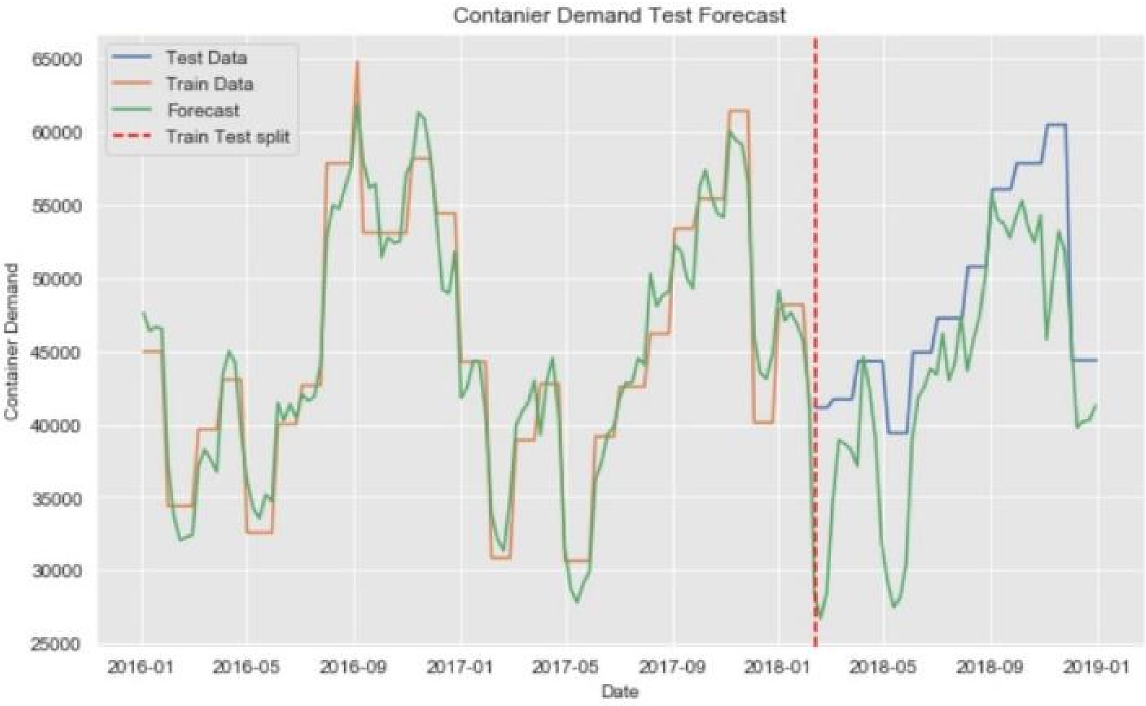
(c)


Long-term Demand Forecast
Train Data
Test Data
Train Forecast
Test Forecast
Train Test split
Test Forecast split
Long-term Forecast
Container Demand
65000
60000
55000
50000
45000
40000
35000
30000
2016-01
2016-07
2017-01
2017-07
2018-01
2018-07
2019-01
2019-07
2020-01
Date

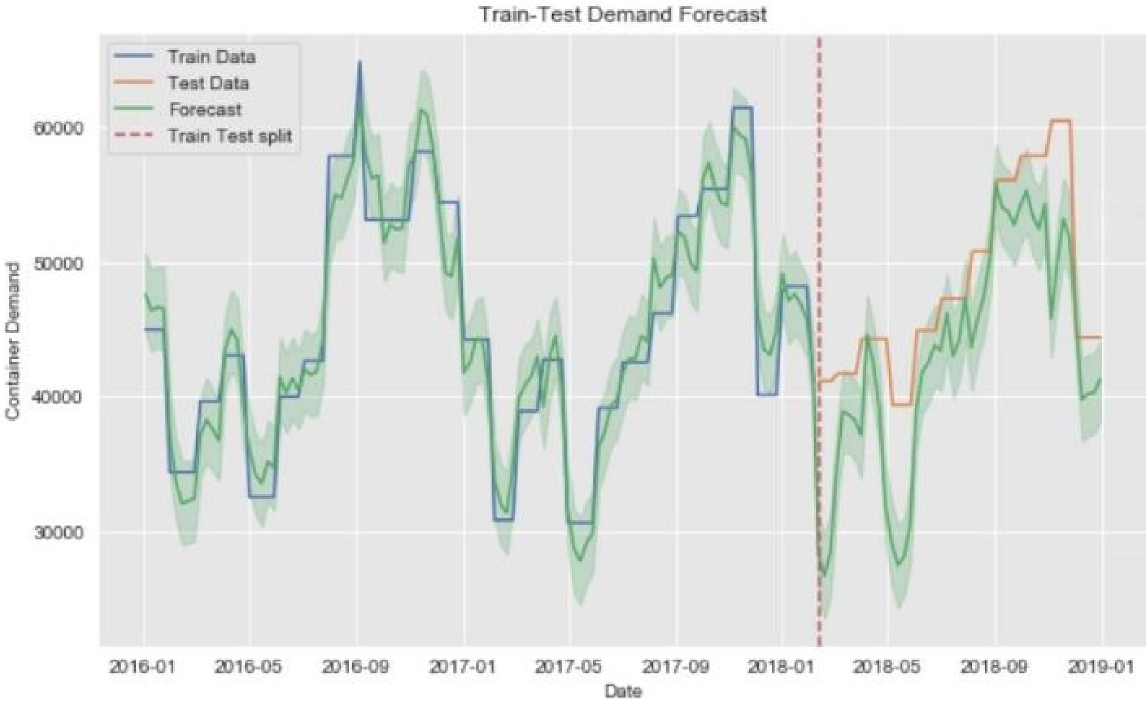
(d)

Fig. 7-11 Demand Forecast by Seasonal Holt-Winters' (a) Test forecast (b) Train test forecast (c) Short-term forecast (d) Long-term forecast

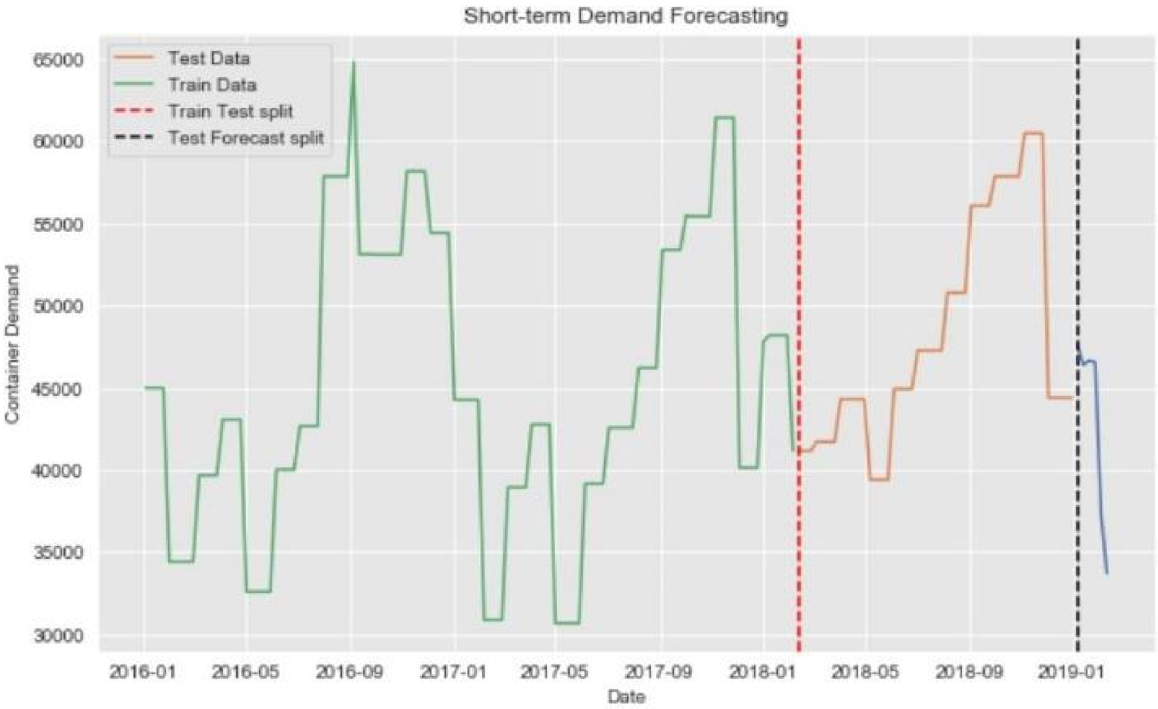
In the third step, Facebook Prophet is trained using the parameters shown in figure 7.8 (a) and (b). It can be seen that Prophet offers two additional flexible tuning parameters for researchers, i.e., custom-defined holidays (see figure 7.8 (a)) and a custom change point definition. In addition to this, the model provides ability to impose dominant or recessive effect of holidays based on requirements by modifying the parameters name 'Holidays prior scale (see figure 7.8 (a)). Greater values of the said variable, greater would be the effect of holidays over the forecasts. Furthermore, the model provides an edge over other models by offering the control over weekly, monthly and yearly seasonality. Based on the dataset, same can be set true and false (refer to Figure 7.8 (a)). The test forecast provided by Facebook Prophet is shown in Figure 7.12(a). The collective train and test forecast are shown in Figure 7.12(b). The short-term i.e., 6 weeks and long term i.e., 52 weeks forecast are presented in Figure 7.12 (c) and (d) respectively.


