Demand Forecasting in Production Planning for Dairy Products Using Machine Learning and Statistical Methods.

Mr. Chayuth Vithisoontorn

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2021

การพยากรณ์อุปสงค์ในการวางแผนการผลิตสำหรับผลิตภัณฑ์ที่ทำจากนมโดยการเรียนรู้ของเครื่อง และวิธีทางสถิติ

นายชยุตม์ วิทิตสุนทร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

Thesis Title	Demand Forecasting in Production Planning for Dairy Products Using Machine Learning and Statistical Methods.
By	Mr. Chayuth Vithisoontorn
Field of Study	Computer Science
Thesis Advisor	Professor PRABHAS CHONGSTITVATANA

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

.... Dean of the FACULTY OF ENGINEERING

(Professor SUPOT TEACHAVORASINSKUN)

THESIS COMMITTEE

.... Chairman

(Assistant Professor SUKREE SINTHUPINYO)

.... Thesis Advisor

(Professor PRABHAS CHONGSTITVATANA)

.... External Examiner

(Associate Professor Worasait Suwannik)

ชยุตม์ วิทิตสุนทร : การพยากรณ์อุปสงค์ในการวางแผนการผลิตสำหรับผลิตภัณฑ์ที่ทำจากนมโดยการเรียนรู้ของเครื่องและวิธีทางสถิติ. (Demand Forecasting in Production Planning for Dairy Products Using Machine Learning and Statistical Methods.) อ.ที่ปรึกษาหลัก : ศ. ดร.ประภาส จงสถิตย์วัฒนา

การพยากรณ์อุปสงค์เป็นส่วนงานหนึ่งที่สำคัญในการวางแผนการผลิตของทุกอุตสาหกรรม การพยากรณ์ที่มีประสิทธิภาพช่วยบรรเทาปัญหาสินค้าคงคลังเกิน, สินค้าคงคลังขาดและลดการสูญเสียรายได้ งานวิจัยนี้ใช้วิธีการพยากรณ์โดยวิธีทางตรง (Direct approach) เพื่อพยากรณ์อุปสงค์ของผลิตภัณฑ์นมทั้ง 8 ผลิตภัณฑ์จาก 5 โรงงาน จากข้อมูลที่มีทั้งหมด 5 ปี โดยใช้วิธีการทางสถิติที่เป็นที่นิยมและวิธีการโครงข่ายประสาทเทียมแบบลึก (Deep learning) คือ ARIMA และ LSTM. โดยเปรียบเทียบการพยากรณ์ในหลายมิติ ดังนี้ เปรียบเทียบระหว่างการใช้จุดข้อมูลเป็นรายเดือนและรายสัปดาห์ เปรียบเทียบระหว่างวิธีการตัวแปรเดียวและหลายตัวแปร เปรียบเทียบระหว่างวิธีการทางสถิติและวิธีการทางโครงข่ายประสาทเทียมแบบลึก โดยใช้เกณฑ์ชี้วัดทั้งสองประเภท คือ เกณฑ์จากความคลาดเคลื่อนจากการทำนายและเกณฑ์จากมุมมองทางธุรกิจ ผลลัพธ์จากการทดลองชี้ว่าทั้งวิธีการทางสถิติและวิธีการโครงข่ายประสาทเทียมแบบลึกเหมาะกับการพยากรณ์อุปสงค์ทั้งคู่ โดยพบว่าโมเดล ARIMA มักจะพยากรณ์ออกมาเป็นเส้นตรงค่าเฉลี่ยของข้อมูล ทำให้เหมาะกับชุดข้อมูลที่ไม่ค่อยแปรปรวนมากและ LSTM จะพยากรณ์ตามลักษณะการขึ้นลงตามฤดูกาลของชุดข้อมูลและมีแนวโน้มการขึ้นลงของข้อมูลที่ชัดเจน การทำนายโดยใช้จุดข้อมูลเป็นรายเดือนให้ค่าความคลาดเคลื่อนที่น้อยกว่ารายสัปดาห์อันเนื่องมาจากลักษณะของข้อมูลรายเดือนมีความแปรปรวนที่น้อยกว่ารายสัปดาห์ ทำให้ง่ายต่อการทำนาย

สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์	ลายมือชื่อนิสิต	....
ปีการศึกษา	2564	ลายมือชื่อ อ.ที่ปรึกษาหลัก	....

6270055421 : MAJOR COMPUTER SCIENCE

KEYWORD: Demand Forecasting, Deep Learning, Long Short-Term Memory, LSTM, Autoregressive Integrated Moving Average, ARIMA, Time Series Forecast

Chayuth Vithisoontorn : Demand Forecasting in Production Planning for Dairy Products Using Machine Learning and Statistical Methods.. Advisor: Prof. PRABHAS CHONGSTITVATANA

Demand forecasting is an essential task in manufacturing of every industry. Efficient forecasting relieves the excessive stock and out-of-stock problem, reducing revenue loss. This research performs a direct multistep forecast approach of demand forecasting on 8 dairy products of 5 different dairy production plants with 5-year data. Widely used traditional statistical method and the state of the art deep learning method for sequence problems are picked. ARIMA and LSTM. The models are compared in many aspects, monthly observations against weekly observations, univariate against multivariate, and statistical against deep learning using model error and business metrics. The result shows that both statistical and deep learning method are reliable and are suitable to be used in demand forecasting. There is no single best optimization algorithm. ARIMAs predict the future in an average smoothed straight line. It shows the best result on few wavering series, whereas LSTMs predict the future value follow the seasonal of series. It beats ARIMAs on strong trend series. Training the model on monthly observations provides lower error score because of monthly series generally has lower fluctuation than weekly series which is easier to forecast.

Field of Study: Computer Science		Student's Signature ....
Academic Year: 2021		Advisor's Signature ....

ACKNOWLEDGEMENTS

Thank my supervisor, Prof.Prabhas Chongstitvatana , for patience, guidance, and support. I have benefited greatly from his wealth of knowledge which have been conveyed it to me in a simply manner and turned difficulties into ease. I am extremely grateful that he took me in as a student and continued to have faith in me over the years.

Thank my committees and chairman, Prof.Sukree Sinthupinyo and Asst. Prof. Worasait Suwannik . Your encouraging words and thoughtful, detailed feedback have been very important to me.

Thank Dairy Farming Promotion Organization Of Thailand (DPO) for generously providing the real-world dataset for the experiment.

Thank computer science classmate of Chulalongkorn university for reminding of important events and sharing useful information.

Lastly, Thank my family, for endless love, care and raising me to be me today.

Chayuth Vithisoontorn

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Chapter 1: Introduction

1.1 Introduction

With a rise in the number of competitors, entrepreneurs need to adapt their business to attract new customers and keep their existing guests, to expand or keep their market share, one needs to know the customer demand and handle them wisely, providing decent customer experiences from every aspect of a business. From the manufacturing aspect, the best thing we can do is satisfy the order demand of customers to reduce the churn rate. There are ways to suffice the demand: (1) applying minimum stock limit policy, (2) substituting customer order with similar products, (3) collecting the customer's demands at the lowest chain of the supply chain, (4) cooperating with the lower chain to know the customer demand, and last (5) forecasting the demand which is in the scope of this research. This study uses a real-world dataset for comparative analysis of demand forecasting methods between traditional statistical method and recent method: deep learning. Also, to investigate the feature candidates and provide insight from the analyzed historical and expected to be useful for dairy product forecasting or typical time series forecasting.

To assess the forecast difficult degree, we can refer to the book of forecasting principles and practice [1].They define the predictability of the demand into 4 major factors.

1) how well we understand the factors that contribute to it?
2) how much data is available?
3) how similar the future is to the past?
4) whether the forecasts can affect the thing we are trying to forecast.

For the first question, as far as we know, there exist a seasonal from open and close of schools, a seasonal from monthly end promotion sale and open and close of the Modern trade due to the pandemic.

For the second question, the data is intermittent sale data that is recorded for 5 years and the goal is to forecast 3 months ahead.

Next the third question, the future demand is not that fluctuate except only on new products and on discontinued products. However, the supplier, government policy and the surprised pandemics strongly disrupt the demand.

For the last question, unlike stock trading where the demand directly affects the price. The dairy product's price is non-volatile.

For the research, the amount of data available and the ability to collect external factor is the most crucial part. The demand prediction is still yet quite hard but possible to do so.

1.2 Objectives

- To investigate real-world demand forecasting problem solving using both statistical and machine learning methods.
- To compare between multivariate and univariate prediction methods.
- To compare between different frequency of the data points.
- To investigate the appropriate features for diary product demand forecasting.

1.3 Contributions

- The research demonstrates the utilization of the machine learning method to aid real-world production planning problems comparing with the traditional statistic method.
- The research studies and reveals the insight from the analyzed historical data
- The research studies the effect of frequency of the data points against prediction accuracy

1.4 Scope of works

- Daily sale transaction from eight different products from five plants are used as training and testing data to forecast three months ahead demand for each product.

- The model is trained and predict using local model approach method for each series.
- LSTM algorithms and ARIMA represent the Machine learning method forecasting and Statistic method forecasting, respectively.
- Monthly and weekly frequency data points are used

Chapter 2: Background

2.1 Time series Transformation

Time series is a series of data points where the time variable play a major role. It contains three characteristics: Trend, Seasonality, and cycle. The time-series data tend to have a repetitive pattern comparing to typical data. Understanding these data patterns is also like understanding the business. Identifying trend, seasonality and cycle lead to easier forecasting or business counter measure.

