Prediction Models for Beauty Products Monthly Sales with Price Promotion

Miss Nichakan Phupaichitkun

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Industrial Engineering
Department of Industrial Engineering
Faculty Of Engineering
Chulalongkorn University
Academic Year 2023

แบบจำลองการทำนายยอดขายรายเดือนของผลิตภัณฑ์ความงามที่มีการส่งเสริมการขายด้านราคา

น.ส.ณิชกานต์ ภู่ไพจิตร์กุล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมอุตสาหการ ภาควิชาวิศวกรรมอุตสาหการ
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2566

Thesis Title	Prediction Models for Beauty Products Monthly Sales with Price Promotion
By	Miss Nichakan Phupaichitkun
Field of Study	Industrial Engineering
Thesis Advisor	Associate Professor NARAGAIN PHUMCHUSRI, Ph.D.

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

..... Dean of the FACULTY OF ENGINEERING
(Professor SUPOT TEACHAVORASINSKUN, D.Eng.)

THESIS COMMITTEE

..... Chairman
(Associate Professor DARICHA SUTIVONG, Ph.D.)

..... Thesis Advisor
(Associate Professor NARAGAIN PHUMCHUSRI, Ph.D.)

..... Examiner
(Assistant Professor NANTACHAI KANTANANTHA, Ph.D.)

..... External Examiner
(Assistant Professor Siravit Swangnop, Ph.D.)

ณิชกานต์ ภู่ไพจิตร์กุล : แบบจำลองการทำนายยอดขายรายเดือนของผลิตภัณฑ์ความงามที่มีการส่งเสริมการขายด้านราคา. (Prediction Models for Beauty Products Monthly Sales with Price Promotion) อ.ที่ปรึกษาหลัก : รศ. ดร.นระเกณฑ์ พุ่มชูศรี

กลยุทธ์การตั้งราคาเพื่อการส่งเสริมการตลาด (Promotional pricing) เป็นเครื่องมือทางการตลาดสำคัญสำหรับผู้ค้าปลีกส่วนใหญ่ อย่างไรก็ตาม การคาดการณ์ยอดขายเมื่อมีการเสนอส่วนลดอาจเป็นเรื่องยาก เนื่องจากมีปัจจัยอื่นๆ ที่ทำให้ปริมาณความต้องการมีความไม่แน่นอนหรือมีความผันผวนสูง งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อศึกษาแบบจำลองการคาดการณ์ที่เหมาะสมที่สุดสำหรับปริมาณยอดขายของผลิตภัณฑ์ความงามในการร้านค้าปลีกและศึกษาผลกระทบของปัจจัยที่ส่งผลต่อยอดขาย โดยชุดข้อมูลที่ใช้ในการศึกษาคือข้อมูลยอดขายรายเดือนของสินค้าความงามจากบริษัทกรณีศึกษาตั้งแต่เดือนมกราคม 2563 ถึงเดือนธันวาคม 2565 เป็นจำนวนรวมทั้งสิ้น 36 เดือน ตัวแบบการพยากรณ์ที่ศึกษา คือ การวิเคราะห์การถดถอยเชิงเส้น (Linear regression) ตัวแบบการเรียนรู้ของเครื่อง ได้แก่ ตัวแบบการสุ่มป่าไม้ (Random forest) ตัวแบบ Extreme Gradient Boosting (XGBoost) ตัวแบบโครงข่ายประสาทเทียม (Artificial Neural Networks: ANN) และตัวแบบผสม ในการประเมินตัวแบบการพยากรณ์จะพิจารณาจากร้อยละค่าเฉลี่ยเคลื่อนที่สมบูรณ์ (MAPE) และเปรียบเทียบประสิทธิภาพโดยรวมด้วยร้อยละค่าเฉลี่ยเคลื่อนที่สมบูรณ์แบบถ่วงน้ำหนัก (WMAPE) นอกจากนี้ปัจจัยที่ใช้ในตัวแบบการเรียนรู้ของเครื่องพิจารณาทั้งที่ใช้ตัวแปรอิสระทั้งหมดหรือใช้ปัจจัยที่ได้จากวิธีการวิเคราะห์สมการถดถอยแบบเป็นขั้นตอน รวมถึงพิจารณาหรือไม่พิจารณาปัจจัยของผลิตภัณฑ์อื่นที่อยู่ในกลุ่มเดียวกันโดยจัดกลุ่มสินค้าตามประเภท (Category) ประเภทย่อย (Subcategory) หรือวิธีแบ่งกลุ่มแบบเคมีน (K-means clustering) จากผลการศึกษาพบว่าตัวแบบผสมของ Random forest และ XGBoost มีความแม่นยำในการทำนายสูงที่สุดโดยมีค่า WMAPE อยู่ที่ 27.65% ซึ่งน้อยกว่า WMAPE ของตัวแบบตัวแบบ Random forest 0.5% แต่ใช้เวลาในการประมวลผลนานกว่าประมาณ 5 เท่า การศึกษาครั้งนี้จึงเลือกตัวแบบ Random forest เป็นตัวแบบที่เหมาะสมที่สุด และเมื่อพิจารณาปัจจัยที่ส่งผลต่อยอดขาย พบว่าปัจจัยช่วงเดือนที่มีการส่งเสริมการขาย (Promotion period) มีความสำคัญมากที่สุด รองลงมาคือปัจจัยเปอร์เซ็นต์ส่วนลดและราคาขาย

สาขาวิชา วิศวกรรมอุตสาหการ ลายมือชื่อนิสิต

.....

ปีการศึกษา 2566 ลายมือชื่อ อ.ที่ปรึกษาหลัก

.....

6470340421 : MAJOR INDUSTRIAL ENGINEERING
KEYWO Model prediction, Machine learning, Hybrid model, Retail
RD:

Nichakan Phupaichitkun : Prediction Models for Beauty Products Monthly Sales with Price Promotion. Advisor: Assoc. Prof. NARAGAIN PHUMCHUSRI, Ph.D.

Promotional pricing strategy is a major marketing tool for most retails. However, predicting sales when discount is offered can be difficult since there are other factors causing demand to be uncertain or highly fluctuating. The objective of this research is to identify the most suitable prediction models for beauty product unit sales in retail and capture the effects of factors impacting sales. The dataset provided by the case study retail company was available from January 2020 to December 2022 (36 months). The prediction models, including linear regression, random forest, XGBoost, artificial neural networks (ANN), and hybrid models, are constructed and evaluated using the mean absolute percentage error (MAPE). Then, to select the most appropriate model, the weighted MAPE was calculated and compared for overall performance. Moreover, factors used in machine learning models are either using all the independent variables or using significant factors from the stepwise method, and either considering or not considering factors of exogenous products in the same cluster grouped by category, subcategory, or K-means method. The result shows that the series hybrid model of random forest and XGBoost outperformed with a weighed MAPE of 27.65%, which had 0.5% lower weighted MAPE and around 5 times longer runtime than the random forest model. Thus, the most suitable model is the random forest model. Considering factors affecting sales, it was found that the promotion period factor was the most important, followed by discount percentage and price factors.

Field of Study:	Industrial Engineering	Student's Signature .....
Academic Year:	2023	Advisor's Signature .....

ACKNOWLEDGEMENTS

The completion of this thesis could not have been possible without the participation and support of so many people, whose names may not all be enumerated. Their contributions are sincerely appreciated and truly acknowledged, particularly the following:

I would like to express my deep and sincere gratitude to my advisor, Assoc. Prof. Naragain Phumchusri, Ph.D., for giving help and support, providing valuable guidance, insights, and advice, as well as motivation.