Contanier Demand Test Forecast
Test Data
Train Data
Forecast
Train Test split
Container Demand
65000
60000
55000
50000
45000
40000
35000
30000
25000
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
2018-09
2019-01
Date

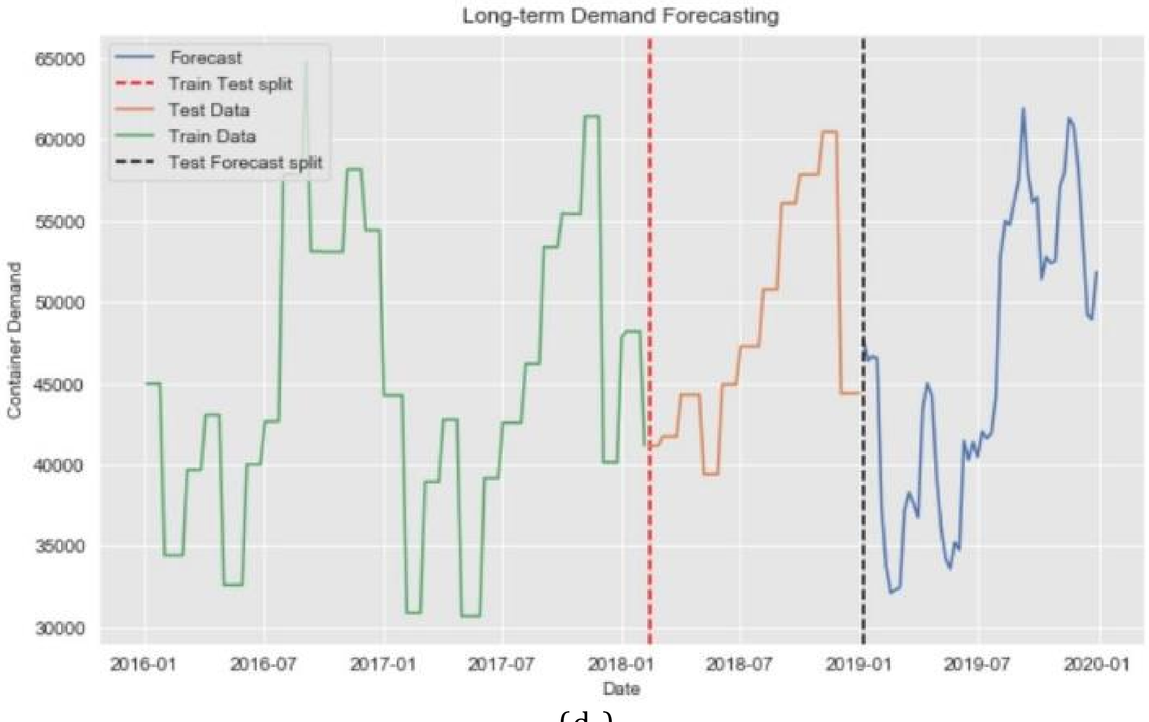
(a)


Train-Test Demand Forecast
Train Data
Test Data
Forecast
Train Test split
Container Demand
60000
50000
40000
30000
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
2018-09
2019-01
Date

(b)


Short-term Demand Forecasting
Test Data
Train Data
Train Test split
Test Forecast split
Container Demand
65000
60000
55000
50000
45000
40000
35000
30000
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
2018-09
2019-01
Date

(c)


Long-term Demand Forecasting
Forecast
Train Test split
Test Data
Train Data
Test Forecast split
Container Demand
65000
60000
55000
50000
45000
40000
35000
30000
2016-01
2016-07
2017-01
2017-07
2018-01
2018-07
2019-01
2019-07
2020-01
Date

(d)

Fig. 7-12 Demand Forecast by PROPHET (a) Test forecast (b) Train test forecast (c) Short-term forecast (d) Long-term forecast

7.5 Evaluation:

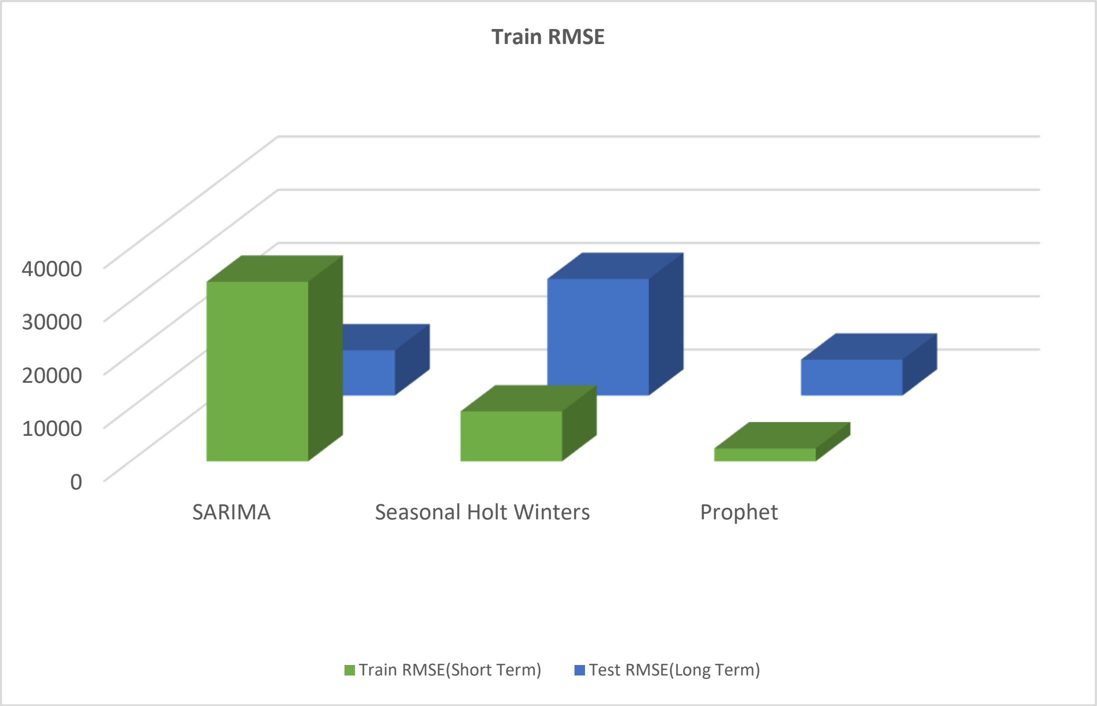
An accurate evaluation is fundamental to conclude the best fit model. We have selected root mean squared error (RMSE) and mean absolute percentage error (MAPE) to evaluate the performance of the selected models. RMSE can be computed using equation (7.8) given below.

$$RMSE = \sqrt{\overline{(f - o)^2}} \tag{7.8}$$

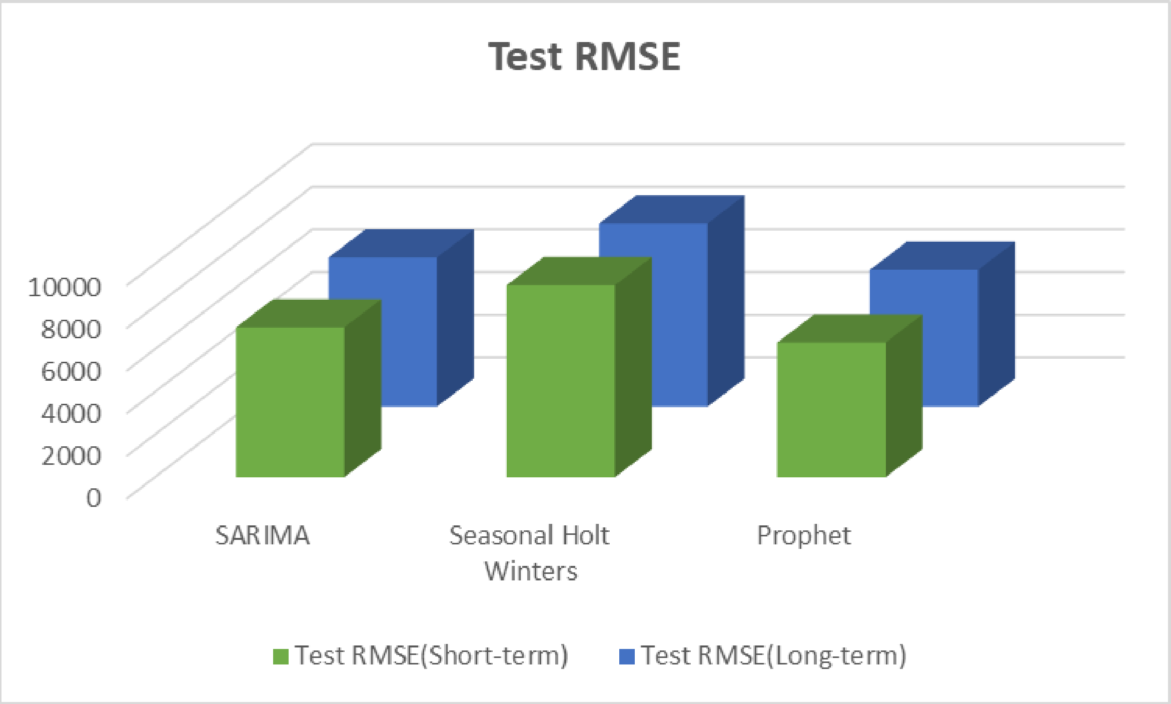
In (7.8) f is the forecast, and o is the observed value. MAPE measures forecast accuracy as a percentage. MAPE can be calculated as equation (7.9).

$$MAPE = \frac{1}{N} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \tag{7.9}$$

In (7.9) A_t is the actual value and F_t is the forecast. Figure 7.13 shows the comparison of RMSE achieved by the selected models (for both short-term and long-term). Figure 7.13 (a) and (b)shows the RMSE achieved using the train and test data respectively. It can be seen that Facebook Prophet offers lesser RMSE for both train and test data in comparison to the other models i.e., SARIMA and Holt-Winters' Seasonal Model. However, Holt-Winters' Seasonal Model provides lesser RMSE than SARIMA over training dataset but loses its performance over test data. Thus, it can be concluded that Prophet outperforms both SARIMA and Holt-Winters' Seasonal model.


Train RMSE
40000
30000
20000
10000
0
SARIMA
Seasonal Holt Winters
Prophet
🟩 Train RMSE(Short Term)
🟦 Test RMSE(Long Term)

(a)


Test RMSE
10000
8000
6000
4000
2000
0
SARIMA
Seasonal Holt Winters
Prophet
🟩 Test RMSE(Short-term)
🟦 Test RMSE(Long-term)

(b)

Fig. 7-13**Error! Use the Home tab to apply 0 to the text that you want to appear here.** Comparative RMSE of forecasting models

This conclusion can be supplemented by looking at the comparative MAPE values of the selected algorithms presented in Figure 7.14. It can be seen that Prophet reaches approximately 4.55% and 11.21 MAPE for train and test dataset respectively. On the other hand, the train and test datasets MAPE provided by SARIMA is 62.69.46% and 36.88% respectively while Seasonal Holt-Winters' attains 17.46% and 12.95% MAPE for train and test respectively. Henceforth, it can be concluded that Facebook Prophet outperforms both Seasonal Holt-Winters' and SARIMA.