Trend

A trend is a long-term increasing or decreasing in the series. In other word, the direction of the data. For example, saying that the sale of recent years is higher than the older years is consider at a trend.

Seasonal

A seasonal is a pattern of increasing or decreasing of the series at a fixed or known period. For example, the increase in flower sale on the Valentine Day each year.

Cycle

A cycle is an increasing or decreasing of the series without fixed or known period. These changes are usually due to economic conditions. The cycle is happened less often than the seasonal.

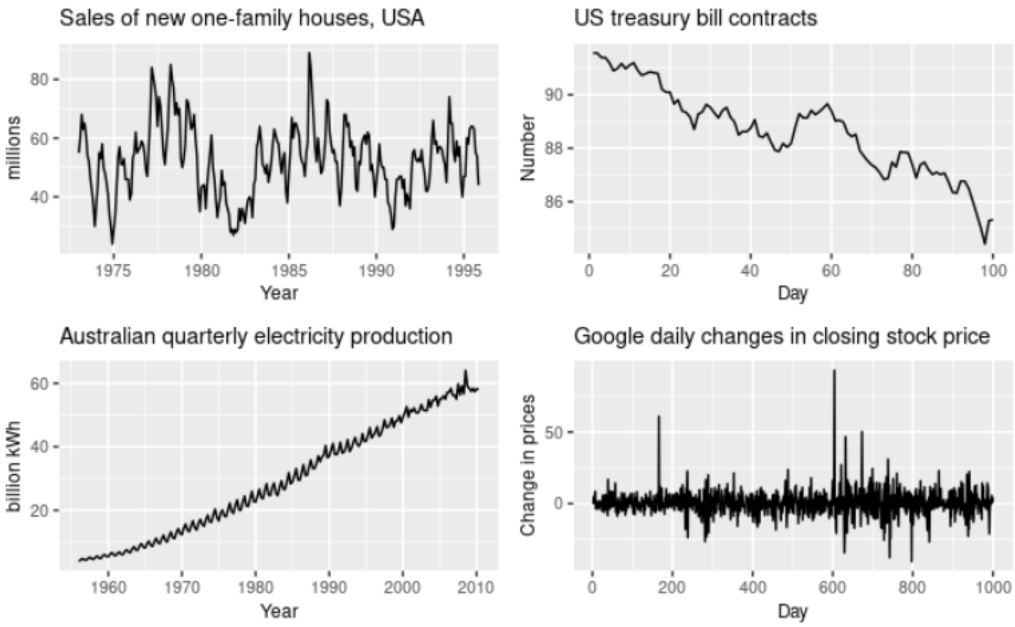


Figure 1 Time series data examples

One can simplify time series pattern by adjusting the raw series data to a simpler form. Known methods are calendar adjustments, popular adjustments, infatuation adjustments and mathematical transformation. The purpose of these adjustments is to remove the know variation sources so the model will only need to learn the unknown remainder sources. Simpler time series pattern led to easier to model and lead to more accurate forecast. This simpler state is often called the **stationary state** series where the data do not have a change in mean or variance over time. In this research, we use cox-box transformation to transform the data.

2.1.1 Augmented Dickey-Fuller Test (ADF)

A statistical method is used to check whether the series data has a unit root which is the property of non-stationary data. There are two values to check Dickey-Fuller statistical value (DF) and the p-value. if the DF value is more negative than critical values where p-value is the smallest significance level that null hypothesis would be rejected (5%).

2.1.2 One-parameter Cox-Box Transformation

The method is one of power transformation family. It transforms nonnegative data to have different scaling, stabilize the variance and make the data more normal distribution-like. The normal distribution like data is more valid to use in the statistical actions such as measure the association of Pearson correlation or any others statistical methods that required normalized data as an assumption.

The formula of cox-box transformation is given below.

$$x_i^{(\lambda)} = \begin{cases} \frac{{x_i}^{\lambda}-1}{\lambda} \ if\ \lambda \neq 0, \\ \ln(x_i) \ \text{if}\ \lambda\ = 0, \end{cases} \quad (1)$$

Where

- x_i is a value of observation i
- λ is an optival value ranging between [-5,5] which result in the best approximation of a normal distribute curve.

2.1.3 Min-Max Scaling

Min-Max scaling is one of the feature scaling methods. It is aimed to normalize the range of the input variables to having the appropriate range. Especially required in model training as many classifiers calculate the distance between two points and if any of the feature's range has broad range of values, the distance will be governed by this feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. Moreover, normalized the features helps the gradient descent converge much faster or speed up the model training

2.2 ARIMA

ARIMA is a one of the most wildly used statistical approach for forecasting the time series data. The domain of use cares is very diverse. Any time series problems are fit with the ARIMA such as future demand in production planning, future price of stock. ARIMA or "Auto-Regressive Integrated Moving Average" uses given past time series values, lags, and lagged forecast error to forecast the future values.

The pre-condition to using the ARIMA model is to prepare the non-seasonal time series values from the given time series. It requires the data not to have patterns and white noise.

ARIMA characterize by three terms: ARIMA (p, d, q)

where

- p (Auto Regressive: AR) is number of lag observations
- d (Integrated: I) is number of times requires to differentiate data to make it stationary.
- q (Moving Average: MA) is a size of the moving average or number of lagged forecast errors.

Then a linear regression model is constructed using the above-specified terms.

2.3 LSTM

LSTM is one of the machines learning based model. It is an improved version of recurrent neural network (RNN) to improve the system robustness and versatility [2]. RNN is a type of neural network model that can maintain an internal state to process a long sequence of inputs. Because it is expertise in handling sequential data, [3] it is popular among the video, NLP, and time series domains. However, during back propagation, RNN suffer from the vanishing gradient problem. Gradients are values used to update neural network weights. The vanishing gradient problem is when the gradient shrinks as it back propagates through time. If a gradient value becomes extremely small, it will not contribute too much learning. Earlier layers of RNN suffer most from the problems so the RNN could not have long term memory. Juergen Schmidhuber invented LSTM to solve this problem by introducing new gates for better control of gradient flow and better handling long-term sequence dependencies by enabling the network to either write, forget, or read the memory. Figure 2 and Figure 3 are shown to compare between RNN layers and upgraded version of RNN called LSTM layers

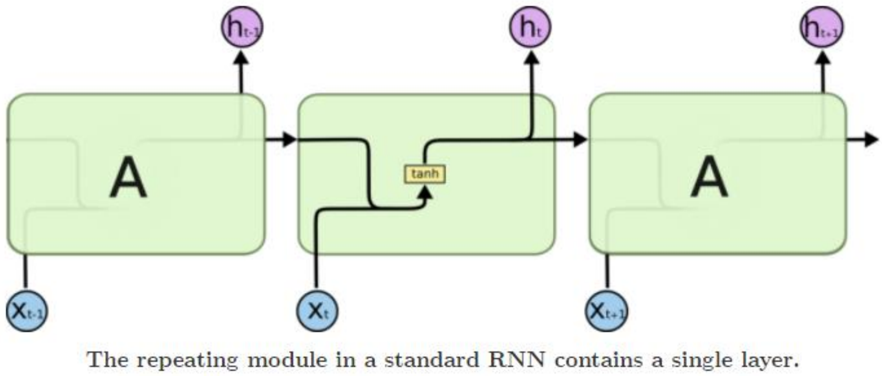


Figure 2 RNN layers

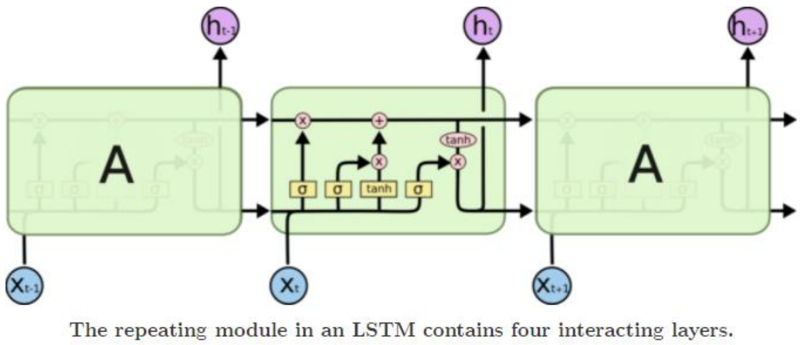


Figure 3 LSTM Layers

The core concept of the any recurrent neural network is simple. The data is transported from the start to the end of layer on the horizontal line and interact with different activation functions, gates, and operations along with their path depend on type of architecture. For LSTM architecture, 2 different type of activation functions are utilized. Tanh and Sigmoid, these activation functions construct the fundamental gates used in LSTM. Forget gate, Input gate, cell state and output gate

Tanh

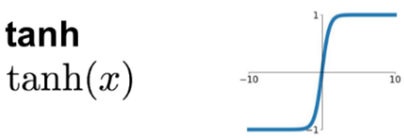


Figure 4 tanh function

The tanh activation is used to help regulate the values flowing through the network. The tanh function squishes values to always be between -1 and 1. The function prevent specific values from exploding and causing other values to become insignificant.