Besides my advisor, I am grateful to the members of my thesis committee: Assoc. Prof. Daricha Sutivong, Ph.D.; Asst. Prof. Nantachai Kantanantha, Ph.D.; and Asst. Prof. Siravit Swangnop, Ph.D., for their helpful comments, feedback, and suggestions on this thesis.

Finally, my sincere thanks to all my relatives, friends, and others who, in one way or another, shared their support and kindness.

Thank you again for your help and support.

Nichakan Phupaichitkun

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1 Introduction

1.1 Background of Research

Retail businesses or retailers are one of middlemen in retail supply chain that connect suppliers or other intermediaries to end consumers by purchasing a large number of products from suppliers or other intermediaries and gradually selling goods or services directly to end consumers for their personal use, as shown in Figure 1. These businesses make it convenient for consumers to easily and quickly buy products in small amounts (Chiewpanich & Mokkhamakkul, 2019).

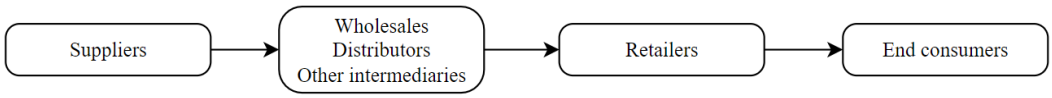


Figure 1 A simple retail supply chain
(Source: (Ayers & Odegaard, 2017; Development, 2021))

Retailers in Thailand formerly was dominated by small family-owned stores, called traditional trade, which obtained products from middlemen and distributors. Nowadays, the traditional retail stores format has transformed to modern stores or modern trades which are rapidly expanding throughout Bangkok and the non-urban locations. Various product categories can now be purchased from several different retail formats (Fox & Sethuraman, 2006). Economic Intelligence Center (EIC) of Siam Commercial Bank PCL (SCB) reported the annual retail markets sales in the Thailand since 2015 to 2022 as shown in Figure 2. For the first few years, it tended to increase continuously because of continued growth in the tourism sector and the investment in expanding branches. Although sales declined in 2020 due to the heavy impact of the coronavirus disease 2019 (COVID-19) outbreak leading to lockdowns and a changes of customer behaviors such as working from home and other online activities, the sales will be expected to grow continually because of support factors including the gradual increase in foreign

tourists coming to Thailand, the Thai government's subsidies and welfare programs, and high urbanization rates. As the result, business competition is likely to become more intense due to new entrants both domestic and international who see potential growth in Thailand retail and competitors from online stores (KResearch, 2022).

Figure 2 Annual retail markets sales in Thailand (Trillion bath)
(Source: (EIC, 2022))

Many retailers have tried to reduce the cost to gain more profits due to high competition in retail industry (Carter, 2019; Fox & Sethuraman, 2006). Moreover, they pay extra attention to product availability and customer satisfaction since they directly face customers. Therefore, it is important for retailers to understand what customers want, as well as when, where, and how much demand occurs (Wen et al., 2019).

One of approaches that could help businesses in handling these concerns is demand forecasting. An accurate sales forecasting model benefits in supply chain management, eventually boosting profitability. It also helps retailers to make promotion plans or strategies of products sold in stores by simulate the demand of different promotional mixes. On the other hand, poor sales forecasting may lead in an under- or over-stocking problem, consumer dissatisfaction, and impact financial gain. Therefore, in order to produce reliable and accurate forecasting

results, it is important to build effective sales forecasting models (Lu et al., 2012; Nunnari & Nunnari, 2017).

Demand forecasting has been studied and applied in many works on energy consumption (Ghalehkhondabi et al., 2017; Lahouar & Ben Hadj Slama, 2015), tourism and hotel demands (Archer, 1987; Laaroussi et al., 2023), fashion products (Loureiro et al., 2018; Nenni et al., 2013; Singh et al., 2019), health-care services (Azadi et al., 2023; Jones et al., 2009), retail sales (Almeida et al., 2022; Aye et al., 2015; Falatouri et al., 2022; Nguyen et al., 2023), etc. Prior research studied and compared the performance of many forecasting methods, including traditional statistical models (time series models or causal models like linear regression models), machine learning models (e.g., decision tree, random forest, extreme gradient boosting (XGBoost), support vector regression (SVR), neural networks, etc.), deep learning (e.g., LSTM), or hybrid models (e.g., clustering and regression models, linear and non-linear models, time-series and regression or machine learning models, etc.). According to the previous papers, there is no forecasting method that outperforms all other models under all conditions for all datasets (E et al., 2022; Fadillah et al., 2022). For multivariate problems, linear regression is usually used as the traditional model for prediction because it is a fast-to-fit method and easy to understand the relationship between variables and the predictions (Khan, 2020; Prabhakar et al., 2018; Wang, 2020). Unlike linear regression, machine learning methods can deal with big and complex non-linear datasets and provide highly accurate predictions (Kiran et al., 2022). Random forest regression and XGBoost have been widely studied for retail prediction because of their high forecasting accuracy and easy-to-interpret algorithms (Almeida et al., 2022; Mitra et al., 2022; Priyadarshi et al., 2019). Artificial neural networks (ANN) also have been used to forecasted or predicted product sales due to high accuracy (Auppakorn & Phumchusri, 2022; Güven & Şimşir, 2020).

However, retail product sales have highly fluctuating demands due to many factors, including price promotions, weather and seasonality, holidays and weekends, COVID-19, and economic indicators (Badorf & Hoberg, 2020; Huber & Stuckenschmidt, 2020; Ma & Fildes, 2021; Tian et al., 2021). Therefore, it could be a big challenge to forecast and predict retail demand.

The case study retailer is one of the modern trades in Thailand and sells an extensive variety of goods. The company now has many branches and plans to expand them to cover all provinces in Thailand. Currently, the company focuses on beauty products from seven categories, including facial moisturizers, body moisturizers, facial cleansing, hair coloring, sunscreen, body scrubs, and beauty accessories, as shown in Figure 3. The total number of products is 333.

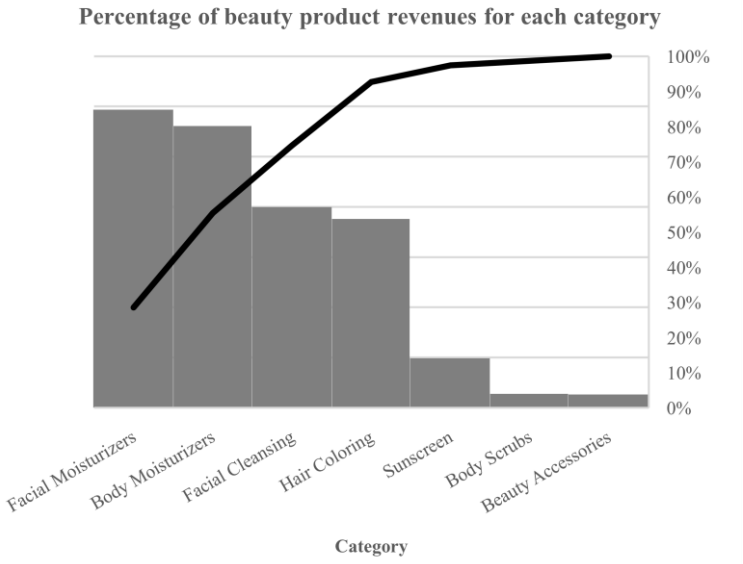


Figure 3 Percentage of beauty product revenues for each category

1.2 Problem Statement

For promotion planning, the company selects some of the products in some periods and offers short-term attractive promotions, such as price reductions, to stimulate demand for the products and boost sales by using experience and personal judgments.