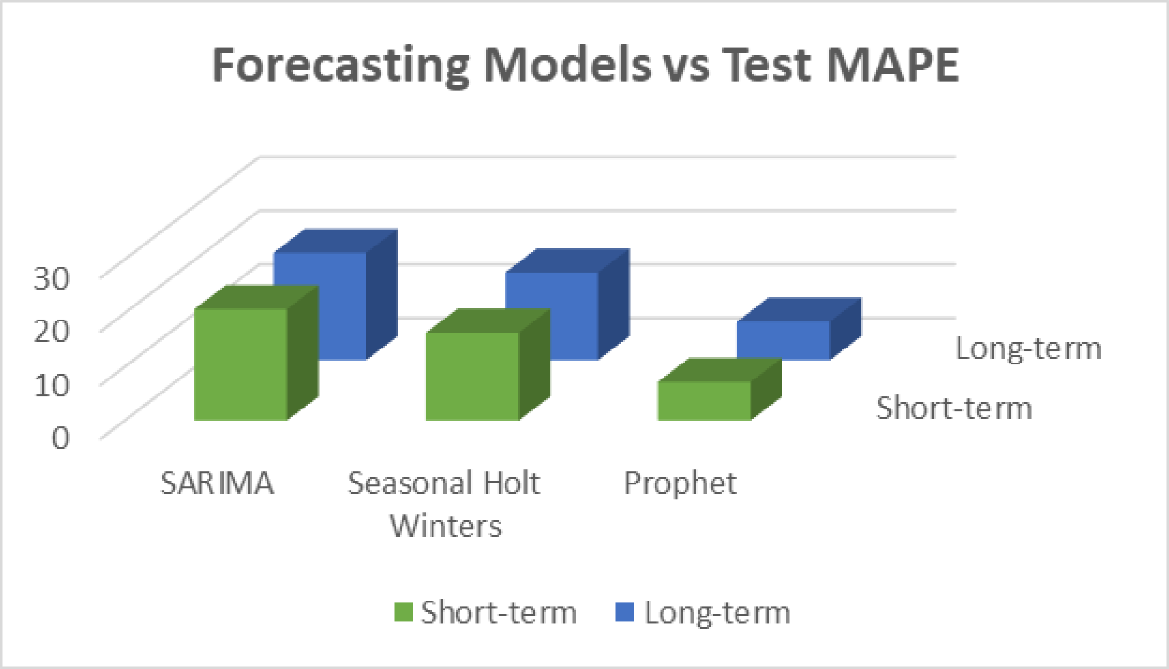


Fig. 7-14 Comparative MAPE values of SARIMA, Prophet and Seasonal Holt-Winters

7.6 Summary:

In this chapter we have described the methodology using which we forecasted the shipment demand forecasting for both short and long-term.

We have applied three existing state of the art time-series forecasting models on shipment dataset and found FB Prophet the most effective one.

In the next chapter we explain the how the price prediction models are incorporated for shipment demand prediction in the shipping industry.

7.7 References:

1. Chen, R., J.-X. Dong, and C.-Y. Lee, Pricing and competition in a shipping market with waste shipments and empty container repositioning. Transportation Research Part B: Methodological, 2016. 85: p. 32-55.

2. Larry Montan, T.K., Julie Meehan, Getting Pricing Right The value of a multifaceted approach. Deloitte University Press.

3. www.portofmelbourne.com/about-us/trade-statistics/monthly-trade-reports/. Port of Melbourne, VIC, Australia. 27 May 2019].

4. Fremantleport ,WA, Australia. [cited 2019 27 May]; Available from: www.fremantleports.com.au/trade-business/container-traffic-reports

5. www.portbris.com.au/Operations-and-Trade/Trade-Development/. Port of Brisbane, QLD,Australia. [cited 2019 27 May].

6. www.flindersports.com.au/ports-facilities/port-statistics/. Flinders Port, SA, Australia. [cited 2019 27 May].

7. www.nswports.com.au/resources/trade-results/. Port Botany, NSW, Australia. [cited 2019 27 May].

8. Ubaid, A., F. Hussain, and J. Charles, Modeling Shipment Spot Pricing in the Australian Container Shipping Industry: Case of ASIA-OCEANIA trade lane. Knowledge-Based Systems, 2020. 210: p. 106483.

9. [cited 2020 18 May]; Missing value handeling]. Available from: https://scikit-learn.org/stable/modules/impute.html.

10. Timse Series Split. 2020 [cited 2020 25-08]; Available from: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html.

11. Arlt, J. and P. Trcka, Automatic SARIMA modeling and forecast accuracy. Communications in Statistics - Simulation and Computation, 2019: p. 1-22.

12. Pongdatu, G. and Y. Putra, Seasonal Time Series Forecasting using SARIMA and Holt Winter's Exponential Smoothing. IOP Conference Series: Materials Science and Engineering, 2018. 407: p. 012153.

13. Facebook Prophet. 22/09/2019]; Available from: https://facebook.github.io/prophet/.

14. Mizzen Group Pty Ltd. Available from: https://www.mizzengroup.com/.

8 Price Prediction Models for The Australian shipping Industry

8.1 Introduction:

In this chapter, we have addressed the research problem identified in objective four of this thesis. We have trained, tested and selected an optimal Machine Learning model for price prediction models for container pricing based on demand and supply for the Australian container shipping industry. The sourcing of demand, supply and pricing data has been done from Australian ports, Sea-Intelligence maritime analysis and the Shanghai Freight Index (SCFI) respectively. Data-driven predictions have been realized by applying three different regression models that include support vector regression (SVR), random forest regression (RFR) and gradient booster regression (GBR) over the gathered datasets after initial feature engineering. A comparison of research outcomes shows that GBR outperforms all the other models by offering a test accuracy of 84%. This chapter is structured as follows. In section 8.2 research design is explained followed by research outcomes in section 8.3. Section 8.4 concludes the chapter.

8.2 Background and Motivation:

The global supply chain is complex, with cargo volumes that are highly seasonal, driven by consumer-related events such as Christmas and Chinese New Year, agriculture harvests that are further impacted by extreme weather occurrences and changes to the geo-political regulatory environment affecting trade. In contrast, supply or shipping capacity is fixed in the short term. This results in periods of mismatched supply and demand and therefore shipping price volatility [1]. The shipping lines are trapped by the current spot market pricing practice of using data validity and not vessel voyage. This causes lack of positive relationship between price and supply and demand, a problem which is compounded by the shipping line's enterprise systems. Entrenched operational silos within the shipping lines result in missed revenue opportunities through the lack of real-time visibility into the availability of equipment and vessel space which are the key inputs to pricing decisions. Forty percent of all shipping containers moved around the world are purchased in the short-term spot market and their commercial terms are set manually via emails and phone calls [2]. The spot market price should be governed by supply and demand; however, the industry as a whole has little visibility into the state of the market in real-time, hence carriers are making sub-optimal pricing decisions and customers are making poor procurement decisions because of this situation.

The purpose of this research is to empower the current manually operating Australian shipping industry with a data-driven price prediction capability driven by demand and supply. To achieve this goal, machine learning (ML) models are applied which are currently being used in almost every area such as business, industrial engineering, medical, physics and statistics to predict future events such as disease diagnosis [3].

8.3 Experiment Design:

This section explains our research scenario in order to develop a price prediction model in the Australian container shipping industry using machine learning algorithms. We started by collecting real-time demand, supply and pricing datasets from multiple sources. In the second step, the datasets are consolidated into a single dataset. Three ML regression-based algorithms are selected and tested in Python. To determine the performance of each regression-based model, three evaluation measures are observed: (1) root mean square error (RMSE); (2) R2 score and (3) accuracy.

Data sourcing, cleansing and integration:

Datasets from three different sources were gathered. Demand data was collected from several Australian ports, namely Port Botany [4], the Port of Melbourne[5], the Port of Brisbane [6], Flinders Port [7] and Fremantle Port [8]. Supply data was gathered from Sea-Intelligence maritime analysis. (Sea-Intelligence is an analytical reporting company which publishes weekly supply related trade summaries for the container shipping industry). Supply data contains weekly total capacity (in TEUs) of shipping companies operating to and from Australia. Pricing data for imports was gathered from the Mizzen group propriety database and Shanghai Freight Index (SCFI).

Selecting the time horizon is the first step in data cleansing. Historic data from the last three years was selected for analysis i.e., 2016- 2018. The acquired dataset contains data of all the trade lanes operating to and from Australia. Since the scope of this research work is to develop a price prediction model specifically for the Asia-Oceania Trade Lane, in the second step, trade-lane-specific data is filtered from the original datasets. The dataset was filtered using domain-driven selective wrapping (DSW)[9], which incorporates a domain expert's knowledge for the selection of subsets from the original dataset. Once the trade lane's data is segregated for demand, all the empty and full containers are added to get the total demand for the Asia-Oceania trade lane. This demand dataset is a combination of both export and import demand. In the next step, the import and export data are separated using the port data ratios, thus providing us with the total import demand for the Asia-Oceania Trade Lane, which is required for our research. The supply and pricing information was also filtered from their respective datasets. Tables 8.1, 8.2 and 8.3 show the cleaned data from the demand, supply, and pricing datasets. Missing values are handled using the back-fill methodology [10]. The data integration process involves combining supply, demand and price datasets into a single unified dataset. Table 8.4 shows the final dataset used for predicting the weekly import container price, based on demand and supply.

Table 8-1 List of features selected from the demand dataset.

Region	Feature Name	Description
Asia Oceania	Imports	Number of inbound containers
	Full	Total quantity of filled inbound container
	Empty	Total inbound empty container
	Date	Starting Date of Week

Table 8-2 List of features selected from supply dataset.

Region	Feature Name	Description
Asia Oceania	Total Supply	Total Available Capacity
	Date	Starting Date of Week

Table 8-3 List of features from pricing dataset

Region	Feature Name	Description
Asia Oceania	Price	Weekly container pricing
	Date	Starting Date of Week

Table 8-4 Dataset used in research (import price prediction for Asia-Oceania trade lane in the Australian shipping industry).

Region	Feature Name	Description
Asia Oceania	Total Supply	Total available capacity
	Total Demand	Total import Demand

	Price	Shipment Pricing
	Date	Starting Date of Week

8.4 Result and Evaluation:

In this section, we describe the machine learning models which we apply to the shipping dataset. The first step is to perform exploratory data analysis (EDA). The relationship between supply, demand, and price is shown in Figure 8.1. The data summary is presented in Table 8.5. The scatter plot in figure 8.1 depicts that the demand and supply has no linear dependency over the dataset. The prices have no link with shipment supply and demand and are randomly assigned. Thus, predicting price using this dataset will provide a random shipment price which is not going to provide prices based on shipment demand and available supply (capacity). On the other hand, we have modified the relationship between price and shipment demand and supply based on a mathematical model based on historic data (please refer to chapter 6). Thus, the relationship is provided by Opti-Price for pricing based on shipment capacity and demand respectively (please refer to figure 6.5). Therefore, we have applied machine learning models on both the pricing datasets and come up with the prediction model that can predict shipment prices based on shipment demand and capacity.