Sigmoid

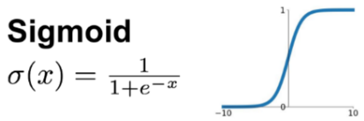


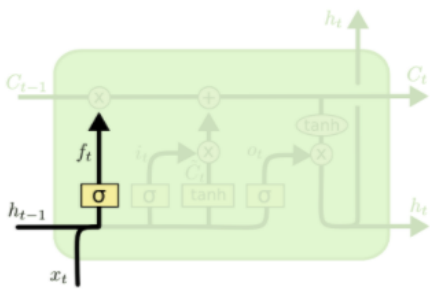
Figure 5 sigmoid function

A sigmoid activation is like the tanh activation. Instead of squishing values between -1 and 1, it squishes values between 0 and 1. That is helpful to update or forget data because any number getting multiplied by 0 is 0, causing values to disappear or be "forgotten." Any number multiplied by 1 is the same value therefore that value stay's the same or is "kept." The network can learn which data is not important therefore can be forgotten or which data is important to keep.

LSTM pseudo code

1. First, the previous hidden state (h_{t-1}) and the current input (x_t) are combined.

2. Combined is fed into the forget layer. This layer removes non-relevant data.

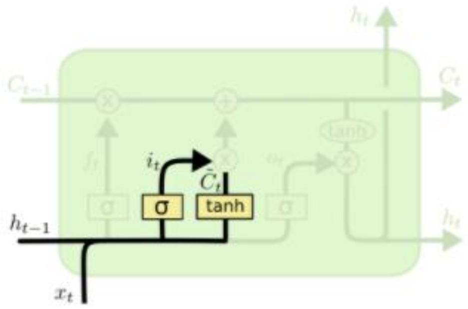


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Figure 6 Forget gate

3. A candidate layer is created using combined. The candidate (c_t) holds possible values to add to the cell state.

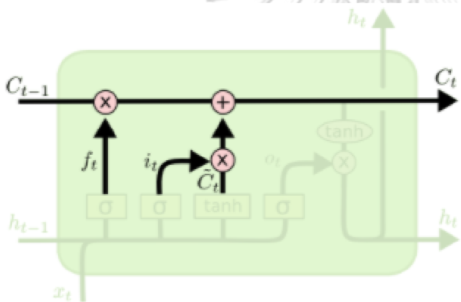
4. Combine (i_t) also gets fed into the input layer(σ). This layer decides what data from the candidate should be added to the new cell state.



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

Figure 7 Input gate

5. After computing the forget layer, candidate layer, and the input layer, the cell state (c_t) is calculated using those vectors and the previous cell state.

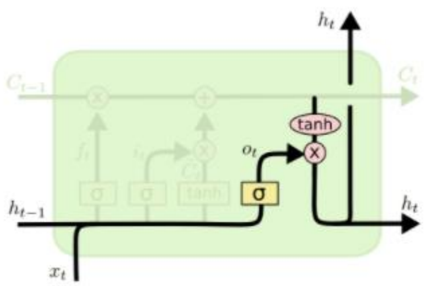


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Figure 8 Update cell state

6. The output (o_t) is then computed.

7. Pointwise multiplying the output (o_t) and the new cell state provides the new hidden state (h_t).



$$o_t = \sigma\left(W_o \ [h_{t-1}, x_t] \ + \ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Figure 9 Output gate

2.4 Persistence Algorithm (Naïve forecast)

Naïve forecast[4] is a forecasting technique which the last data points are used as current forecast without adjusting. This method is commonly used as a baseline for regression forecasting problems. Naïve forecast is a good technique for making a baseline forecast because

Simple: A method that requires little or no training or intelligence.

Fast: A method that is fast to implement and computationally trivial to make a prediction.

Repeatable: A method that is deterministic, meaning that it produces an expected output given the same input.

Naïve forecast is also used to overcome the difficulty in model evaluation of intermittent data by applying the **MAE** score of the method to the new evaluation formula named **MASE** [5].

2.5 Performance Evaluation

To complete the comparing analysis objective of the research, three error metrics are chosen; (RMSE, MAE and MASE) follow with two business metrics [6]; (fill rate and revenue loss). The MASE is introduced to overcome intermittent data where the data partly contains zeroes, scale invariance, interpretability and recommended to be used as forecast accuracy measurement. Because of the above benefits, the research will focus on this metric. The business metrics are simply used to show non-technical user of how the model behave.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)}{N}} \tag{2}$$

$$MAE = \frac{\sum_{i=1}^{N}|x_i - \hat{x}_i|}{\mathrm{N}} \tag{3}$$

$$MASE = \frac{MAE}{MAE_{naive}} \tag{4}$$

$$Fill\ rate = \frac{\sum_{i=1}^{N} \underset{+}{\min}\{x_i, \hat{x}_i\}}{\sum_{i=1}^{N} x_i} \tag{5}$$

$$Revenue\ loss = \sum_{i=1}^{N} \underset{+}{\max}\{(x_i - \hat{x}_i), 0\} \times\ price_i \tag{6}$$

Time series cross-validation

Time series cross validation [1] or walk forward validation is introduced to reduce bias in model evaluation. For common problems, many researchers use k fold-cross validation to evaluate model performance. The method equally split data into n fold, allocate one-fold for testing and others for training, iteratively shift the fold use for testing, ignoring the order of the data

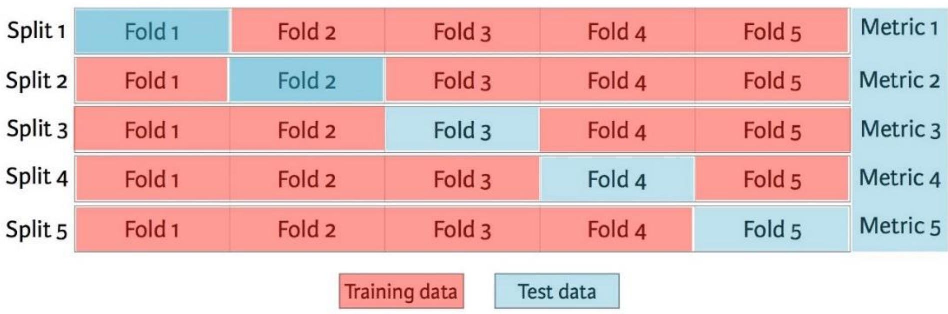


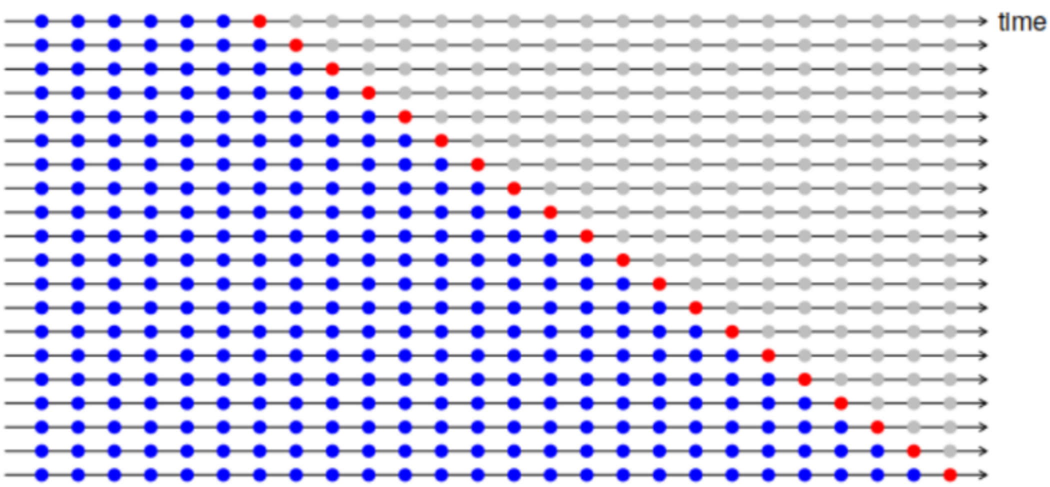
Figure 10 K fold cross-validation

Nevertheless, for the time series problem, the order of the data does matter because in typical, we use historical data to forecast the unseen future data. This is when walk forward validation is introduced as a validation method. The method imitates real-world time series forecasting by perform following steps

1) Reserve a minimum **n** observation as an initial training data
2) Use next observation **(n+1)** as a test data
3) Append a tested data to the training data
4) Repeat step 2 and 3 until it reaches the last observation is met

The size of training dataset can be either fixed window size (constant where the first observation shift to next observation for each iteration) or non-window size (increment where the first observation is always the first training observation.)

The following diagram illustrates the series of training and test sets, the blue observations form the training sets, and the red observations form the test set



Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.

Figure 11 Walk forward validation

Chapter 3: Literature Review

The arrival of the industry 4.0, the available of numerous digital data, the increase in competitors, cheaper but higher computing resources have made the machine learning approaches to the production planning. Many manufacturers began to adapt the machine learning to tackle the forecasting tasks. Nevertheless, the traditional forecasting methods are remaining used. Many studies compare the capability between those two. However, from the study reveal that both have their own stage to be a star performer. So, some invent the new approaches by combining them into the hybrid models to gain dominance from both methods. Usuga Cadavid, J.P., et al's machine learning in production planning survey [7] collect many related production planning researches, analyzed and reveal that only 75% of Production planning research domains are barely explored or not addressed as all. Their conclusion statistics show that RNN and LSTM have joined to the production planning tools party recently and they suggest categorizing the machine learning tasks for production planning into 11 activities.

Long short-term memory or LSTM is gaining the popularity over time, especially on time series domain. The LSTM is originated from Recurrent neural network or RNN. [2] Hochreiter, S. and J. Schmidhuber invented it since 1997 to overcome the flaw in RNN by introducing the either write, forget, or read allowing it to maintain their long memory, solving the vanishing gradient problem. Thanks to the easier accessible of high computing resources of current era, this invention is brushed up and start shrining.