Nevertheless, promotions may not always be effective on all products, as sales do not always increase when doing promotions. Figures 4 and 5 show examples of beauty products A and B for which promotion works well in increasing sales and a product for which promotion may not always work in increasing sales, respectively. In Figure 4, the product sales when promoted are significantly higher than when not

promoted. While Figure 5 shows that at the same reduced price, sales may increase as much as or more than when without promotion. From May 2021 and August 2021, sales declined, possibly as a result of the COVID-19 outbreak. Additionally, even at the same promotional price, sales may not always increase by the same amount due to other factors that may affect sales, including the Thai government's fund-subsidized or welfare programs, the COVID-19 pandemic, etc.

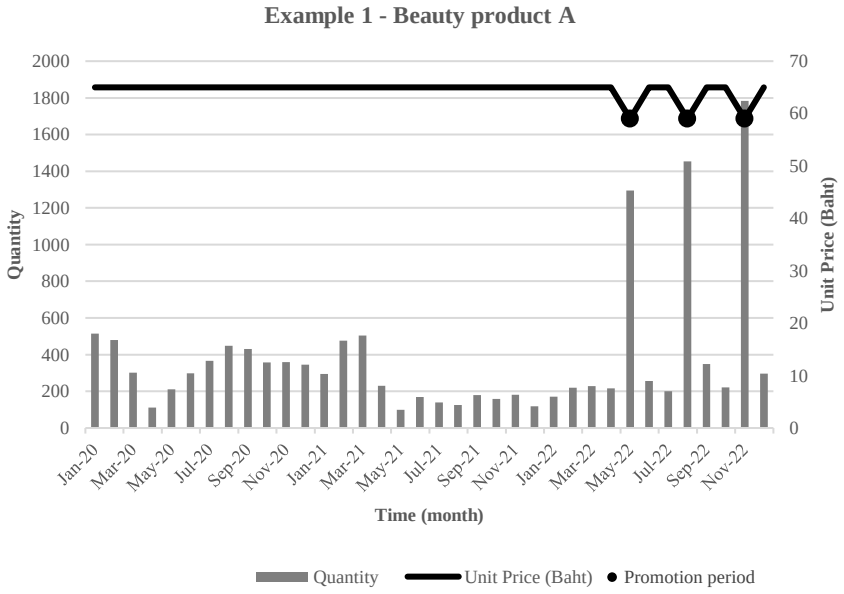


Figure 4 Example 1 – Sales of beauty product A with and without promotion

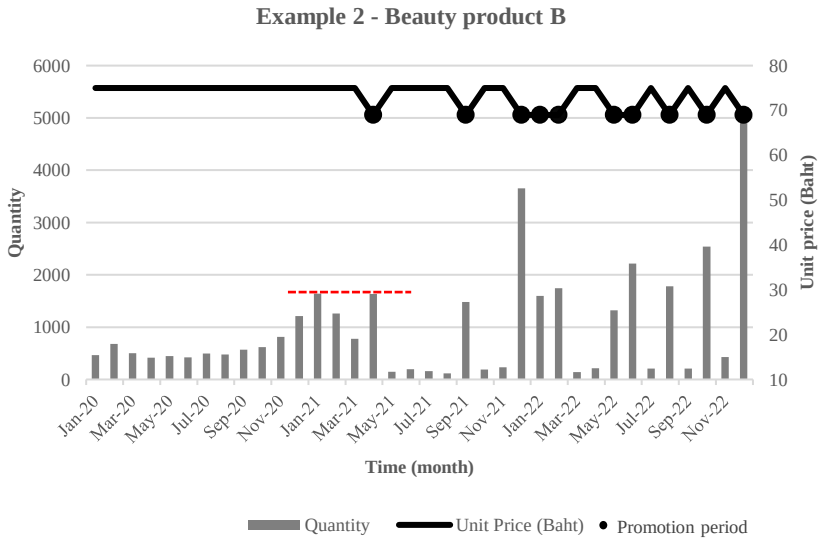


Figure 5 Example 2 – Sales of beauty product B with and without promotion

Therefore, the companies need sales predictions to understand demand behaviors of products with or without promotions or other sales-impacting factors in order to decide which goods are worth promoting to enhance sales and profitability in promotional planning.

However, the company currently does not have a sales prediction model to understand sales behaviors for products with or without promotions or other sales-impacting factors.

1.3 Research Objectives:

- To identify prediction models which can accurately predict monthly sales of beauty products sold in a retail offering price promotion.
- To determine exogenous variables which are significant for sales prediction model.

1.4 Scopes of Research:

- The top 10 best sellers in facial moisturizers category are considered in this study.
- The monthly sales dataset from a case-study retail company in Thailand from January 2020 to December 2022 for 36 months of beauty products from a case study retail company in Thailand is used, with 30 months for training the prediction models and the rest 6 months for testing the models' performances.
- The variables in this study include:
    - Independent variables or features: price, promotion characteristics (e.g., discount percentage, promotion period and lagged of promotion period), time period, number of stores, number of COVID-19

pandemic new cases growth in Thailand and the Thai government's subsidies and welfare programs.

- Dependent variable: sales quantity data is used.
- Other variables: price and promotion characteristics factors of other products in the same group.

- Prediction models are explored in this research including linear regression, random forest, XGBoost, ANN and hybrid models.
- The effects of clustering data before running models will be observed to determine whether clustering products can increase prediction accuracy.
- The 10-fold cross validation is used for hyperparameter tuning.
- The model performance is captured by weighted mean absolute percentage error (WMAPE) where the weight is determined by revenue of the products.

1.5 Research Outcomes:

- The accurate sales prediction model of overall top 10 best sales beauty products
- The important factors which affect sales of beauty products
- Relationship between exogenous variables and product's sales of the top 10 beauty products of the case study retailer

1.6 Benefits of this research:

- The accurate demand forecasting method helps them manage promotion planning for the retail business.
- The retail company can understand the important factors and relationships between variables and sales of the studied beauty products.

Chapter 2 Literature Review

2.1 Related Theory

2.1.1 Forecasting techniques

The forecasting techniques can be grouped into two types which are qualitative forecasting and quantitative forecasting (Thoplan, 2014). The qualitative methods are mainly based on judgments, opinions, or personal experiences from experts, executives, staff members or consumers. They are useful when there is a lack of historical data or an unavailable or inexistent data such as the data of new technology or new products, but they can be affected by personal biases. While the quantitative methods are based on mathematical models from existing data to determine or predict the outcomes. They help the business understand the relationship between dependent and independent variables and track patterns that appear over time or the possible impact on the business from the changes. However, historical data which available and enough is require. There are two types of quantitative methods: time series forecasting and causal method forecasting. Time series forecasting is models that predict the future based on their past data patterns such as trends, seasons or cyclicals and the occurrence of variables over time. Causal method forecasting assumes that the forecasted variable is related to other factors or variables. It takes a mathematical relationship between the dependent and independent variables to forecast the future.

This work focused on the quantitative forecasting methods used to predict the demand for beauty products in the retail case study.

2.1.2 Linear regression

Linear regression is a traditional, easy, and fast-to-fit method to study relationships between independent variables and dependent variable by fitting a straight line. The coefficients of the model were chosen using the ordinary least squares (OLS) method to minimize the sum of squares of error between the predicted

and actual values. The general equation can be expressed in Equation (1) (Khan, 2020).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon \quad (1)$$

where Y is the dependent variable

$X_1, \ldots, X_m$ are the independent variables

$\beta_0, \ldots, \beta_m$ are the coefficients calculated by the model

m is the number of independent variables

ε are the residuals

This method assumes the relationship between predictors and the target variable to be linear. The predictors are not highly correlated, and the residuals are independent and identically normal distributed. After fitting the model, the predictions were calculated.