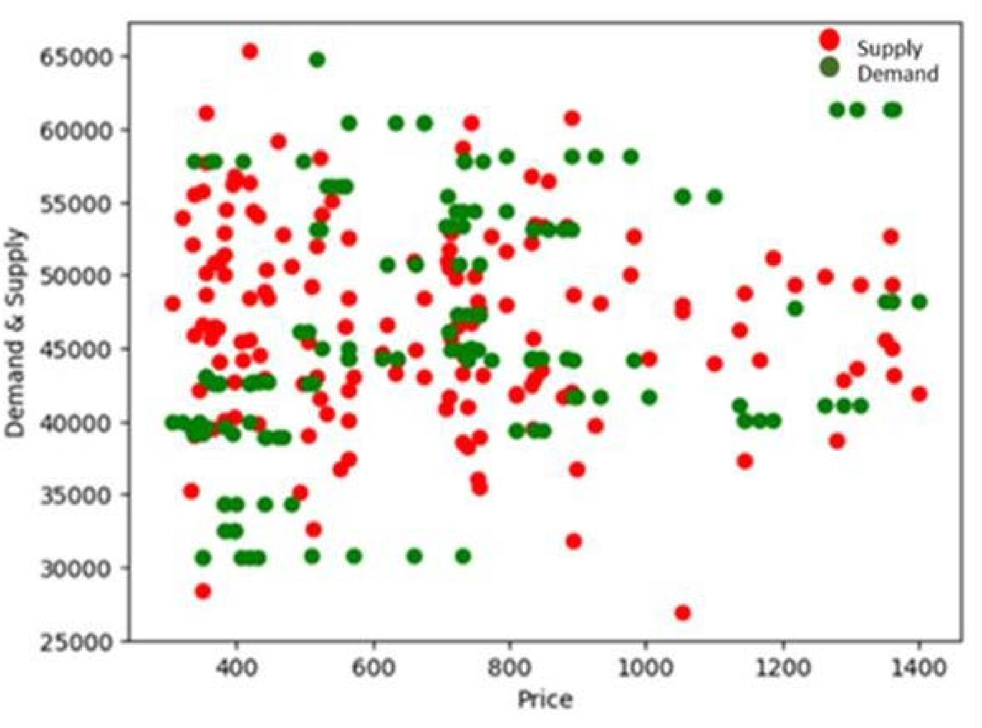
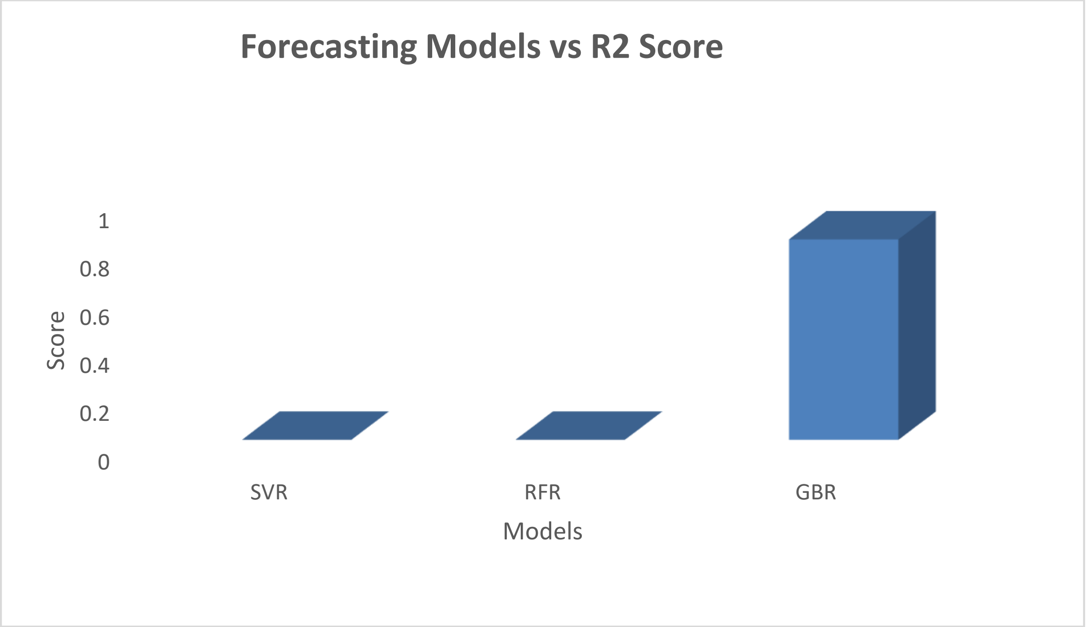


Fig. 8-1 Demand, supply and pricing data distribution

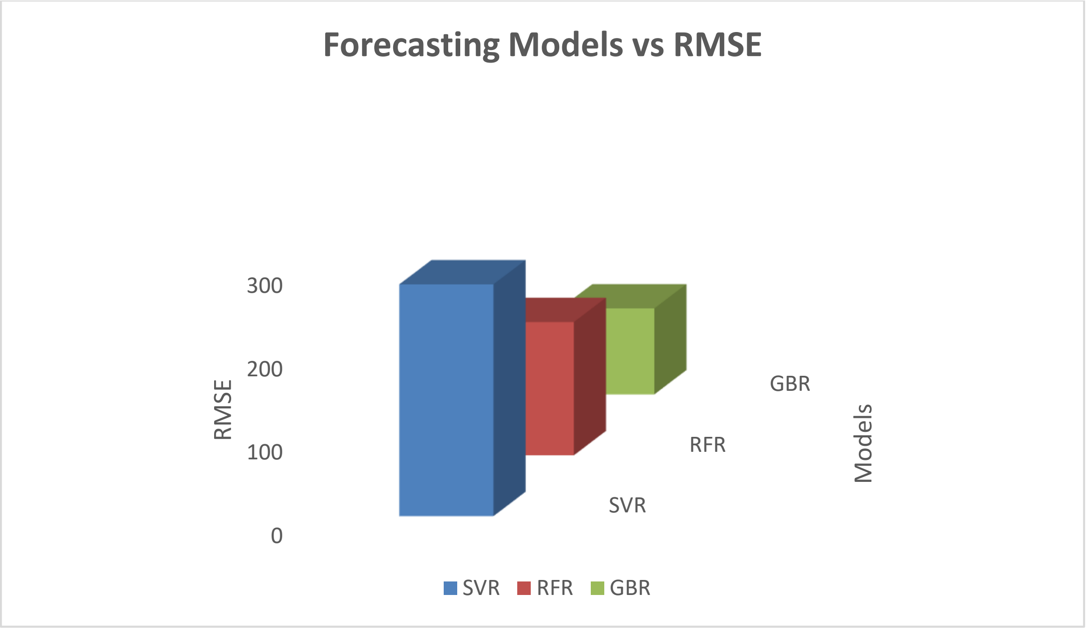
Table 8-5 Data summary of key variables

Variable	Mean	Std	Min	Max	70%
Supply	65669.67	6588.804	8373.84	83733	50992.96
Demand	47891.7	8430.8	39992.08	59300.12	53381.48
Price	825.05	285.188	532	1399	833.0

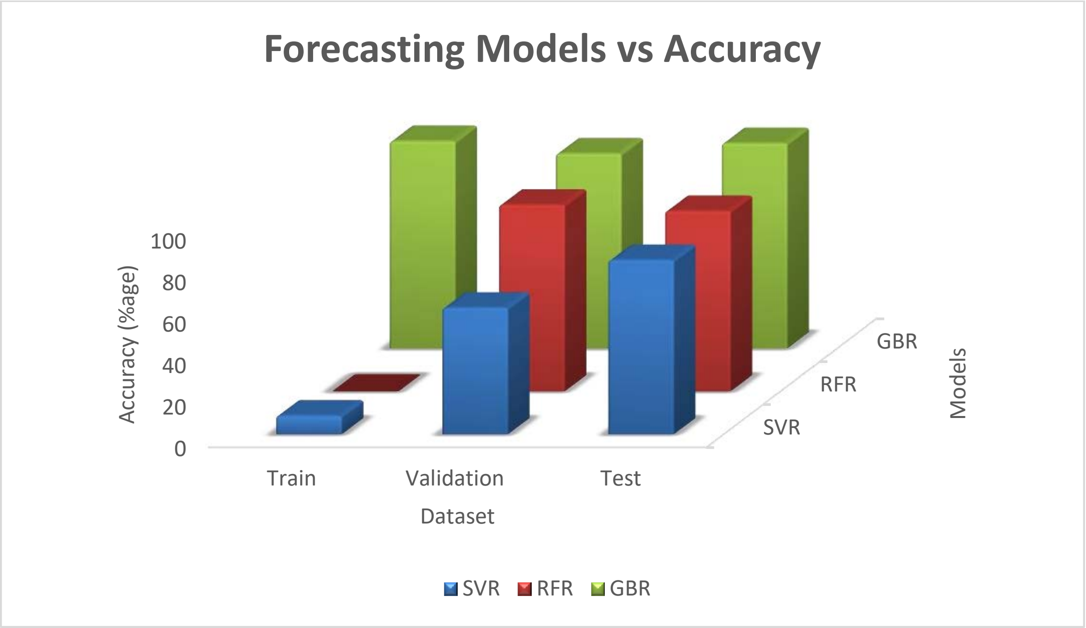
A careful evaluation of the regression model is very important to determine the best fit model. To do so, three-way data splitting is performed (test, train and validation datasets). We divide the dataset into three parts and each part is split into three. 70% of the data is used for training, 20% of the data is used for validation and the remaining 10% of the data is used for testing. RMSE, R2 score and accuracy are used to evaluate the models. Figure 8.2 shows the comparative RMSE, accuracy and R2 score of the models. It can be observed that GBR has the smallest RMSE value amongst other selected regression models and the greatest R2 score. However, SVR fits the training data well but performs poorly on the validation and testing datasets. Both RFR and GBR are tree-based structures and perform well for data coverage for both the training and validation dataset. However, RFR fails to generalize the test dataset. GBR, on the other hand, achieves the best performance for generalizing the testing dataset of the selected regression models. Thus, we select GBR for price prediction in the Australian shipping industry. The training, validation and testing accuracies of the applied models are summarized in Table 8.6.



(a)



(b)



(c)

Fig. 8-2 Comparison results for models (a) RMSE (b) R2 Score (c) Accuracy

Table 8-6 Accuracy percentage of regression models

Model	Train Accuracy (%)	Validation Accuracy (%)	Test Accuracy (%)
SVR	100	0	9
RFR	94	90	61
GBR	99	87	84

From the results discussed above, it is evident the GBR outperforms all other algorithms in its comparison by offering minimum RMSE, and maximum R^2 score. Additionally, it also has maximum prediction accuracy for all train, validate and test dataset.

8.5 Summary:

In this chapter we have explained the application of three regression-based ML models on real-time datasets to predict the price of import shipment containers in the Australian shipping industry, specifically for the Asia-Oceania Trade Lane. The container price is predicted based on current demand and available supply. To analyze the performance of these models, R2 score, accuracy, and RMSE are measured. The evaluation results show that GBR performs best over the available dataset.

In the next chapter we explain the process flow of the software prototype designed to demonstrate the working of the research model.