Autoregressive Integrated Moving Average or ARIMA are a form of Box Jenkins models. The terms ARIMA and Box-Jenkins are sometimes used interchangeably. Statisticians George Box and Gwilym Jenkins developed systematic methods for applying them to business & economic data in the 1970's and named it Box-Jenkins models. [8]The model relies on assumptions past occurrences influence future ones. It has using three principles: autoregression, differencing, and moving average. Despite the old age, the model is believed to be the most used traditional statistical method and be counted as the top candidate of the time series forecasting problems.

Hyndman, R.J. and G. Athanasopoulos wrote a forecasting principle and practice book. [1]. They explain the time series fundamental and tips in the forecasting. which are very useful in doing the research, they cover all the necessary details from data exploration, data transformation, data modeling until the evaluation. They emphasize the use of time series cross-validation for train/test dataset split and evaluation which is more appropriate for time series data.

da Veiga, C.P., et al. compares the accuracy of demand forecasting in retail sales (ton) of liquid dairy products collected between 2005-2013 between statistical models and neural network-based models [9]. three evaluation metrics are used to compare: MAPE, fill rate and revenue loss. We took the fill rate [6]and revenue loss metrics to use in our research as well because they can illustrate the model performance in business manner. The result shows that neural network-based models outperform the statistical model. Although previous research show that statistical model got beaten but, [10] Fattah, J., et al. prove that ARIMA model can be used to forecast the demand for food manufacturing. Another study of single model forecasting is performed by Chniti, G., H. Bakir, and H. Zaher [11]. Two neural network-based models are chosen: Support vector machine regression (SVR) and Long Short-Term Memory (LSTM). Both univariate and multivariate are used. The dataset is daily phone prices from amazonfr marketplace collected for 4 years. The result reveal that LSTM perform better than SVR in Multivariate match but got beaten in univariate fight. The average multivariate error is lower than univariate. They noticed that LSTM model prediction is very accurate when the time series is stationary but does not provide the perfect solution if the series is fluctuated or suddenly peaks.

Recent studies began to use the hybrid model in prediction.[12] Fan, D., et al. use ARIMA, LSTM and hybrid between them to predict 3 group of well production. For single model, LSTM provides slightly better error than ARIMA but with less non-linear or smoother declining trend, ARIMA clearly outperform the LSTM. This research fulfills their objection perfectly because using hybrid model provide the best scores for all the series. Another usage of hybrid model is studied by Livieris, I.E., E. Pintelas, and P. Pintelas[13]. They show that LSTM perform the best among other single models: Neural network and Support vector machine regression with gold price prediction problem. Unlike the previous research, they use the couple of 2 machine learning method to improve the prediction: LSTM and CNN. The CNN is used specially to preprocess the data, filter the input, and extract the useful feature before passing to the LSTM layer. The duo between these two acquire the best prediction accuracy.

Apart from focusing on the model, other suggestions from the research are helpful. Ren, S., H.-L. Chan, and T. Siqin [14] suggest using both qualitative and quantitative methods for the demand forecasting. Firstly, predict the result with the model then follow with a judgmental forecast by a domain expert. It provides a method to capture fast-changing market information problems. Cadavid, J.P.U., et al., Hyndman, R.J. and G. Athanasopoulos [1] suggest using number of days in month and warn us to consider the factors we normally overlook such as calibrate of inflation, or population size in order to reduce bias in the forecasting model. Hyndman, Rob J., and Anne B. Koehler [5] proposed a new easily interpretable measure named Mean Absolute Scale Error (MASE). Their error metrics have many advantages and overcome the difficulty of measuring the intermittent data and the difference in scale of the series. Molnar, Christoph.[15] Explains the importance of model interpretation and provide interpretation guideline.

Chapter 4: Research Method

We applied the research methods listed in the survey [7] for this work.

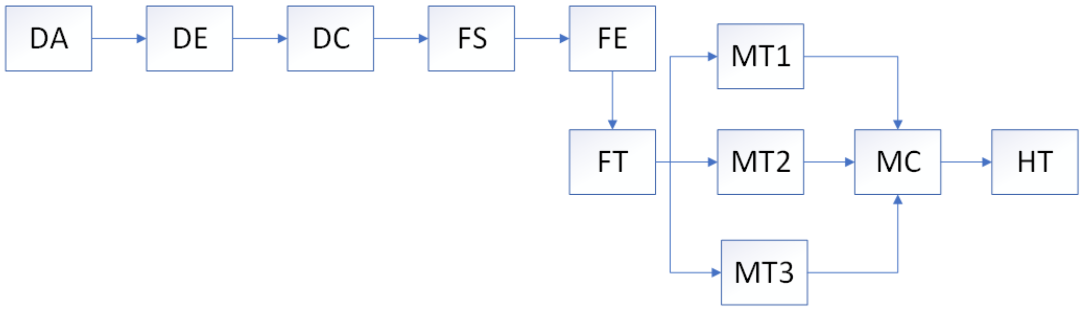


Figure 12 Research methods workflow

4.1 Data Acquisition (DA)

The diary sale transaction data are collected from the ERP system of the Dairy Farming Promotion Organization of Thailand (DPO) within period Oct 2016 – Sep 2021 (5 financial years) recorded by 5 plants. Over 100 products are on sale, but only 25 products require the forecast because they cover 90% of the total revenues. The series with less than 5 missing monthly observations and products produced by all plants is selected. 8 products remain or a total of 40 different series are remained. Sale volume of the series is highly fluctuated ranging from 10 unit/month to over 800 thousand unit/month. The holiday data are collected from the Bank of Thailand.

4.2 Data Exploration (DE)

The sale in volume of each series is explored to check seasonality, trend, outlier, and their statistical values. The exploration provides benefits in two aspects. One help in creating the prediction models and another provide the insight data to the business user. From the exploration, we can highlight traits of single variables, and reveal patterns and relationships between variables. With the patterns, we can select the appropriate model and fine tune the modeling to fit the data and with the relationships we can answer the business question and help provide the business strategy.

Table 1 List of raw data

Features	Sources	Description	Selected
Sale in volume	ERP System from production plant	Sale in volume	✓
Sale in ton	ERP System from production plant	Sale in ton	
Sale price	ERP System from production plant	Sale price (May contain 0 and negative values)	✓
Customer	ERP System from production plant	Buyer name	
Customer type	ERP System from production plant	Type of customer	
Customer province	ERP System from production plant	Buyer registerd province	
Plant	ERP System from production plant	Plant code	✓
Product name	ERP System from production plant	Name of product	✓
Product category	ERP System from production plant	Category of product	
Discount	ERP System from production plant	Reduction of price in percent	✓
Transactio n date	ERP System from production plant	ERP Transaction date. Can be either sale, return or give away.	✓
Holiday list	Bangkok of Thailand website	5 years Holidays date	✓

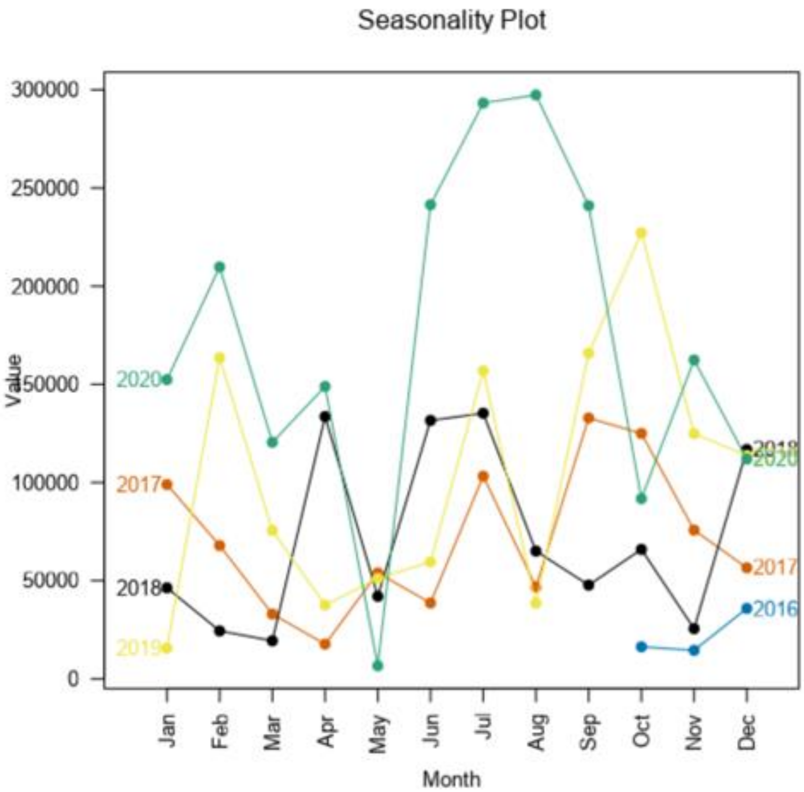


Figure 13 Sample of seasonality plot

Figure 13 shows sample of seasonality plot. Monthly aggregated sale in volume is plotted against each year to compare the rise and fall pattern of each year to identify the seasonal. We can get the insight of month impact to sale and provide countermeasure to it.