2.1.3 Machine Learning

Machine learning (ML) is a subset of artificial intelligence (AI) that teaches machines how to learn and handle data automatically with minimum human intervention by using algorithms. In recent years, ML techniques have been broadly used in companies to help businesses increase sales and make better decisions. ML can generally be divided into three types (Auppakorn & Phumchusri, 2022; Mahesh, 2019), as follows:

- Supervised learning uses both input data and labels to develop a predictive model to predict discrete output in classification problems or continuous output in regression problems.
- Unsupervised learning uses an unlabeled dataset to analyze and identify some pattern or structure in the data for clustering or association.
- Reinforcement learning is based on trial and error and taking actions aiming to obtain the reward.

In this paper, the supervised machine learning algorithms were considered as follows:

2.1.3.1 Random Forest (RF)

Random Forest is a supervised learning algorithm that can be used for both classification and regression problems. The approach uses the bagging ensemble technique to randomly choose the samples from the entire dataset with replacement for individually and independently training each decision tree. Several randomized decision trees are then combined and aggregated for their outcomes or predictions by majority voting for classification problems and averaging for regression problems. Figure 6 shows the random forest algorithm for the regression problem. The method enhances model performance compared to a single decision tree model and has a low overfitting problem (Biau & Scornet, 2015).

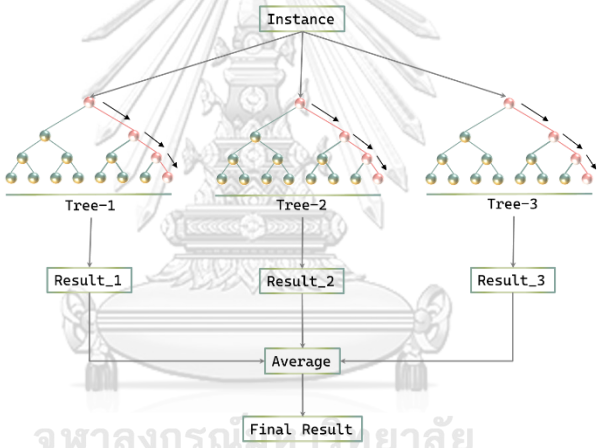


Figure 6 Random forest diagram
(Source: (Wang et al., 2020))

2.1.3.2 Extreme gradient boosting (XGBoost)

XGBoost is a boosting ensemble learning method for supervised machine learning that can be used in both classification and regression problems. The method combines many weak learners into one strong learner by implementing gradient boosting decision trees to create a series of decision tree models sequentially and trying to minimize the errors of the previous decision trees as shown in Figure 7. The model can achieve high accuracy and process in parallel to reduce runtime. Besides, XGBoost includes

regularization hyperparameters for setting to reduce overfitting and improve overall performance.

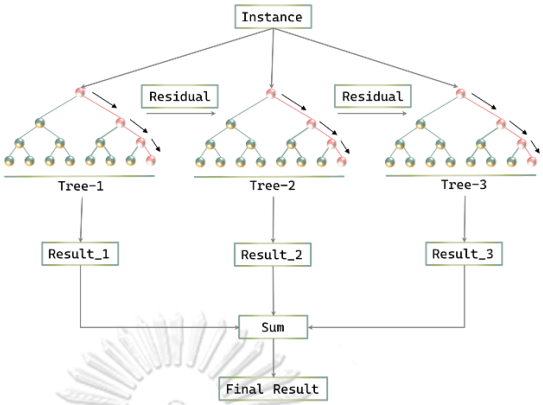


Figure 7 XGBoost diagram
(Source: (Wang et al., 2020))

2.1.3.3 Artificial neural networks (ANNs)

ANNs are computational networks simulated by the human–brain processes. ANN consists of hundreds of single units or nodes called artificial neurons, transmitting a signal to other neurons through the connections called edges containing coefficients (weights), which constitute the neural structure and are organized in layers (Agatonovic-Kustrin & Beresford, 2000). After the neuron receiving the signals, each input is multiplied by a corresponding weight, and these weighted inputs are summed up. The weighted sum is then passed through an activation function, which provides non-linearity into the network, to provide the desired output. The activated output is sent to the next layer or used as the final output of the network. Figure 8 shows the components of a single neuron.

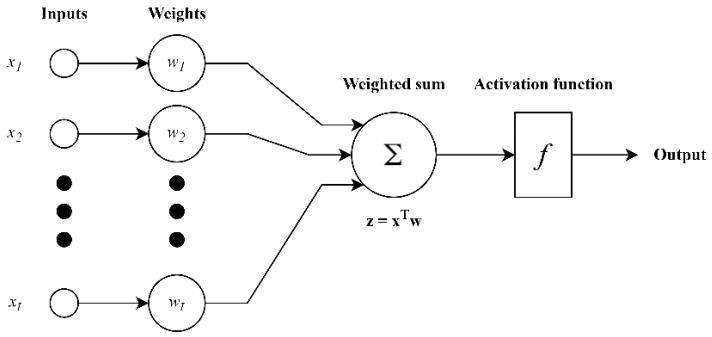


Figure 8 The component of an artificial neuron
(Source:(Boukadida et al., 2011))

There are many types of ANNs, e.g., feed forward ANN, feed backward ANN or competitive ANN. Generally, ANN architecture that has been popularly used to forecast is a Multi-Layer Perceptron (MLP) network which consists of three main types of layers: input layer, hidden layer(s), and output layer(s) as shown on Figure 9 (Lek & Park, 2008; Samang, 2020).

The input layer receives the initial input data and passes it to the next layer. It does not perform any computations. The number of neurons in the input layer depends on the number of independent variables. The hidden layer may be a single layer or multiple layers. These layers are to receive input data and process it through their neurons, performing complex computations, then send the output to the next layer or the output layer. The optimal number of hidden layers are selected by trial and error. It commonly has one hidden layer because it provides enough accurate prediction. However, for modelling complex problem, more hidden layers may be used (Khan et al., 2023). The output layer will have a single neuron for regression problem and a neuron or several neurons depended on class label for classification problem. The output layer provides the predicted output and is then compared to the actual value. The data from the input layer is passed through to the output layer by forward propagation technique and after training the model, the model uses the backpropagation technique, which is a method that optimizes the weights connected between nodes based on the difference between the obtained results

and the desired results, to improve the performance of the network (Auppakorn & Phumchusri, 2022; Ghafari et al., 2014; Samang, 2020).

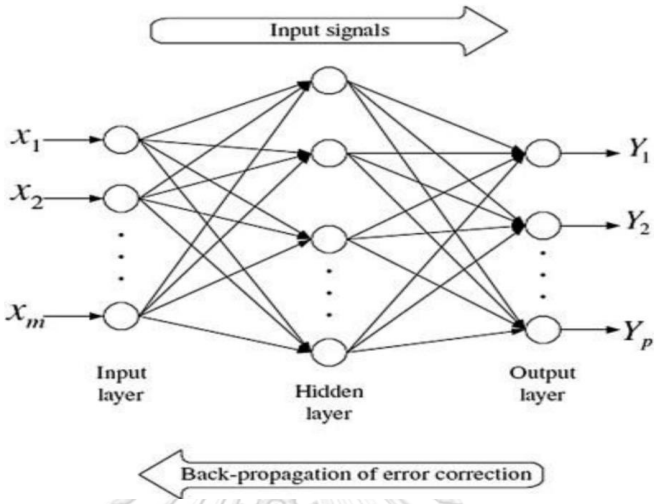


Figure 9 The structure of Multi-Layer Perceptron (MLP) network (source: (Ghafari et al., 2014))

2.1.3.4 Hybrid forecasting approaches

A hybrid model combines several forecasting techniques to improve prediction accuracy, lower the chance of selecting the inappropriate model because of the combination method and reduce the complexity of the model selection process. Based on the hybrid model structure proposed in several research, it can be divided into three types: parallel, series and parallel-series (Hajirahimi & Khashei, 2019).