8.6 References:

1. Munim, Z. and H.-J. Schramm, Forecasting container shipping freight rates for the Far East – Northern Europe trade lane. Vol. 19. 2017. 106-125.

2. Dey, A., *Machine Learning Algorithms: A Review* International Journal of Computer Science and Information Technologies, 2016. **Vol. 7 (3)**(1174-1179).

3. Ebrahimian, H., et al., *The price prediction for the energy market based on a new method.* Economic Research-Ekonomska Istraživanja, 2018. **31**(1): p. 313-337.

4. www.nswports.com.au/resources/trade-results/. *Port Botany, NSW, Australia*. [cited 2019 27 May].

5. www.portofmelbourne.com/about-us/trade-statistics/monthly-trade-reports/. *Port of Melbourne, VIC, Australia*. 27 May 2019].

6. www.portbris.com.au/Operations-and-Trade/Trade-Development/. *Port of Brisbane, QLD,Australia*. [cited 2019 27 May].

7. www.flindersports.com.au/ports-facilities/port-statistics/. *Flinders Port, SA, Australia*. [cited 2019 27 May].

8. www.fremantleports.com.au/trade-business/container-traffic-reports. *Fremantleport ,WA, Australia*. [cited 2019 27 May].

9. Ubaid, A., F. Dong, and F.K. Hussain. Framework for Feature Selection in Health Assessment Systems. in Advanced Information Networking and Applications. 2020. Cham: Springer International Publishing.

10. www.pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html. *Pandas Official Website*. [cited 2019 25 July].

9 Illustrative Examples of Predicting Shipment Demand and Shipment Price Prediction in Container Shipping Industry

9.1 Introduction:

In this chapter, we present the software prototype description designed to provide interface between shipment demand predictions and price prediction models (based on available shipment capacity and current shipment demand).

For calculating shipment demand ML model is trained using the historic shipment demand dataset, which then predicts short-term i.e., 6 weeks ahead of shipment demand and long-term shipment demand for 12 weeks. These predictions are saved in the csv file in order to save the predictions for multiple weeks. From this file, the UI is populated to display the predictions to the users.

In order to predict price for the shipment price, the predicted shipment demand and available capacity is used as input. The user uses the predicted shipment demand and available shipment capacity (already published by shipping stakeholders) to the price prediction UI that calculates the shipment spot/contract pricing.

In the next section, we provide details of the software prototype we have designed to demonstrate the functionality of our research work. Subsequently in section 9.3, we explain the price prediction model's working along with the user interface snapshots. Following which, the flow charts of the intelligent data driven methods for shipment demand and capacity are shown in section 9.4. In section 9.5, the formal user interface of the shipment demand and price launched by Mizzen Group is presented.

The access to this tool is paid. Hence, we provide only the screen shot of the user interface. Finally, section 9.6 concludes this chapter.

9.2 Intelligent Data Driven Methods for shipment demand and price prediction:

In this section with the help of a case study, we will explain the process determining shipment price based on current shipment demand and available capacity. The case study is as follows:

Let us assume that Alice wants to arrange shipment of her materials. Alice not only want to know the contract price of the shipment but also has some important stuff to be shipped as soon as possible. Alice is an old customer of the shipping company XYZ. Joe is the consultant in XYZ and has been working from many years in XYZ. He not only has information of Alice being a loyal customer but also has knowledge of new data driven models being implemented in the company. His aim is to make Alice happy and earn maximum revenue for his business.

As can be seen from the above description, Alice is looking for good prices for her shipment items. She is looking for cheaper shipment cost for contract prices shipments along with the urgent shipment order. However, XYZ Company has two aims, (1) make their loyal customer happy and (2) earn maximum revenue from both contract and spot prices.

To achieve the business goal, XYZ will use intelligent data driven models for shipment demand and price prediction in the shipping industry proposed in this thesis. The first step will be looking at the shipment demand forecast provided by the deployed ML mode, the details of which are discussed in the next section.

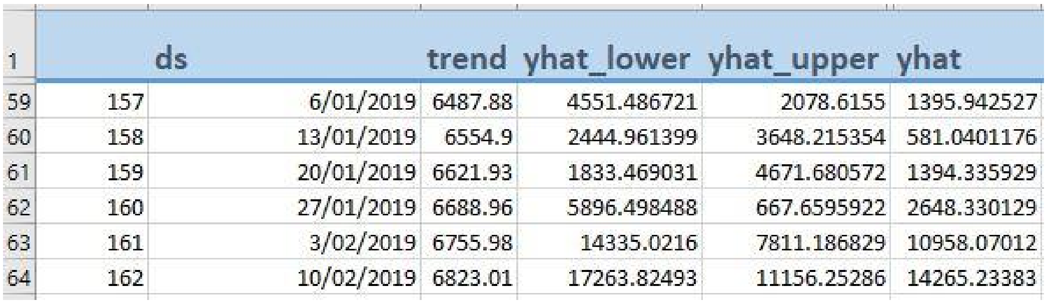
9.3 Prediction of Short-term and Long-term Shipment Demand in the Container Shipping Industry:

As discussed in chapter 7, the shipment demand forecasting model will provide the short-term and long- term demand forecast and the result will be saved in the csv file. From the csv file, the results are loaded into the XYZ's tool. The snapshot of the result produced by the forecasting models are shown in figure 9.1 given below. In figure 9.1 (a), ds shows the forecasted week date starting from Monday, trend shows the possible future trend of shipment demand i.e., whether it would be an increasing variable or decreasing. The yhat_lower and yhat_upper shows the possible minimum lower and upper boundary for the shipment demand and yhat depicts the forecasted demand. Once Joe has an industry wide view of the shipment current and future demand, Joe will use these demand stats into the price prediction tool for predicting the spot and contact prices. For spot pricing, the current shipment demand is observed, while for contract prices, Joe will look into the long-term demand forecast for any specific week. The process of performing shipment price prediction is explained in section 9.4.

9.4 Prediction of Shipment Price in the Container Shipping Industry:

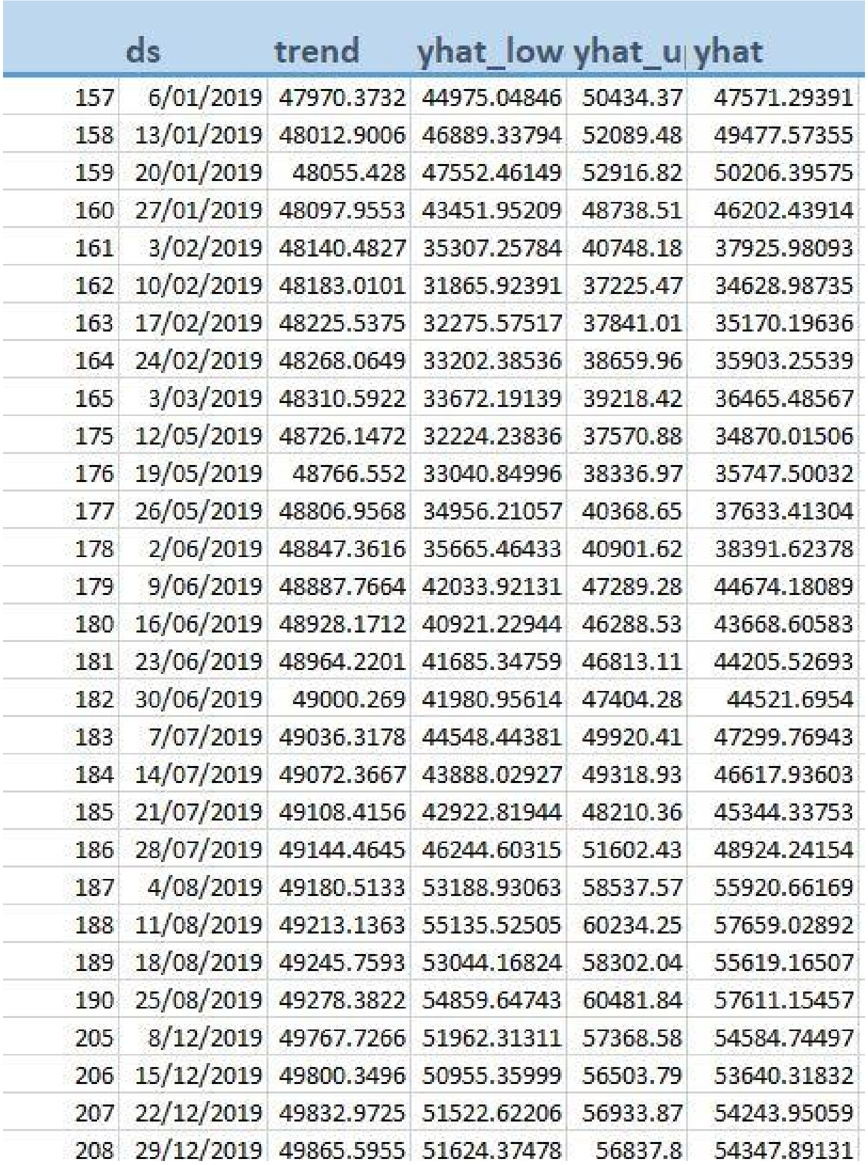
As discussed in chapter 6, and 8, the shipment prices are based on shipment demand and capacity. However, the current prices are not in line with the shipment demand. Thus, we presented Opti-price in chapter 6 to cover the research gap. Joe will have access to shipment current/future demand from the forecasting model (as explained in section 9.3).

In order to quote spot and contract price for Alice, Joe will use the price prediction tool. Figure 9.2 below shows the user interface (UI) of the software prompting user to input the shipment demand and capacity at a particular interval of time. When the UI loads, the user has the option to enter shipment demand and capacity as shown in figure 9.2 (a). Once the user enters the demand and capacity value they predict button (see figure 9.2 (b)). Once the user enters the shipment demand and available shipment capacity (supply) and clicks predict, the machine learning model running at the backend is activated and the predicted shipment price is displayed. This provides the pricings that are usually quoted in the industry without considering shipment demand and capacity as shown in figure 9.2 (c).