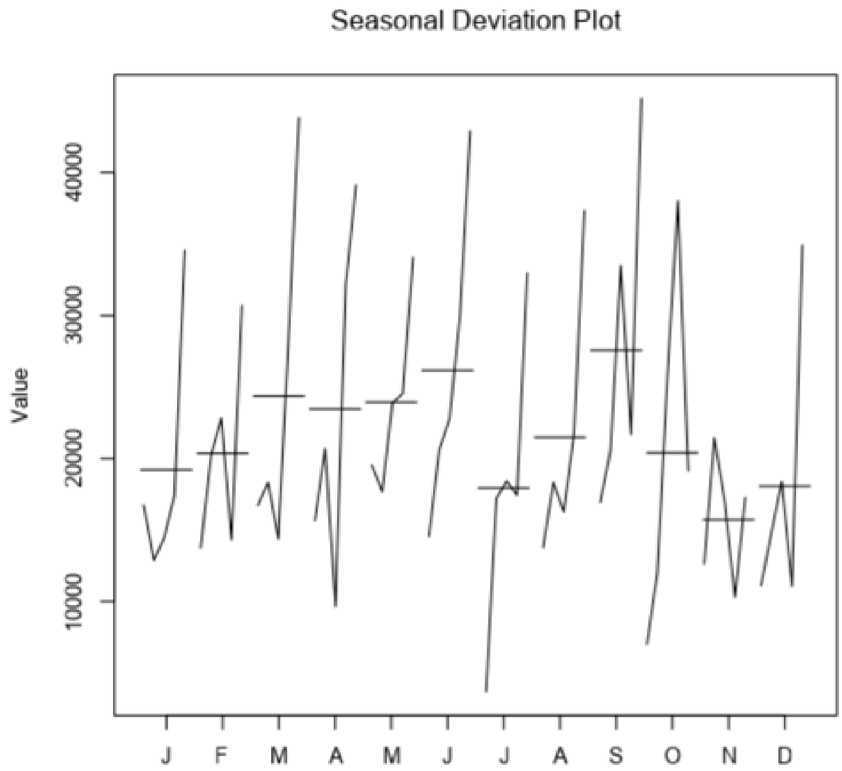


Figure 14 Sample of Seasonal Deviation Plot

Figure 14 shows sample of seasonal deviation plot. Monthly aggregated sale in volume is plotted consecutively to identify trend and the average value. The graph explains how to measure the performance of the business.

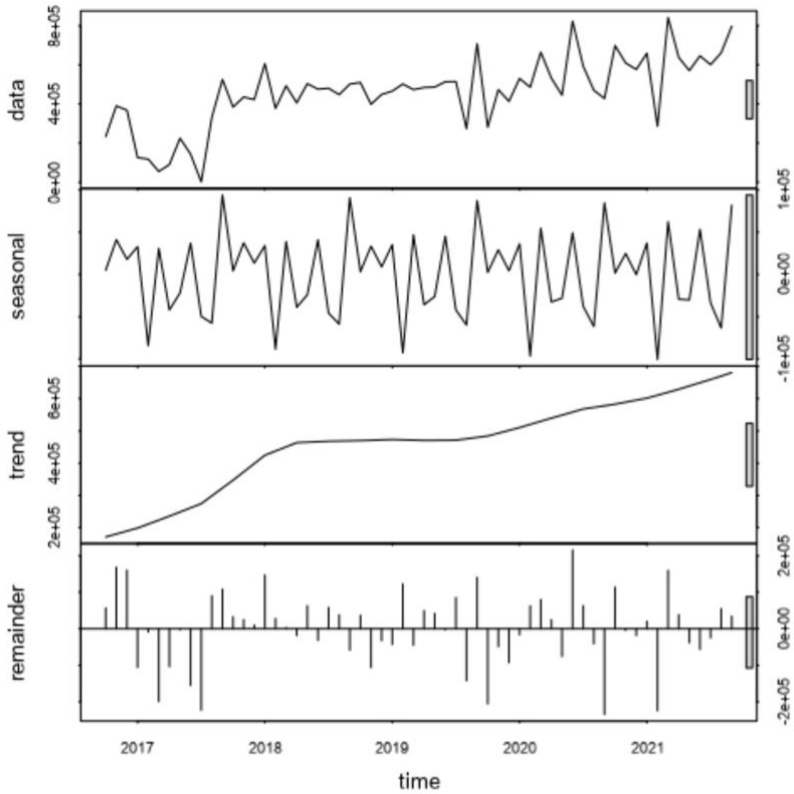


Figure 15 Sample of time series decomposition

Figure 15 shows sample of time series decomposition using loess method. The original series is decomposed into seasonal, trend and the remainder.

Max	Min	STD	MEDIAN	Mean	Kurtosis	Skewness	Product
7,427	4,470	1,069.394361309	5,291.5	5,671.5	-3.135294436	-0.186389214	ProductA
13,040	5,890	2,506.290918203	11,485	10,590.833333333	-3.163319000	-0.166569644	ProductB
136,678	36,106	38,352.856662401	108,937	98,474.833333333	-3.323155251	-0.018967261	ProductC
378,417	166,015	75,702.587260058	210,962	232,237.666666667	-2.961116353	0.208564645	ProductD
444,230	280,771	54,605.744943183	330,813.5	345,451	-1.658605474	0.836992184	ProductE

Figure 16 Statistical value of each product

Figure 16 shows sample of series statistical values. These values can also be used to analyze the characteristic of the series but is more into numerical manner. These numeric explains the distribution of values where we need to know this distribution to percept the fluctuation. We calculate and use these values twice. First with the original series to know the characteristic and second time to check the stationary state after applying the transformation.

4.3 Data Cleansing

To improve the quality of the model, the missing values and outlier values of original series are cleaned based on individual features and they are cleaned using mean and mode calculated from their own series.

Table 2 Cleansing policy

Features	Cleansing Policy	Description
Sale price	• Missing values are imputed with mode • 0 values are converted to small integer closer to 0	Sale price
Discount	• Missing values are imputed with mode	Reduction of price in percent
Holiday list	• Change to holiday flag 0 or 1	Holidays date
Transaction Date	• Manually check for invalid date format and clean case by case	Sale date

4.4 Feature Extraction and Selection (FE and FS)

The initial variables known as features are transformed to more powerful features. in feature extraction step. The extracted versions are then selectively picked using many criteria in feature selection step to filter only related useful predictor.

Table 3 Feature Extraction

Initial features	Extracted features	Description
Sale price>0	Average price per unit	Average of price per unit
Discount	Average discount percent	Reduction of price in percent

Sale price =0	Giveaway	Giveaway volume
Sale price <0	Return volume	Return volume
Holiday flag	Holiday count	Holiday count (month or week)
Date	Day count in month	Day count
Date	Financial year	Financial year
Date	Month	Month of year
Date	Week no	Week of month

Feature selection refers to techniques that select a subset of the most relevant features for a dataset. Fewer features allow machine learning algorithms to run more efficiently (less space or time complexity) and be more effective. Some machine learning algorithms can be misled by irrelevant input features, resulting in worse predictive performance. For this research, the processed series after cleansing is aggregated, test with 4 feature selection methods and follow with judgmental selection. The remained features are shown in Table 4 Feature selections where X denotes selected feature. The remained features are the survivor used for creating the prediction models for multivariate models.

Table 4 Feature selections

Features	Description	Week	Month
Average price per unit	Average of price per unit	X	X
Average discount percent	Reduction of price in percent		
Giveaway	Giveaway volume		
Return volume	Return volume		
Holiday count	Holiday count (month or week)		
Day count in month	Day count		
Financial year	Financial year	X	X

Month	Month of year		
Week no	Week of month	X	

4.4.1 Pearson correlation coefficient

Pearson's correlation is s a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations. The selection ratio is between -0.25 and 0.25

$$r = \frac{n(\sum(xy) - \sum x \sum y}{\sqrt{[n\sum_x 2 - (\sum x)^2][n\sum_y 2 - (\sum y)^2]}} \quad (7)$$

Where

r = Pearson correlation coefficient

n= number of observations

4.4.2 Spearman rank correlation

Unlike Pearson's correlation, spearman's correlation assesses monotonic relationships. The selection ratio is between -0.25 and 0.25

$$r_s = 1 - \frac{6\sum_d 2}{n(n^2-1)} \quad (8)$$

Where

r = Spearman rank correlation

n= number of observations

D= difference between the two ranks of each observation

4.4.3 Low variant Filter

The variance is a statistical measure of the amount of variation in the given variable. If the variance is too low, it means that the feature can be ignored. An input parameter is important only if its value changes significantly. The selected variance threshold is 0.16

$$\sigma^2 = \frac{\sum(x-\bar{x})^2}{n} \quad (9)$$

Where

σ = Variance

n= number of observations

4.4.4 Recursive Feature Elimination (Gradient boosting)

The method works by searching for a subset of features by starting with all features in the training dataset listed in Table 3 Feature Extraction. The elimination has the following pseudocode.

1) Select the initial features
2) Train the model using designed algorithms (Gradient boosting tree, Random Forest and Decision tree for this research)
3) Evaluate the model performance and rank the features by feature importance.
4) Remove the least important feature for desired number (remove by 1 per iteration for this research)
5) Repeat step 2 until the desired features are remained (3 features for this research)

4.5 Feature Transformation (FT)

The features are transformed into different spaces and scaling. The following steps are performed before feeding the data into training algorithms.

4.5.1 Power transform using Cox-Box Transformation

The series are transformed using cox-box transformation and standardizing to have zero mean and one std in attempt to remove trend and seasonal.

4.5.2 Convert time-series data to supervise format

For LSTM model, the shape of data from time series data needed to be converted to supervised data. To convert the data to supervised data, one step lookback into the past of each attribute is taken, added as attributes of current timestep and follow with the target value like the figure below.