For parallel hybrid structure as shown in Figure 10, the forecasts are weighted and integrated by averaging, employing linear or nonlinear function as given in Equation. (2).

$$f_{combined,\,t} = \varphi\left(w_1\hat{f}_{1,t}, w_2\hat{f}_{2,t}, \ldots, w_n\hat{f}_{n,t}\right) \;;\; t = 1, 2, \ldots, T \qquad (2)$$

where φ is the hybrid function

$w_i\hat{f}_i$ are the weighted forecasted value of each individual model

T is the number of data

n is the number of the based models or components

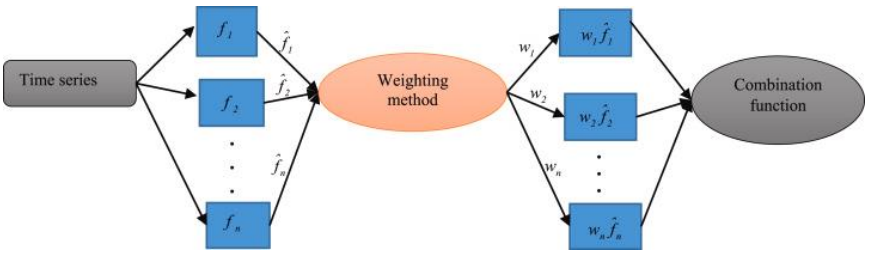


Figure 10 The parallel hybrid structure

(source: (Hajirahimi & Khashei, 2019))

Figure 11 shows the series hybrid structure, which uses the concept of sequential modeling procedure. It is usually divided into linear and non-linear part. To construct the model, firstly, the first model is fitted and the forecasted value is calculated. After that, residuals from the first model are used as an input for the second model, which then fits the second model. The final forecast value is the sum of different forecasts, as given in Equation. (3) (Hajirahimi & Khashei, 2019).

$$f_{combined,\,t} = \hat{f}_{1,t} + \hat{f}'_{2,t} + \ldots + \hat{f}'_{n,t} \quad ; t = 1, 2, \ldots, T \qquad (3)$$

where $\hat{f}_{i,t}$ are the weighted forecasted value of each individual model

T is the number of data

n is the number of the based models or components

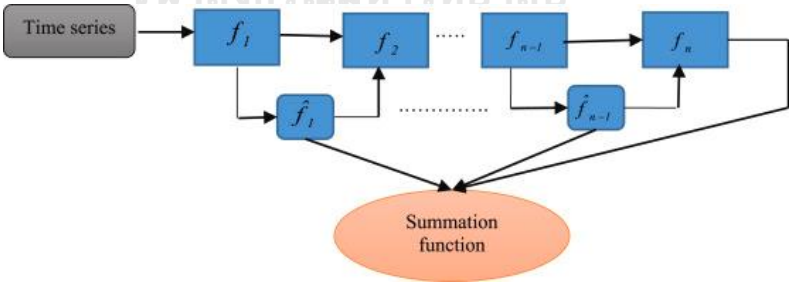


Figure 11 The series hybrid structure

(source: (Hajirahimi & Khashei, 2019))

For the parallel-series hybrid structure, the model is constructed based on the combination of parallel and series hybrid concepts to extract the advantages of both structures. However, there will be a disadvantage as it

increases the complexity of the model, which takes longer to compute (Hajirahimi & Khashei, 2019).

2.1.4 Data standardization

Data standardization is a step for machine learning algorithms to rescale or standardize the features in a dataset to have a mean of 0 and a standard deviation of 1. It uses the formular in Equation (4) to rescale data.

$$Z = \frac{X - \mu}{\sigma} \quad (4)$$

where μ is the mean and σ is the standard deviation

2.1.5 Hyperparameter Tuning

A hyperparameter is a parameter whose value is set manually to control the learning process before applying a training algorithm to a dataset, for instance, the learning rate in a neural network, the number of groups in k-means clustering, or the number of trees in a random forest. Grid search is the simplest method for hyperparameter tuning techniques, which test all combinations of possible defined hyperparameter values to find the optimal parameters producing the best results. However, it is possible to miss a better hyperparameter.

2.1.6 K-fold cross validation

Cross-validation is one of the most popular data resampling techniques for evaluating the performance of machine learning models and optimizing model parameters to prevent overfitting (Arlot & Celisse, 2010; Berrar, 2019). It usually splits a dataset into two parts, one for training models and another for validating models' performance, to help in comparing models and selecting a suitable model for a particular problem.

K-fold cross validation is one of the cross-validation techniques that divides data into k disjoint, equal-size subsets or folds. The model is trained by using k-1 folds as a training set and measured for performance by using the rest of the folds as a validation set. This process is repeated with different validation folds until each fold is

used as a validation fold. The cross-validation performance is measured by averaging k performance measurements on k validation sets, as shown in Figure 12. This method provides reliable results because training and testing are performed on several different parts of the dataset. It is useful for small datasets or when the model has many hyperparameters to be tuned. For a larger k, each model is trained on a larger training set (closer to the entire dataset) and tested on a smaller test set, which may result in a lower prediction error as the models see more of the available data. However, an increase in k leads to time-consuming Commonly, k-fold cross validation is done using k = 5 or k = 10. It is recommended using 10-fold cross validation (k = 10) (Arlot & Celisse, 2010).

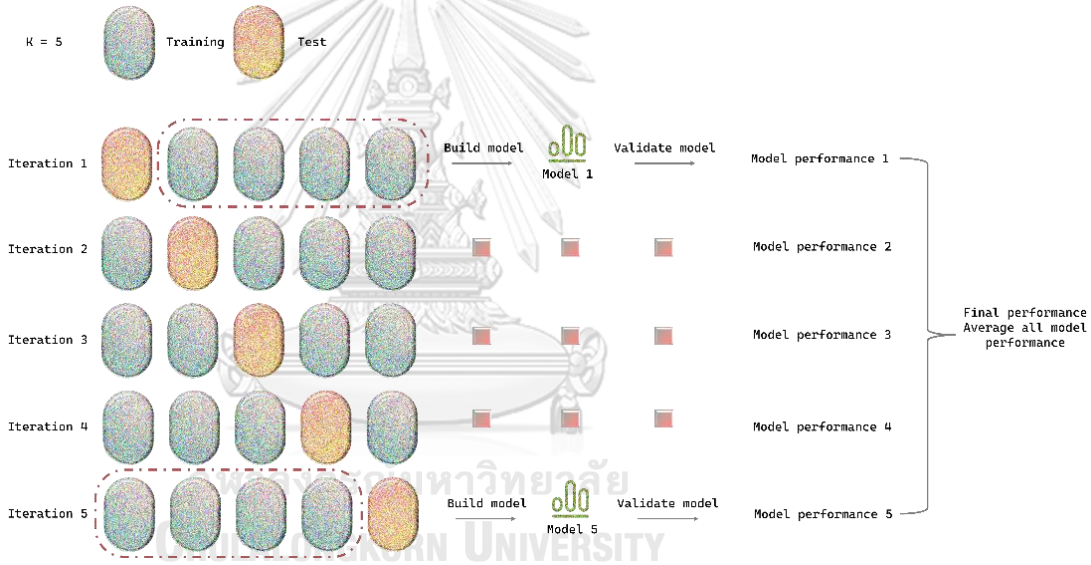


Figure 12 5-fold cross validation
(Source: (García et al., 2019))