1		ds	trend	yhat_lower	yhat_upper	yhat
59	157	6/01/2019	6487.88	4551.486721	2078.6155	1395.942527
60	158	13/01/2019	6554.9	2444.961399	3648.215354	581.0401176
61	159	20/01/2019	6621.93	1833.469031	4671.680572	1394.335929
62	160	27/01/2019	6688.96	5896.498488	667.6595922	2648.330129
63	161	3/02/2019	6755.98	14335.0216	7811.186829	10958.07012
64	162	10/02/2019	6823.01	17263.82493	11156.25286	14265.23383

(a)



	ds	trend	yhat_low	yhat_u	yhat
157	6/01/2019	47970.3732	44975.04846	50434.37	47571.29391
158	13/01/2019	48012.9006	46889.33794	52089.48	49477.57355
159	20/01/2019	48055.428	47552.46149	52916.82	50206.39575
160	27/01/2019	48097.9553	43451.95209	48738.51	46202.43914
161	3/02/2019	48140.4827	35307.25784	40748.18	37925.98093
162	10/02/2019	48183.0101	31865.92391	37225.47	34628.98735
163	17/02/2019	48225.5375	32275.57517	37841.01	35170.19636
164	24/02/2019	48268.0649	33202.38536	38659.96	35903.25539
165	3/03/2019	48310.5922	33672.19139	39218.42	36465.48567
175	12/05/2019	48726.1472	32224.23836	37570.88	34870.01506
176	19/05/2019	48766.552	33040.84996	38336.97	35747.50032
177	26/05/2019	48806.9568	34956.21057	40368.65	37633.41304
178	2/06/2019	48847.3616	35665.46433	40901.62	38391.62378
179	9/06/2019	48887.7664	42033.92131	47289.28	44674.18089
180	16/06/2019	48928.1712	40921.22944	46288.53	43668.60583
181	23/06/2019	48964.2201	41685.34759	46813.11	44205.52693
182	30/06/2019	49000.269	41980.95614	47404.28	44521.6954
183	7/07/2019	49036.3178	44548.44381	49920.41	47299.76943
184	14/07/2019	49072.3667	43888.02927	49318.93	46617.93603
185	21/07/2019	49108.4156	42922.81944	48210.36	45344.33753
186	28/07/2019	49144.4645	46244.60315	51602.43	48924.24154
187	4/08/2019	49180.5133	53188.93063	58537.57	55920.66169
188	11/08/2019	49213.1363	55135.52505	60234.25	57659.02892
189	18/08/2019	49245.7593	53044.16824	58302.04	55619.16507
190	25/08/2019	49278.3822	54859.64743	60481.84	57611.15457
205	8/12/2019	49767.7266	51962.31311	57368.58	54584.74497
206	15/12/2019	49800.3496	50955.35999	56503.79	53640.31832
207	22/12/2019	49832.9725	51522.62206	56933.87	54243.95059
208	29/12/2019	49865.5955	51624.37478	56837.8	54347.89131

(b)

Fig. 9-1 Forecasted Shipment Demand (a) Short-term shipment demand forecast (b) Long-term shipment demand forecast

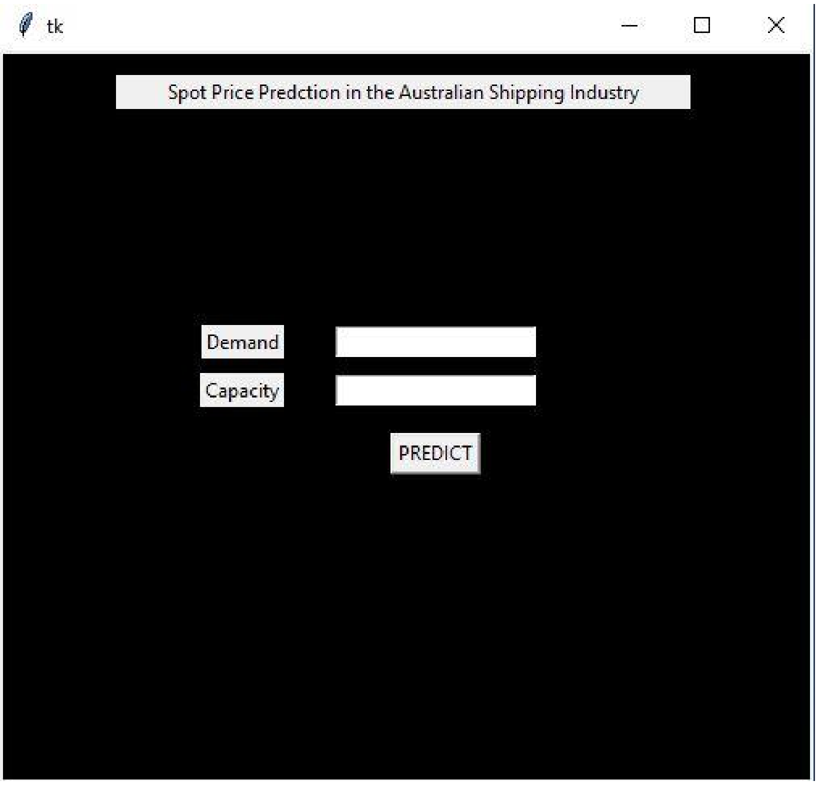
In order to get spot price for Alice, Joe will use current shipment demand values and capacity (from records).

In addition to this, the designed software is also capable top provide optimal prices that can increase the business revenue. Figure 9.3 below shows the UI similar to the UI shown in figure 9.3. However, once the predict button is pressed after entering the respective value of shipment demand and capacity (see figure 9.3 (b)), it provides the optimal shipment cost calculate by the Opti price explained in chapter 7 (see figure 9.3 (c)). To get Opti-price for Alice, Joe will use the same demand and capacity values in Opti Price UI.

In this way, Joe will have two comparative spot prices for the same shipment date. Now Joe can see how much maximum profit XYZ can earn from this shipment. Now it is up to Joe to quote the prices of his own choice. Joe is a smart employee. Joe will quote Opti-Price to Alice. Alice will negotiate the price and Joe will reduce some price to make Alice happy. But not as low as the original prices. Alice is happy. Joe is happy. XYZ's CEO is happy and the company has earned some added cash. Alice is still a loyal and happy customer while business is prospering.

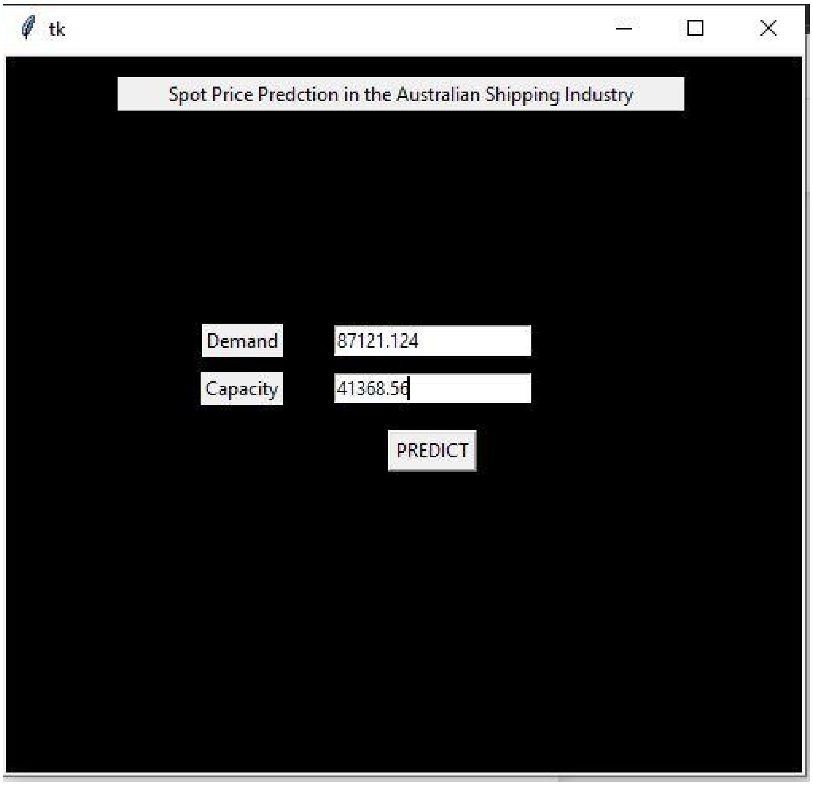
Joe will go through similar steps for contract prices for Alice but instead of current shipment demand, he will use forecasted shipment demand for the requested week and will work out with Alice for the pricing that would be beneficial for the company (i.e., keeping the customers happy while earning more revenue).

Thus the comparative prices provides the shipping stakeholders with the industry wide view of prices, shipment demand and capacity and based on their business legislations, they can quote any price to their customer based on the current scenario.



tk
Spot Price Predction in the Australian Shipping Industry
Demand
Capacity
PREDICT

(a)

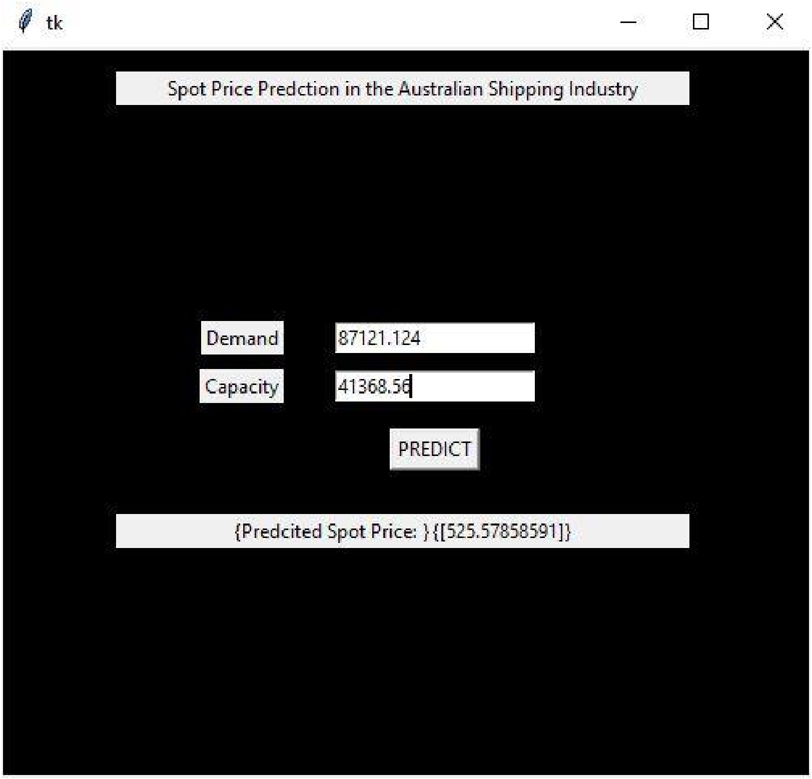


Spot Price Predction in the Australian Shipping Industry

Demand 87121.124

Capacity 41368.56

(b)