Table 5 Supervised format

var1(t-1)	var2(t-1)	var3(t-1)	var4(t-1)	var5(t-1)	var1(t)
-0.103540	8.702170	5.283305	11	2016	-0.042834
-0.042834	9.666784	5.797149	12	2016	0.070879
0.070879	9.259424	5.606327	1	2017	0.048181
0.048181	9.693994	5.884194	2	2017	0.059133
0.059133	10.096175	6.107782	3	2017	-0.094139
-0.094139	10.464374	6.015370	4	2017	-0.004260

Where var1, var2, …, var n is nth attribute.

4.5.3 Min-max scaling

Rescaling the values of each attribute between [-1,1]. Figure 17 and Figure 18 show effect of the transformation by visualizing the series coup with Q-Q Plot which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. We can notice from the Q-Q Plot that the transformation makes the data closer to normal distributions (The blue observations are better aligned with diagonal line) in which imply the reduction of trend and seasonal.

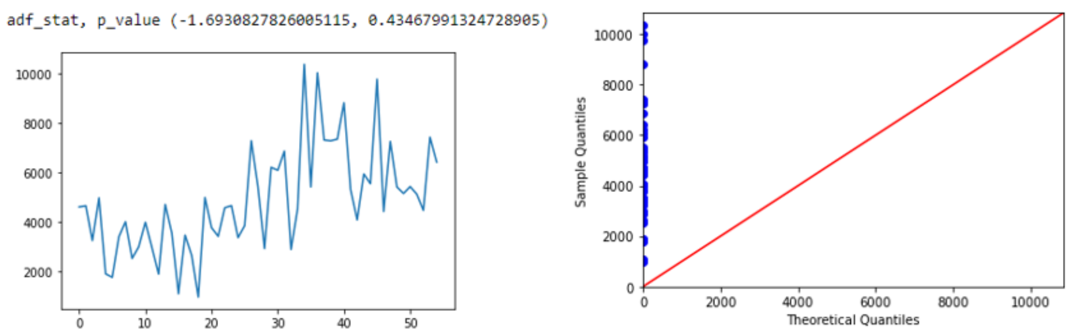


Figure 17 Sample of raw series and QQ Plot

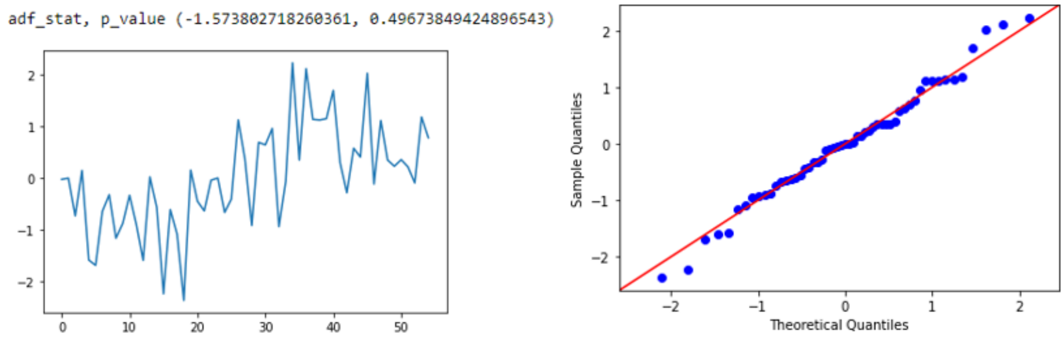


Figure 18 Sample of series after power transformation and QQ Plot

4.6 Hyperparameter Tuning (HT)

We use auto ARIMA and grid search for LSTM models to search for the best parameters of each model of each series.

4.6.1 ARIMA Hyper parameter

- p (Auto Regressive: AR)
- d (Integrated: I)
- q (Moving Average: MA)

4.6.2 LSTM Hyperparameter

- number of neural networks: Test parameters are 3,4,5 and 10
- number of epochs: Test parameters are 50 and 100
- learning rate: Test parameters are 0.005, 0.01 and 0.1

4.7 Model Training, validation, testing, and assessment (MT)

The first 4 years data are reserved as an initial training set and the latter year an as out-of-sample test set. Time series cross-validation [1] is used as an evaluation method. The method reuses tested observations as a trainset. It appends them to the original trainset at the end of each iteration. Sixes forecasting algorithms are used for each series, two statistical methods (ARIMA and ARIMAX) and four deep learning methods (Univariate/Multivariate LSTM and theirs Encoder-Decoder version)

Table 6 Research models

Model name	**Model type**	**Number of variables**
ARIMA	Statistical	Univariable
ARIMAX	Statistical	Multivariable (4)
LSTM	Machine learning	Univariable
LSTM	Machine learning	Multivariable (4)
Encoder decoder LSTM	Machine learning	Univariable
Encoder decoder LSTM	Machine learning	Multivariable (4)
Naïve (Persistent)	Baseline	Univariable

As the machine learning algorithms requires some randomization in training, the machine learning models are trained and evaluated repeatedly for 5 times and their median values are noted. In addition, all models are trained and evaluated twice for monthly observation and weekly observation respectively.

4.8 Model Comparison (MC)

For fair comparison, the models are trained using local approach method where each series are trained and evaluated separately because of ARIMA type models are not capable to be trained with global method models. Forecast values from weekly models are aggregated to monthly forecast value and then be compared with forecast values from monthly models.

Chapter 5 Result

There are 480 individual scores from total 40 series, 6 training algorithms and 2 types of observations so to publish the result, the average scores are presented instead to show the overall of how each model performed. The scores are then interpreted and led to more investigation and conclusion in both Modeling aspect and business aspect.

5.1 Average scoring

The scores of 40 series are computed into average scores for all chosen models.

Table 7 Average scores of monthly series

Model	**RMSE**	**MAE**	**MASE**	**Fill rate**
AR	21501.657	18390.864	0.817	0.821
ARX	21935.72	18843.995	0.83	0.816
U-LSTM	21509.296	18027.518	0.742	0.743
M-LSTM	20693.862	17240.801	0.694	0.763
EU-LSTM	20978.78	17511.478	0.732	0.744
EM-LSTM	21146.189	17835.905	0.697	0.76

Table 7 shows average scores of among each model. ARIMA (AR), ARIMAX (ARX), Univariate LSTM (U-LSTM), Multivariate LSTM (M-LSTM), Encoder decoder Univariate LSTM (EU-LSTM) and Encoder-decoder multivariate LSTM (EM-LSTM). Grey highlights mark the best score of each metric. M-LSTM performed best for all model error metric performance and ARIMA shows the best fill rate score. The reason why ARIMA acquired the best fill rate despite low MASE score is because of the model bias mentioned in section 5.5 Model bias and no penalty in fill rate formula for over forecasting.

5.2 Observation type comparison

MASE is a good choice for comparison metric as it is scale invariance and avoiding imbalance scaling because of zeros demand in intermittent case.

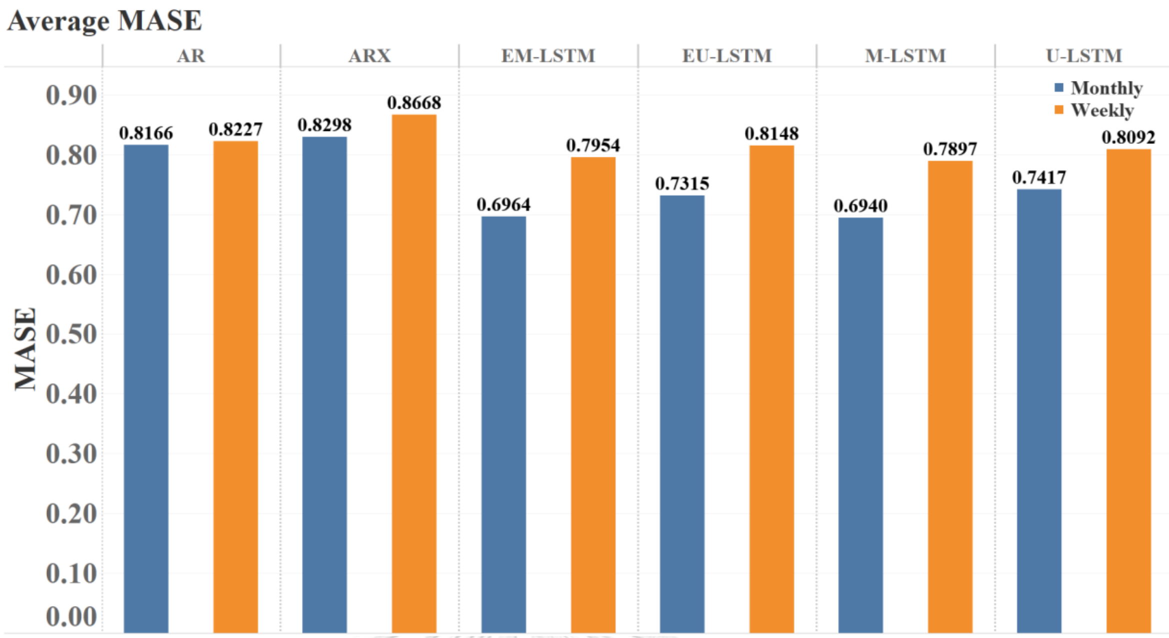


Figure 19 Average MASE score of monthly and weekly models

Figure 19 compares MASE score between each model and observation type. The orange bar denotes weekly observation and blue bar denotes monthly observation. All monthly models perform better weekly models. LSTMs clearly acquires better benefit from using monthly observations. Because intermittent data causes the under forecast bias in weekly observations LSTMs.