2.1.7 Performance Metrics

2.1.7.1 Coefficient of determination or R-Squared (R^2)

The R^2 is an indicator representing the proportion of variance in the dependent variable that is explained by the independent variables in the regression model. It explains how well a model predicts or explains the outcomes. The R^2 can have any value between 0 and 1, which is commonly

expressed in percentages and can be calculated by Equation (5) (Tian et al., 2021). The higher R^2 indicates a better fit for the model.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (5)$$

where y_i is the actual values

$\hat{y}_i$ is the predict values

$\bar{y}$ is the mean of the actual values

2.1.7.2 Mean Absolute Percentage Error (MAPE)

MAPE is an error measurement metric representing the average of the absolute percentage errors of model prediction in relation to actual values as shown in Equation (6) (Hamzaçebi, 2008).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i} \text{ x } 100 \quad (6)$$

where y_i is the actual values

$\hat{y}_i$ is the predict values

n is the number of observations

Table 1 shows the interpretation of MAPE values, the lower MAPE meaning high accuracy prediction (Montaño et al., 2013).

Table 1 The MAPE interpretation

MAPE (%)	**Interpretation**
< 10	Highly accurate forecasting
10 – 20	Good forecasting
20 – 50	Reasonable forecasting
> 50	Inaccurate forecasting

2.1.8 Shapley Additive Explanations (SHAP value)

SHAP is a method that can be used to make machine learning model more explainable including feature importance, influence of the features or explaining the prediction. Figure 13 shows an example of SHAP feature importance.

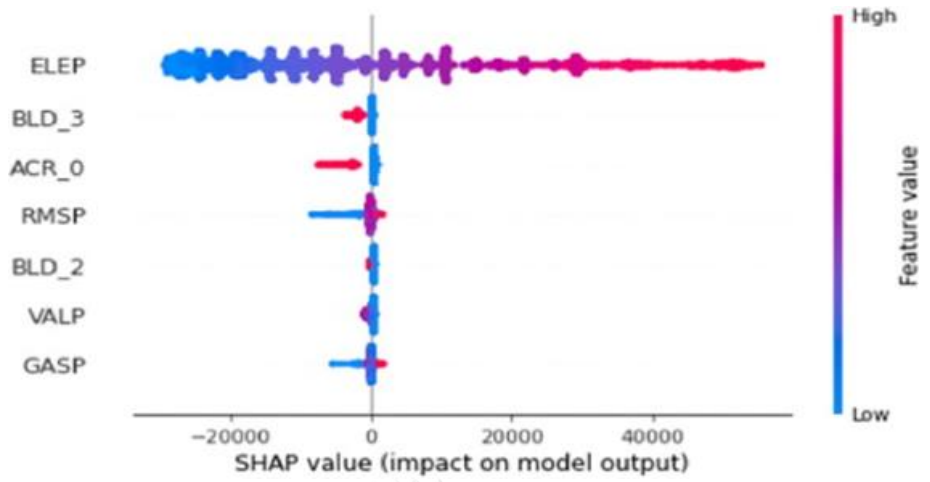


Figure 13 SHAP feature importance
(Source: (Amiri et al., 2023))

2.2 Related Research

Several studies have looked at and compared different forecasting methods, ranging from traditional models to novel approaches. This part describes the influencing factors that should be considered in forecasting, the demand forecasting models used in forecasting, and the indicators that measure the performance of the forecasting model.

2.2.1 Factors influencing sales

In previous works, many factors that may affect sales were studied for sales prediction in retail industries, as shown in Table 2. Those factors can be grouped into five types, including product and store details, time period and seasonal factors, past sales, price and promotion, and other external factors.

For the product and store details, the product data related to number, type, group, or sub-group or the store data related to store number, location, and types are usually used in sales prediction. Loureiro et al. (2018) used product groups, sub-groups, and store types. The number, type and location of store were studied by Huber and Stuckenschmidt (2020), Saha et al. (2022), Wang (2020), and Mitra et al. (2022).

Moreover, store size was included for sales forecasting in research by Boyapati and Mummidi (2020).

Time period and seasonal factors, which may depend on the data timeframe. The factors may be used as day of the week, weekday or weekend, holiday or other special days, week of the month or month of the year. For example, daily sales forecasting studied by Prabhakar et al. (2018) used factors related to date and holiday. Similarly, Huber and Stuckenschmidt (2020) focused on period factors, including day of year, month, weekday, and special days. Mitra et al. (2022) studied the holiday week for the weekly forecasting.

Some research explored sales forecasting using previous sales information as one of the factors. These variables may also be described as shift sales or lagged sales studied by Abolghasemi et al. (2020), Auppakorn and Phumchusri (2022) and Yoon et al. (2023), rolling mean lagged sales investigated by Saha et al. (2022), or moving average sales explored by Auppakorn and Phumchusri (2022) and Yoon et al. (2023).

Most studies have considered price or promotion as factors that impact retail sales. Product price was often used in the studies. Punia et al. (2020) studied relative price, representing competition within a category. Additionally, promotion was mostly considered by using a price discount, a percentage of price reduction or discount, or a binary variable of promotion.

For the external factors, Mitra et al. (2022) studied the average temperature, fuel price, consumer price index (CPI) and unemployment rate. In Punia and Shankar (2022) study also included weather and economic activity index. Considering Thailand's retail sector, Auppakorn and Phumchusri (2022) included the Thai government’s welfare projects and COVID-19 in their study.

Table 2 Features studied in retail industries

Author Name (Year)	Data	Factors
Arunraj and Ahrens (2015)	Daily sales of a perishable food in Germany retail store	• Seasonal effect as dummy variables (e.g., day, month, and special day effect) • Promotion effect (percentage of price reduction, discount) • Weather effect used as the extreme weather conditions (e.g., temperature, precipitation, relative humidity)
Loureiro et al. (2018)	Sales of new individual fashion products in fashion retail	• Product price • Product size and color • Product group and sub-group • Store type
Prabhakar et al. (2018)	Daily unit sales of products for large grocery chain	• Date • Store type and location • Items • Promotion (binary variable) • Holiday
Chiewpanich and Mokkhamakkul (2019)	Monthly sales of 4 bedding product categories	• Number of products doing promotion • Price discount • Feature displays and capacity • Period • Promotion duration • Inventory level • Weather condition • Special date and holiday
Abolghasemi et al. (2020)	Weekly demand of FMCG companies	• Past sales • Promotion effect (promotion type, display type and advertisement type)
Boyapati and Mummidi (2020)	Sales data from different items from different outlets	• Item details: weight, visibility, MRP • Outlet: establish year, size, location
Huber and Stuckenschmidt (2020)	Daily retail demand	• Time: day of year, month, weekday, special calendar days • Store class, store location • Product category • Past sales: lagged sales, rolling median of sales • Binary promotion information
Punia et al. (2020)	Daily sales for online stores and weekly sales for offline stores of	• Products details • Based price and relative price as competition within category • Discount, Display/ feature • Holiday (binary) • Visits • Store Area • Temporary price reduction (TPR) • No. of households (HHS)

Table 2 Features studied in retail industries (Cont.)