tk

Spot Price Predction in the Australian Shipping Industry

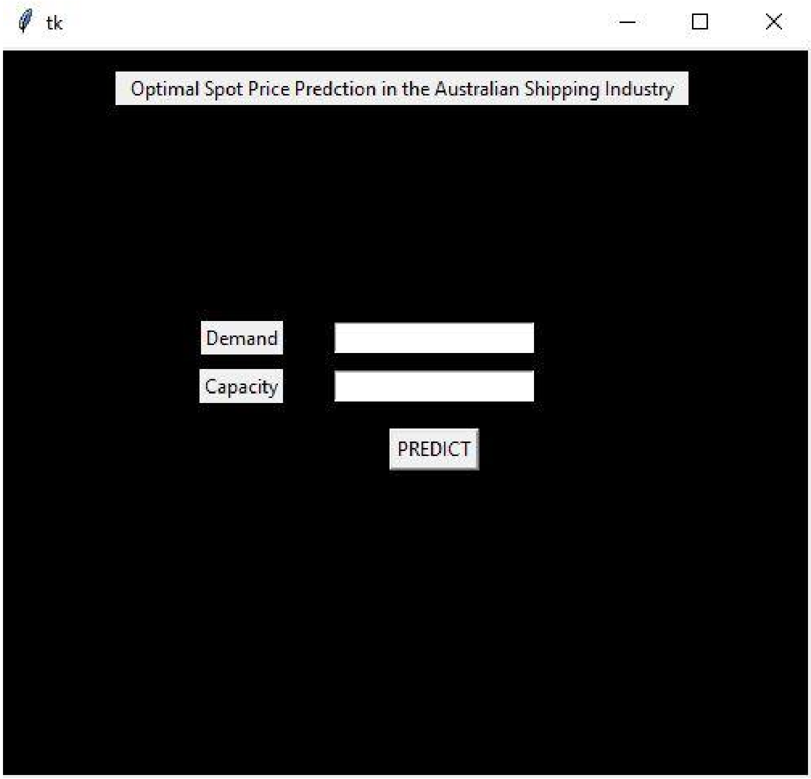
Demand 87121.124

Capacity 41368.56

{Predcited Spot Price: }{[525.57858591]}

(c)

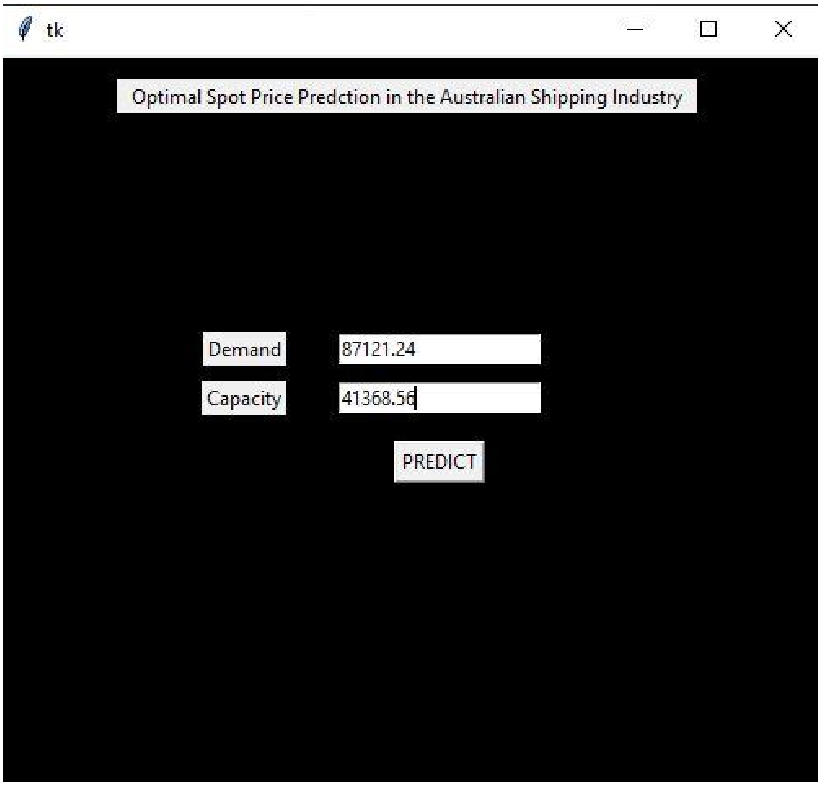
Fig. 9-2 Shipment spot pricing tool (a) Initial UI (b) UI showing the entered shipment demand and capacity (c) UI showing calculated shipment price based on manual pricing dataset.



tk

Optimal Spot Price Predction in the Australian Shipping Industry

(a)

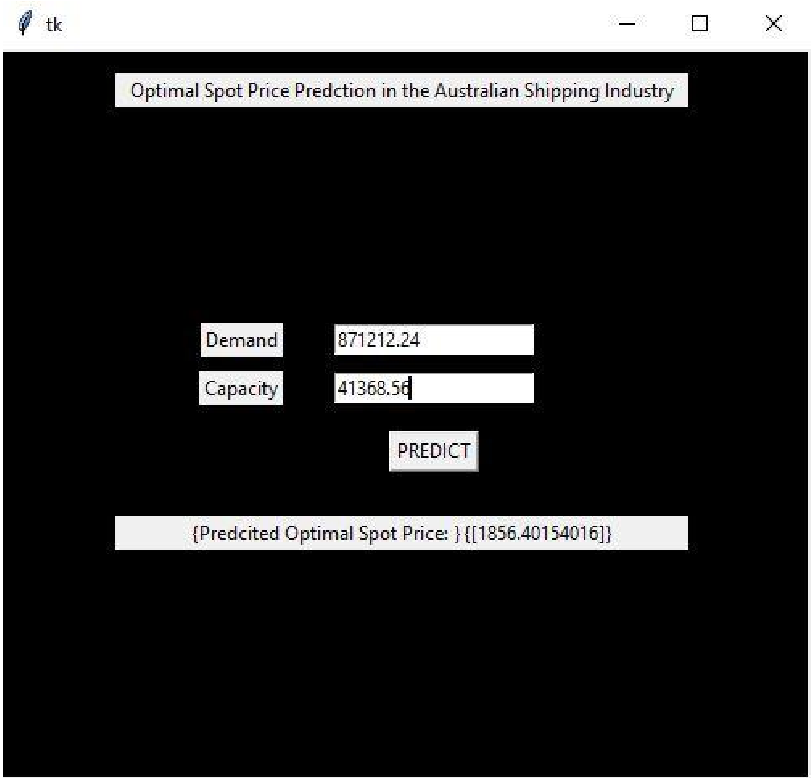


Optimal Spot Price Predction in the Australian Shipping Industry

Demand 87121.24

Capacity 41368.56

(b)



Optimal Spot Price Predction in the Australian Shipping Industry

Demand 871212.24

Capacity 41368.56

{Predcited Optimal Spot Price: }{[1856.40154016]}

(c)

Fig. 9-3 Shipment spot pricing tool (a) Initial UI (b) UI showing the entered shipment demand and capacity (c) UI showing calculated shipment price based on Opti-price

9.5 Abstract Flow Chart of the intelligent data driven methods of demand and price prediction in the Australian shipping industry:

Figure 9.4 below shows the flow chart demonstrating the process flow task adopted for shipment demand and price prediction in the Australian shipping industry. According to the process flow chart, the intelligent data driven methods of shipment demand and price prediction are comprised of three data set. Shipment demand, capacity and price. The data cleansing process is the first step towards developing the respective prediction models. The cleansed data is used to generate forecasting of shipment demand which is further used by the shipment price prediction model along with demand and capacity data set. However, the Opti price uses demand, capacity and prices dataset to provide optimal price derived from historical data statistics.

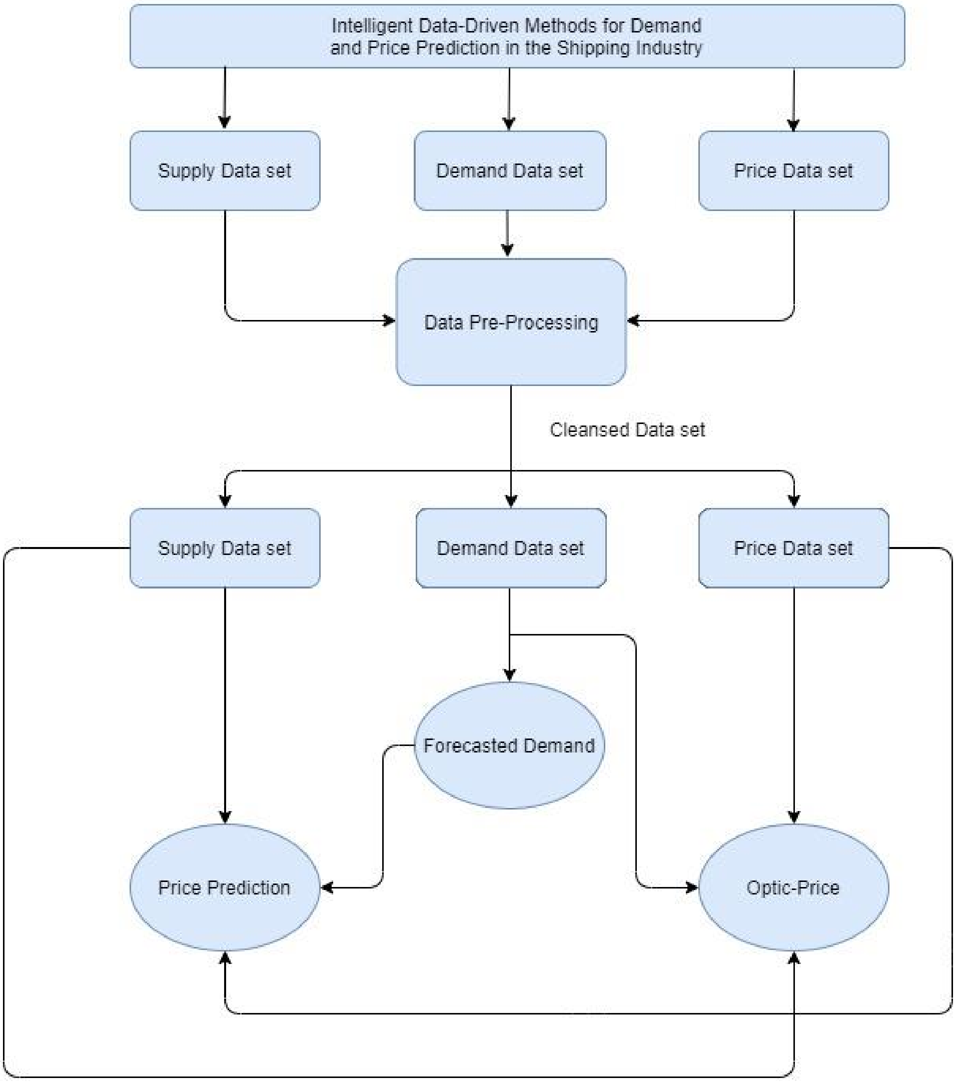


Fig. 9-4 Abstract Flowchart of Intelligent Data Driven Methods for Shipment Demand and Price Prediction in the Australian Shipping Industry

9.5.1 Flowchart of Short-term and Long-term demand prediction:

Figure 9.5 below shows the flow chart demonstrating the process flow task adopted for shipment demand forecasting. As seen from the flowchart in figure 9.5, the dataset is pre-processed following test-train split. After this division of the dataset, the model is trained over the dataset and is verified using test dataset. In our research, we have three comparative models to perform forecasting of shipment demand. Based on model's evaluation, the best performing model's results are saved in the file to be used for the software.