5.3 Model winner distributions

The research tries to figure out whether the machine learning model or statistical model will perform better with the chosen 40 dairy sale series. Below pie chart assists in answering the question.

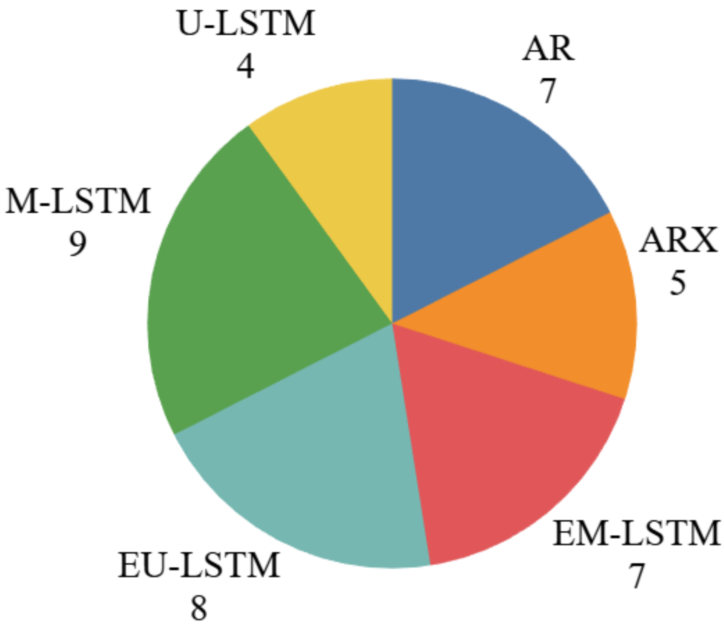


Figure 20 Model winner distributions

Notice from Figure 20, LSTM type models beat ARIMA type models with score 28:12. It is convinced that the LSTM models are more preferred by the sample series.

5.4 Unsatisfied predicted series

This section list all the series that both machine learning and statistical models perform worse than the base line Naïve method and find out the causes.

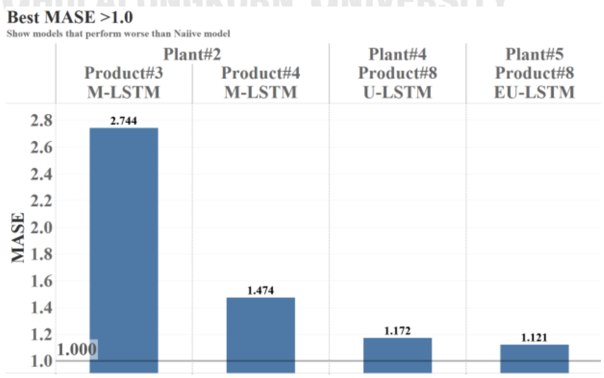


Figure 21 Series with MASE greater than 1.0 of monthly series

Figure 21 shows the worst perform series (MASE >1.0) and the least MASE on that series. There are 4 series out of 40 series. From those series, we investigated further and found an interesting reason below.

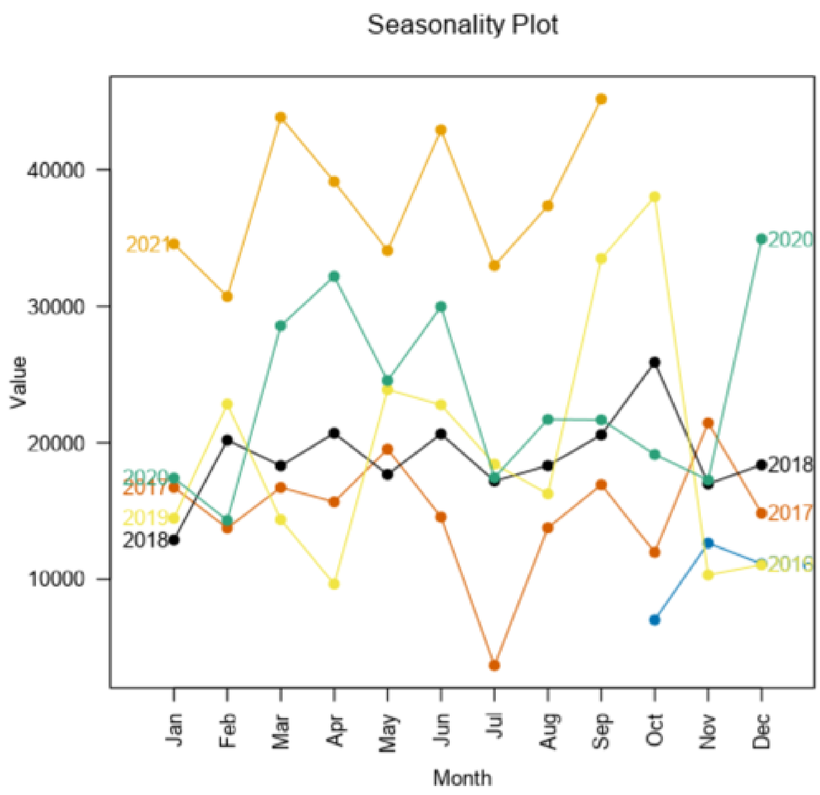


Figure 22 Seasonality plot between of one of low MASE series

From Figure 22, there is no overlap between 2021 seasonal line and any other seasonal line at all. It shows that on the concerned series 2021-year value is clearly an outlier compared to other years.

5.5 Model bias

This section aggregated the forecasting result from all models and identify the forecasting pattern of each model.

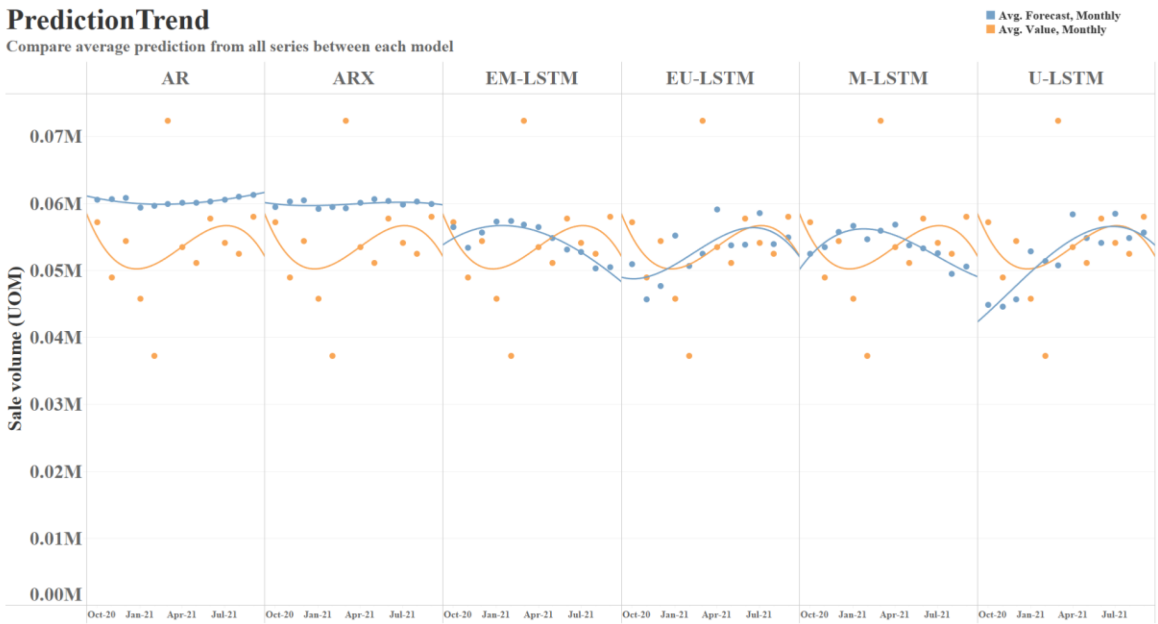


Figure 23 Actual values trend and predicted values trend of monthly series (Average)

From Figure 23, orange line and orange dot represent **actual trend line** and actual values. Blue line and blue dot represent **prediction trend** and predict values. LSTMs try to forecast to follow the seasonal and trend where ARIMAs predicted in straight line with average from past values. Overall forecast of ARIMAs is above the actual values. It explains why ARIMA models acquired higher fill rate ranks in Table 7 with lower average MASE scores.

5.6 Revenue loss

The research converts the error score into business score for convincing non-technical user to understand how the forecasting affects their revenues.