Author Name (Year)	Data	Factors
Wang (2020)	Daily sales of retail products	• Time • Item: level, department, category • Store detail: store ID, state ID • Price and promotion • Special events
Auppakorn and Phumchusri (2022)	Daily sales in Thai retail business	• Selling price, ratio of promotion price and regular price • External factor (COVID-19, the Thai government's scheme) • Time (Day, month, holiday, beginning or ending of month) • About past Sales (Lag Sales, moving average)
Mitra et al. (2022)	Weekly demand of a US-based retail company	• Store and geographic-specific information: store number, size • Time: date mentioning the week, holiday week • External data: region's average temperature, fuel price in the region, CPI (consumer price index), unemployment rate
Punia and Shankar (2022)	Weekly sales for 55 food items sold through 77 retail stores	• Price and promotion: display, temporary price reduction and feature, percentage of discount • The number of customers and the number of purchasing households that visited the store in the given week • External factor: weather, economic activity index
Saha et al. (2022)	Daily sales of American retail company	• Time: day of the week, weekday number, the month of the date, and year of the date • Event name and event type as binary variable • Item: ID, price • Store: ID, location, state • Past sales: rolling mean lagged sales
Yoon et al. (2023)	Weekly unit demand data from US retail companies	• Time period • Moving Average sales, Lag sales • Stores

2.2.2 Demand forecasting in Retail industries

Demand forecasting models, from traditional demand forecasting models to hybrid approaches, were studied for demand forecasting, as presented in Table 3. Previous research revealed that there is no obvious conclusion regarding the best forecasting method for all datasets and under all conditions. For the traditional model, time series and causal models were used in many studies as their base models. The time series model only considered past data, so it may not be an appropriate model to forecast retail sales due to other factors that should be included, such as promotion or store effects. For multivariate variables considered, a causal model such as linear regression was widely used as the base model. However, linear regression may give poor prediction performance if the assumptions are violated, such as when there are outliers in the data. Contrary to linear regression, machine learning provides highly accurate prediction, which is popularly used and can deal with large, non-linear, and complex datasets. Although it may not conclude the best models for retail sales forecasting, most research shows that machine learning methods and hybrid approaches outperform traditional methods.

This study considered on random forest regression, XGBoost, artificial neural networks (ANN) algorithms and hybrid model. Random forest regression and XGBoost have been widely studied for retail forecasting. According to the research, those two models have high prediction accuracy. For random forest regression, Čeh et al. (2018) studied price prediction for apartments using multiple regression and random forest regression. The result showed that random forest regression was better. Loureiro et al. (2018) studied sales forecasting in fashion retail using linear regression, decision trees, random forest regression, SVR, artificial neural networks (ANN), and deep neural network (DNN). Although DNN outperformed in terms of RMSE and MSE, the random forest regression was the best model in terms of R^2, MAPE, and MAE. The result also found that random forest regression and DNN had slightly different prediction measurements, so random forest regression can be concluded to be the best-performing technique. Sales forecasting using linear regression, gradient boosting regression (GBR), SVR and random forest regression were studied by Boyapati and Mummidi (2020), it was recommended that random forest regression was the most appropriate algorithm compared to the other models.

Moreover, Singh et al. (2019) investigated demand forecasting for new items using XGBoost, NNs, and long-short-term memory (LSTM) and concluded that XGBoost was a suitable model. Daily sales forecasting using XGBoost and LSTM researched by Swami et al. (2020) and found that XGBoost fared better than LSTM for this dataset. Similarly, Wang (2020) studied daily sales prediction performance using linear regression, SVR, Light Gradient Boosting Machine (LGBM), XGBoost, and a hybrid model. The result found that XGBoost, LGBM and the hybrid models were all high accuracy for prediction and the hybrid model was barely better than others. Some research studied both random forest regression and XGBoost to predict sales, however, the winning method were different depending on the dataset and features. The XGBoost may perform better than the random forest regression in some cases, such as Jain et al. (2015) research studied daily sales forecasting for pharmacy retail stores using linear regression, random forest regression, and XGBoost and found that XGBoost outperformed. While, random forest regression may have higher performance than XGBoost as reported by Almeida et al. (2022). They studied daily sales data prediction using linear regression, random forest regression and XGBoost and found that the best algorithm was different for each store.

ANN has also been used for forecasting or prediction in the retail industry because of its high accuracy. Güven & Şimşir (2020) studied sales forecasting in the retail garment industry by comparing ANN and support vector machines (SVM). The result showed that ANN forecasting had a lower RMSE than SVM for seven out of ten colorless datasets. It can be concluded that ANN predicts more accurately than SVM. Similarly, Auppakorn and Phumchusri (2022) studied and compared the performance of daily sales forecasting models in retail business and found that for data with or without both trend and seasonal, ANN with data transformation by natural logarithm was outperform.

Additionally, hybrid models, which are a mixed methods of multiple machine learning algorithms, were studied and concluded to outperform other methods for prediction (Arunraj & Ahrens, 2015; Mitra et al., 2022; Punia & Shankar, 2022). Unsupervised machine learning algorithms have been used to improve forecasting performance and provide insight that help retail companies to understand the sales structure (Kadam & Lingras, 2023; Yang & Nguyen, 2022). K-means clustering

technique is the most popular model used to group the data with similar pattern due to its easy to implement and low memory consumption (Chung et al., 2023). Wijaya et al. (2020) using K-means algorithm to group the data and using the other machine learning algorithms for prediction. The finding found that the performance of forecasts after clustering was improved and better fitted than forecasts without clustering. Moon et al. (2022) studied and compared the performance of prediction models and found that using K-means clustering to group the data based on their similarity and build independent machine-learning models specialized for each group had the highest accuracy. Similar to the research studied by Yoon et al. (2023) using K-means algorithm to cluster similar data and then applied forecasting models. The result found that each cluster had different variables that effect sales of the cluster and improved the prediction accuracy.

2.2.3 Performance measurement

According to the research, many performance metrics were used, as shown in Table 3. For demand forecasting metrics in regression approach, R^2 and error measurement including MAE, MSE, RMSE or MAPE were usually used to compare the model's performance.

Prabhakar et al. (2018) studied unit sales prediction and suggested using RMSE and R^2 as measures of model performance for predicting a continuous variable. However, to compare model performance using different datasets, the MAPE which reported as a percentage is more suitable (Ensafi et al., 2022). In this work, R^2 and MAPE were used as performance metrics.

2.3 Research gap

According to all the mentioned studies, the research gaps are as follows:

- The factors of the COVID-19 effect and the Thai government’s subsidies and welfare projects are rarely used in retail demand prediction because these events just recently occurred and greatly

affected product sales. Moreover, the data used in this study is during the COVID-19 pandemic.

- The previous studies focused on finding the accurate model for product, product category or product group prediction by using all the data in the same group to sales (1 model for all products in each category or each group). Regarding investigation, there is little research conducted the concept of clustering the similar product into the same group and then predicting for each product using some of other products' variables as independent variables.

This study focuses on predicting monthly unit sales considering selling price, promotion characteristics, monthly period, number of active stores, number of COVID-19 new cases in Thailand, and Thai government subsidies and welfare programs. Five prediction techniques are constructed, including linear regression with the stepwise method, random forest, XGBoost, ANN and the hybrid model. Moreover, the model considers factors of other products in the same group using the clustering method before prediction is performed and compared to the model without considering them, which may provide useful information such as which products have similarities or how the exogenous variables of other products affect sales. The clustering method includes three types of criteria: by category, by subcategory and by K-means method using sales quantity, selling price and promotion. After constructing the models and predicting the results, the prediction performance is measured by MAPE. For model comparison and selection, the overall performance is measured by weighted MAPE (WMAPE).