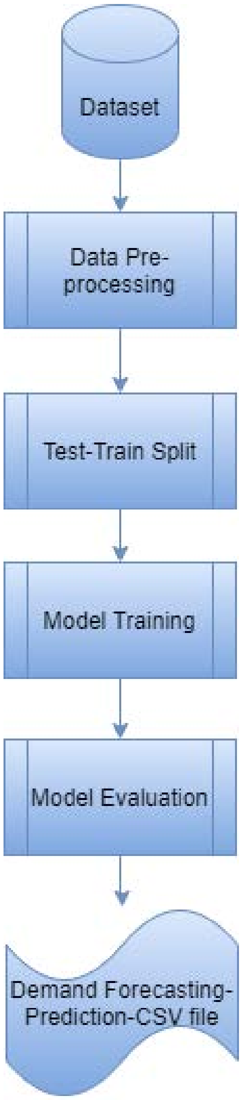


Fig. 9-5 Flow chart for Demand Forecasting Models

9.5.2 Flowchart of shipment price prediction:

Figure 9.6 below shows the flow chart demonstrating the process flow task adopted for shipment demand and price prediction in the Australian shipping industry. It can be seen from the flow chart that the price prediction model has the input from two datasets, the shipment capacity and the forecasted shipment demand as explained in chapter7, section 9.3

and 9.5.1.The capacity dataset requires pre-processing as explained in chapter 5. However, the forecasted demand does not need data pre-processing. Following this step, both the datasets are split into test-train split and are used by prediction model for training. In our research, we have three comparative regression-based models for predicting shipment demand. The best performing model is selected after model evaluation and is deployed for the production.

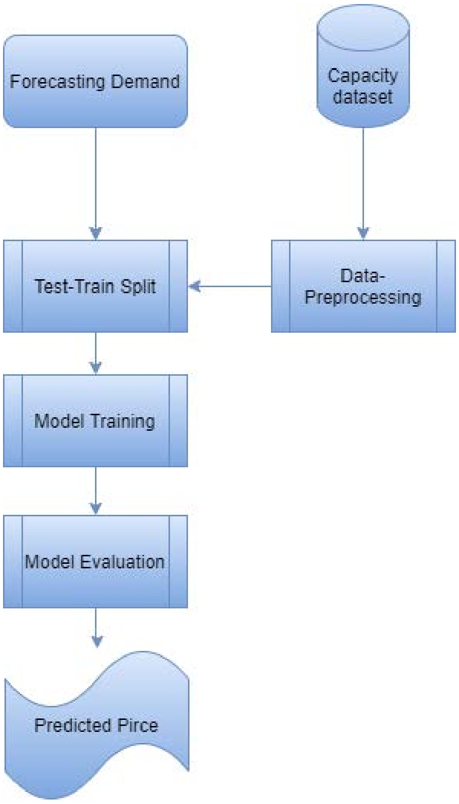
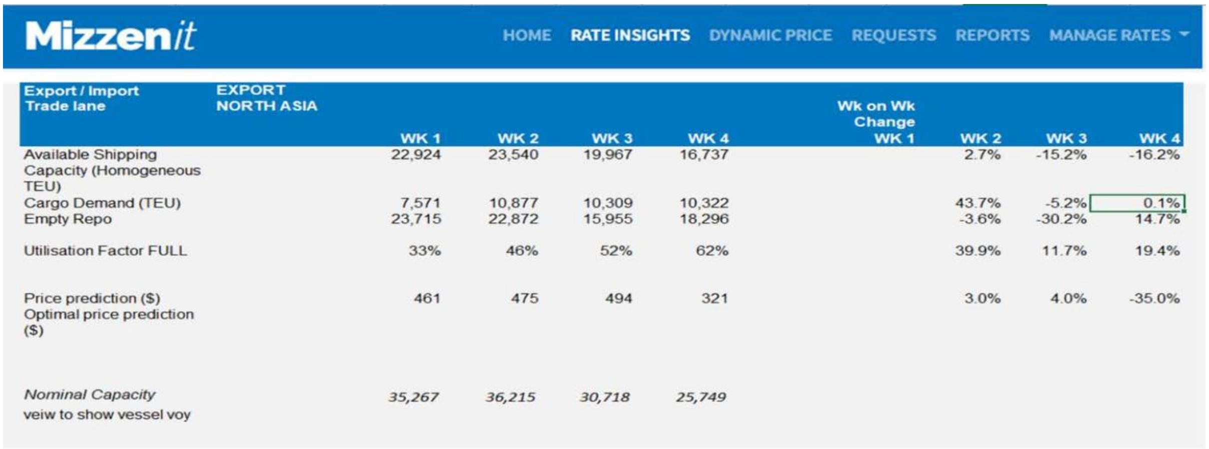


Fig. 9-6 Flow chart for Price Prediction Model

9.6 Mizzen Software User Interface:

The official paid software tool offered by Mizzen Group to its clients is shown in figure 9.7 given below.



Export / Import Trade lane	EXPORT NORTH ASIA	WK 1	WK 2	WK 3	WK 4	Wk on Wk Change WK 1	WK 2	WK 3	WK 4
Available Shipping Capacity (Homogeneous TEU)		22,924	23,540	19,967	16,737		2.7%	-15.2%	-16.2%
Cargo Demand (TEU)		7,571	10,877	10,309	10,322		43.7%	-5.2%	0.1%
Empty Repo		23,715	22,872	15,955	18,296		-3.6%	-30.2%	14.7%
Utilisation Factor FULL		33%	46%	52%	62%		39.9%	11.7%	19.4%
Price prediction ($)		461	475	494	321		3.0%	4.0%	-35.0%
Optimal price prediction ($)									
Nominal Capacity		*35,267*	*36,215*	*30,718*	*25,749*				
veiw to show vessel voy									

Fig. 9-7 Mizzen Software Design

9.7 Summary:

In this chapter, we presented a step-by-step procedure showing how to predict shipment demand and price respectively. We provided the step-by-step illustration of the models from which we can get shipment demand and capacity values and get predicted prices. Furthermore, each step's explanation is augmented by screen shots of software. In the next chapter, we conclude this thesis.

10 Conclusion and Future Work

10.1 Introduction:

This chapter concludes the thesis and provides some directions for future work. The aim of the thesis is to find data driven methods which can help support the shipping industry make informed pricing and supply chain decision resulting in increased business productivity. This study is to come up with new applications of machine learning algorithms in a new domain that is still missing AI and data driven methods for business improvement.

10.2 Thesis Conclusion:

In the first part of the system, we designed a novel feature selection framework that can seek information from domain expert in its selection process to provide a customized set of features. We concluded that the proposed methods work well for the application under observation as compared to the existing methods. The designed framework helped us in performing feature selection methods quick and easily in order to make use of the dataset for data driven methods for the shipping industry. Following the feature selection, we performed data pre-processing which is a necessary step to analysis the dataset. Based on the dataset statistics, the intelligent data driven models were implemented over the shipping dataset.

In the next stage of the thesis, we conducted a research study over the shipment spot pricing strategy. From the study, we designed a data driven method to determine spot price for the shipment. From the research study, it was concluded that the shipment price setting in the shipping industry is quite complicated and is highly seasonally driven. There is a disconnection between current shipment demand, available shipping capacity (supply) and shipment pricing. Hence, there is a strong need to explore the relationship between shipment demand, available shipping capacity (supply) and pricing to set optimal pricing for shipment containers. However, not much research has been done in the past to address this area. In order to fill this gap, we have conducted a research study for the Australian shipping industry. From the conducted research, it is evident that pricing is not dependent on shipment demand and available shipping capacity (supply), which is not the ideal case. There must be a positive relationship between these three factors. There is also a need to have a model to calculate spot shipment prices based on factors that affect pricing. We have proposed a novel mathematical model for setting container shipment spot shipment pricing based on shipment demand and available shipping capacity (supply). The results have shown that the proposed model is able to set prices based on market shipment demand.

In the next stage, we apply an already existing machine learning model over the shipping dataset to provide demand forecasting and price prediction capabilities to the Australian Shipping industry. The Australian shipping industry has no existing machine learning models being applied to their dataset in order to help them plan ahead and make informed marketing decisions for pricing and container shipment demand.

In this research, three regression-based ML models are applied on real-time datasets to predict the price of import shipment containers in the Australian shipping industry, specifically for the Asia-Oceania Trade Lane. The container price is predicted based on current demand and available supply. To analyze the performance of these models, R2 score, accuracy, and RMSE are measured. The evaluation results show that GBR performs best over the available dataset. The study of the literature reveals that there is scant work on price prediction for the container shipping industry. The shipping industry has not digitalized yield management and the current pricing practices of quoting freight rates with a date validity rather than specific vessel voyage result in a disconnect between price and supply and demand resulting in a sub-optimal pricing outcome. Thus, our study shows that using the ML algorithm for predicting shipment prices can positively affect the shipping industry. This research is a first-ever attempt to empower the Australian shipping industry with machine learning predictions.

10.3 Future Research Directions:

Based on the research problems, analyzed in the Australian shipping industry, a few research objectives were identified. Research has been conducted to provide possible solutions to the identified problems. However, there are some limitations of this research work due to research work completion time constraints. Thus, in order to pave path for future researchers, future work for each of the research objective is outlined as follows:

I. We have used only shipping capacity and demand to determine shipment cost in this research work. However, there is a capacity to improve the model by adding more factors that affect pricing such as oil prices, ship utilization factors, political situations such as the US-China trade lane, and environmental factors such as the outbreak of coronavirus which may affect shipping operations. Thus, this research provides a base for future research in the same area and much more sophisticated optimal shipment pricing models can be designed.

II. We have forecasted the shipment demand for a single trade lane's import. The forecasting can be expanded for other operating trade lanes for both imports and exports. Apart from other time series algorithms, deep learning models can also be applied to the dataset to determine their performance and find an even better performing forecasting model.

III. In this research, traditional regression-based models are used to perform price prediction for the first time in the Australian shipping industry. In the future, pricing may be predicted using deep learning models such as Long Short-Term Memory Recurrent Neural Networks

(LSTM). Also, we have worked on a single trade lane i.e., AU-Oceania imports only. In the future, the model can be extended for all trade lanes and for both exports and imports from Australia or even for all the operating trade lanes worldwide.

APPENDIX- A

RESEARCH PAPERS