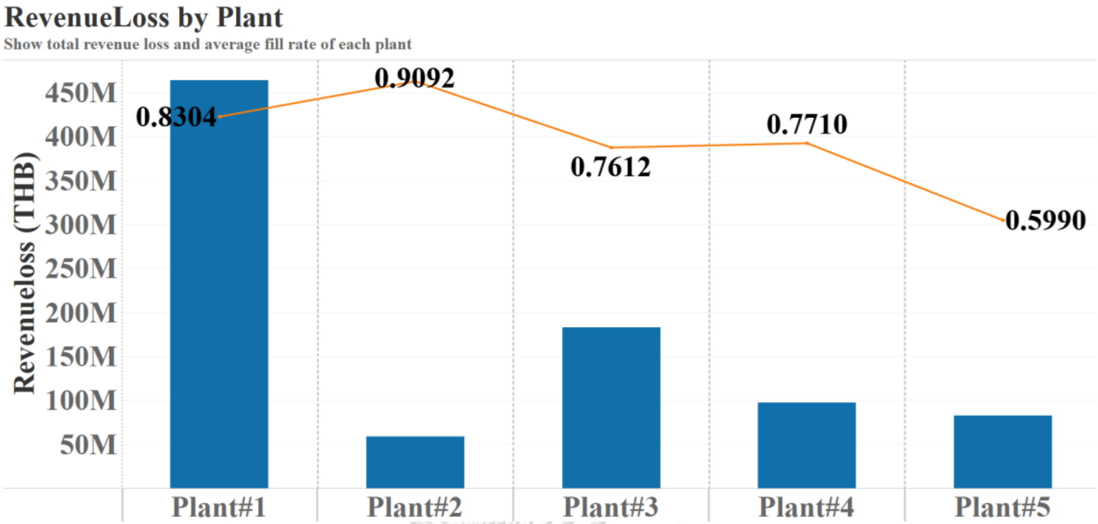


Figure 24 Total Revenue loss by plant and average fill rate by plant of a monthly series

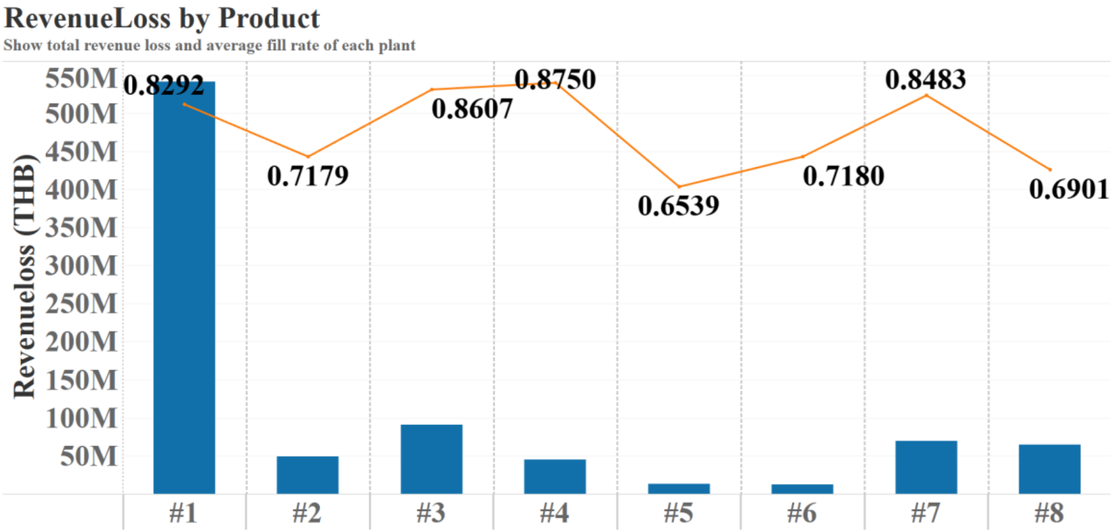


Figure 25 Total Revenue loss by product and average fill rate by plant of a monthly series

Figure 24 and Figure 25 shows possibility in improving the demand forecasting to reduce the revenue loss of each plant and product respectively. For plant perspective, plant#1 has the highest revenue loss (Blue bar) and plant#5 has the least fill rate (orange line) while product#1 has the highest revenue loss and product#5 has lowest fill rate for product perspective. There is a huge room for entrepreneur to investigate the causes of plant#1 and product#1 error and find the countermeasures.

Chapter 6: Conclusion

Although LSTM models acquired about 70% winner share from the experiment series but there is no single best algorithm for all the series. Using the right algorithm for the right situations is important. The ARIMA is a simple model that one can implement with a simple tool such as Excel. It has over forecast behavior, leading to a better fill rate than LSTM because of no penalty on over forecast. It assumes series are stationary, so it works well with unwavering series because of their prediction nature. In contrast, LSTMs predict in a riskier way by following seasonal and trends. This makes LSTM performs better with more complex series, with multivariate models due to a good feature candidate. Encoder-Decoder version of LSTM performs better in univariate LSTM but slightly worse in multivariate LSTM. ARIMA outperforms LSTM on most uptrend series and loses to LSTM on downtrend series or more complex series. Training the model with monthly observations provide a better result because of two main reasons. Monthly series is smoother and batch order frequency is greater than week causing the cavities in weekly observations known as intermittent data. For future work, applying penalty factor for case of over forecast to the formula of fill rate metric is an advice as it will be more concerned if the product's lifespan is short or there is a limit in storage space. Consider adding another candidate; seasonal naïve (previous year value as a forecast value) for benchmark is recommended. Apart from using national holiday as a feature candidate, school holiday is an interested candidate as most of the products aims the students as a main customer. [16] may aids the forecasting in different hierarchy such as product category level. Next, the attention-based mechanism [17] may provide better result intuitively that the model will focus more weight on recent data points. Finally, instead of using a local model for each series, the global model[18] shows equal or better performance compared to the local model and possibly aids outlier in seasonality.

6.1 Summary

1) The chosen statistical methods and machine learning methods are efficient enough to predict the demand forecasting problems. The evidence is based on MASE score. The 36 series out of 40 series got MASE lower than 1.0 which is good enough. To enhance the prediction accuracy further, more complicate methods can be used such as hybrid methods or more layers in neural networks.

2) Multivariate models provide better prediction accuracy than univariate model by about 7% error reduction. This is corresponded to the nature that knowing more factors increase the predictability.

3) Although the intuition that providing more data points to machine learning training method should produce better model but in practical, the result from Figure 19 is opposite because of the intermittent nature of the data causes the model learning from unfair data patterns.

4) The best features apart from the lag of the demand itself are price-related features. Picking one feature among price features are recommended to avoid overfitting from using too correlated features.

6.2 Limitation

1) The research is performed on a single laptop. Training and evaluate with 40 series take about 2 days to complete and it is even longer for weekly datapoint, so the range of hyper parameters are not broad and the sample series are limited (40 series). The specification of working environments is show below.

- CPU 8 core
- Ram 16 GB
- Windows 10
- Anaconda environment on Alteryx Designer

2) There is a fluctuation in demand due to the covid pandemics. Many distributers of the dairy products could not sell the product as usual. The data patterns are then messed up. The model could not learn the actual patterns properly.

3) The data points are quite small (5 years of data) because we use local approach model and monthly data.

REFERENCES

1. Hyndman, R.J. and G. Athanasopoulos, *Forecasting: principles and practice*. 2018: OTexts.
2. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.
3. Sezer, O.B., M.U. Gudelek, and A.M. Ozbayoglu, *Financial time series forecasting with deep learning: A systematic literature review: 2005–2019.* Applied Soft Computing, 2020. **90**: p. 106181.
4. Brownlee, J., *Machine learning mastery.* URL: http://machinelearningmastery. com/discover-feature-engineering-howtoengineer-features-and-how-to-getgood-at-it, 2014.
5. Hyndman, R.J. and A.B. Koehler, *Another look at measures of forecast accuracy.* International journal of forecasting, 2006. **22**(4): p. 679-688.
6. Sobel, M.J., *Fill rates of single-stage and multistage supply systems.* Manufacturing & Service Operations Management, 2004. **6**(1): p. 41-52.
7. da Veiga, C.P., et al., *Demand forecasting based on natural computing approaches applied to the foodstuff retail segment.* Journal of Retailing and Consumer Services, 2016. **31**: p. 174-181.
8. Fattah, J., et al., *Forecasting of demand using ARIMA model.* International Journal of Engineering Business Management, 2018. **10**: p. 1847979018808673.
9. Chniti, G., H. Bakir, and H. Zaher. *E-commerce time series forecasting using LSTM neural network and support vector regression*. in *Proceedings of the international conference on big data and Internet of Thing*. 2017.
10. Fan, D., et al., *Well production forecasting based on ARIMA-LSTM model considering manual operations.* Energy, 2021. **220**: p. 119708.
11. Livieris, I.E., E. Pintelas, and P. Pintelas, *A CNN–LSTM model for gold price time-series forecasting.* Neural computing and applications, 2020. **32**(23): p. 17351-17360.
12. Ren, S., H.-L. Chan, and T. Siqin, *Demand forecasting in retail operations for fashionable products: methods, practices, and real case study.* Annals of Operations Research, 2020. **291**(1): p. 761-777.

13. Usuga Cadavid, J.P., et al., *Machine learning applied in production planning and control: a state-of-the-art in the era of industry 4.0.* Journal of Intelligent Manufacturing, 2020. **31**(6): p. 1531-1558.

14. Molnar, C., *Interpretable machine learning*. 2020: Lulu. com.

15. Punia, S., S.P. Singh, and J.K. Madaan, *A cross-temporal hierarchical framework and deep learning for supply chain forecasting.* Computers & Industrial Engineering, 2020. **149**: p. 106796.

16. Qin, Y., et al., *A dual-stage attention-based recurrent neural network for time series prediction.* arXiv preprint arXiv:1704.02971, 2017.

17. Montero-Manso, P. and R.J. Hyndman, *Principles and algorithms for forecasting groups of time series: Locality and globality.* International Journal of Forecasting, 2021. **37**(4): p. 1632-1653.

VITA

NAME	Chayuth Vithitsoontorn
DATE OF BIRTH	12 July 1992
PLACE OF BIRTH	Phetchaburi, Thailand
INSTITUTIONS ATTENDED	Bachelor’s Degree in Computer Engineering 2014 Master Of Science, Computer Science, Chulalongkorn University 2022
HOME ADDRESS	22/92, Sukhaphiban 3 Mansion, Ramkhamhaeng 58/4 Alley, Hua Mak, Bang Kapi District, Bangkok 10240