Table 3 Summary of research related to demand forecasting

Author Name (Year)	Data	Industry	Demand model/ forecasting method	Measurement Error	Result
Aburto and Weber (2007)	Daily sales of six best-selling SKUs in the store	Retail industry (supermarket chain)	• Naïve • Seasonal Naïve • ARIMA • Neural network (NN) • ARIMA-Neural network	MAPE, NMSE	• Hybrid model was outperformed. • NN require a large number of training examples to obtain reliable results and easily overfitting
Ali et al. (2009)	Weekly SKU-store level sales and promotion time series	Retail industry (grocery retailer)	• Exponential smoothing • Stepwise linear regression • Decision tree • Support Vector Regression techniques with different kernels (SVR)	MAE, MAPE	• One model across all subcategories and stores performs as good or better than others and lowest effort, cost, and time • Traditional model gave best result when non-promotion • Decision tree outperformed when promotion • Considering the large number of features, tree regression employed to forecast SKU sales
Jain et al. (2015)	Daily sales of 1115 stores in Germany	Retail industry (Pharmacy)	• Linear regression • Random forest regression • Extreme Gradient Boosting (XGBoost)	RMPSE	XGBoost was outperform others
Čeh et al. (2018)	Apartment transactions from 2008–2013 (7407 records)	Prices of the Apartments	• Multiple linear regression • Random Forest	R^2 values MAPE coefficient of dispersion (COD)	Random forest was significantly to be better technique to predict price

Table 3 Summary of research related to demand forecasting (Cont.)

Author Name (Year)	Data	Industry	Demand model/ forecasting method	Measurement Error	Result
Loureiro et al. (2018)	Data of 684 types of women bags from Spring-Summer 2015-2016	Retail industry (fashion products)	• Linear Regression • Decision Trees • Random Forest • Support Vector Regression • Artificial Neural Networks • Deep NN	R^2 RMSE MAPE MAE MSE	• DNN was gave better results of RMSE and MSE. • RF was giving better results of R2, MAPE, and MAE
Prabhakar et al. (2018)	Daily unit sales from 2013 and 2017 Data split ratio: 75:25	Retail industry (grocery)	• K-fold cross validation (k=6) • Linear Regression • Gradient Boosting Method • ANN • Support Vector Machine	R^2 RMSE	• SVM high runtime • Gradient Boosting Method was outperformed
Chiewpanich and Mokkhamakkul (2019)	Daily sales of 4 product categories which are Pillow and Bolster, Bed set, Container and Storage and Hanger from March 2016 to February 2018	Retail industry	Multiple Linear Regression	MAPE	Different important features for each product category

Table 3 Summary of research related to demand forecasting (Cont.)

Author Name (Year)	Data	Industry	Demand model/ forecasting method	Measurement Error	Result
Singh et al. (2019)	Data for 2 years	Retail industry (Fashion)	• XGBoost • GBRT • LSTM • MLP	wMAPE	XGBoost with criterion = MSE was outperform others model
Boyapati and Mummidi (2020)	Data from different items from different outlets (8523 instances)	Retail industry	• K-fold cross-validation • Feature selection by data correlation • Simple Linear Regression • Gradient Boosting Regression • Support Vector Regression • Random Forest Regression	Max error, MAE, Accuracy score	Random Forest Regression is the most appropriate algorithm compared to the others.
Huber and Stuckenschmidt (2020)	Daily sales of 8 product categories in 141 stores for 2015 to 2017 (1128 time series)	Retail industry (Bakery chain)	• S-Naïve, S- Median • Exponential smoothing • LASSO regression • LGBM • ANN • LSTM	MASE, MAE, SMAPE, RMSE	ML outperformed traditional models LSTM was giving the best result
Punia et al. (2020)	Data from online platform (daily) and eleven offline stores (weekly)	Retail industry	• neural networks • multiple regression • ARIMAX • LSTM networks • RF • LSTM - RF	Bias/ ME Accuracy/ MAE Variance/ MSE	The proposed model outperformed

Table 3 Summary of research related to demand forecasting (Cont.)

Author Name (Year)	Data	Industry	Demand model/ forecasting method	Measurement Error	Result
Swami et al. (2020)	Daily sales data from Russian software firms	Retail industry	• XGBoost • LSTM	RMSE	XGBoost fared better than LSTM over this dataset
Wang (2020)	Daily data for 4 years from Wal-Mart retail products in three U.S. states (California, Texas, and Wisconsin)	Retail industry	• Linear Regression • SVR • LGBM • XGBoost • LGBM and XGBoost	RMSE	Hybrid method with LGBM and XGBoost was slightly lower RMSE that LGBM and XGBoost
Güven and Şimşir (2020)	Weekly sales of selected products from the 49th week of 2014 and the 52nd week of 2018	Retail industry (apparel)	• ANN • SVR	RMSE	• The ANN outperformed SVM on seven datasets out of ten for the datasets
Auppakorn and Phumchusri (2022)	Daily Sales from 1 January 2019 – 31 August 2021	Retail industry (supermarket chain)	• Day Forward-Chaining Cross-validation • Time Series: TBATS • MLR • ML: XGBoost, ANN • Hybrid: TBATS-ANN, TBATS-XGBoost	MAPE, SHAP	• Data transformation with natural logarithm of dependent variables improved performance • ANN with data transformation for seasonal and trend data and not seasonal and not trend data • Hybrid for seasonal but not trend data • From SHAP Value, the external factor (COVID-19, the Thai government's scheme) and past sales were high impact.

Table 3 Summary of research related to demand forecasting (Cont.)

Author Name (Year)	Data	Industry	Demand model/ forecasting method	Measurement Error	Result
Ahmed et al. (2022)	Daily data of eight different store for 5 years from 2015 to 2019 and predict 2020 pre-pandemic values	Retail industry (supermarket chain)	• Walk-forward cross-validation • Linear regression • Random forests • XGBoost	R^2 score	• The best algorithm varies per store, but for most stores at least one of the methods proves effective.
Mitra et al. (2022)	Weekly sales for all the 45 stores and 99 departments over 3 years	Retail industry	• Random forest • XGBoost • AdaBoost • ANN • RF-XGBoost-LR	MSE, R^2 score, MAE	RF-XGBoost-LR was outperform
Punia and Shankar (2022)	Weekly sales for 55 food items sold through 77 stores and is available for the duration from January 2009 to January 2012. (4235 demand time series)	Retail industry (packaged food products)	• OLS Regression • ARIMA, ARIMAX • Back propagation neural network • Random forest • LSTM • ARIMA+NN, ARIMA+RF • LSTM-RF and weight by GA	ME, MAE, MSE, RME, RMAE, RMSE	The proposed method outperformed the benchmarking methods.

Table 3 Summary of research related to demand forecasting (Cont.)

Author Name (Year)	Data	Industry	Demand model/ forecasting method	Measurement Error	Result
Tekin and Sarı (2022)	Daily demand dataset from 10 different stores and 50 items	Retail industry	• 10-fold cross validation • Linear regression • Huber regression • Ridge regression • Random forest • LightGBM • Catboost • XGBoost	R-squared, MAE, MSE, RMSE, RMSLE, MAPE, runtime	LightGBM and Catboost were outperformed other models

Chapter 3 Methodology

This study aims to predict monthly sales of beauty products using prediction models and study factors influencing sales. The overall procedures include data collection, data preparation, and data exploration to prepare and understand the data, then construct models to predict and compare the performance, as shown in Figure 14.

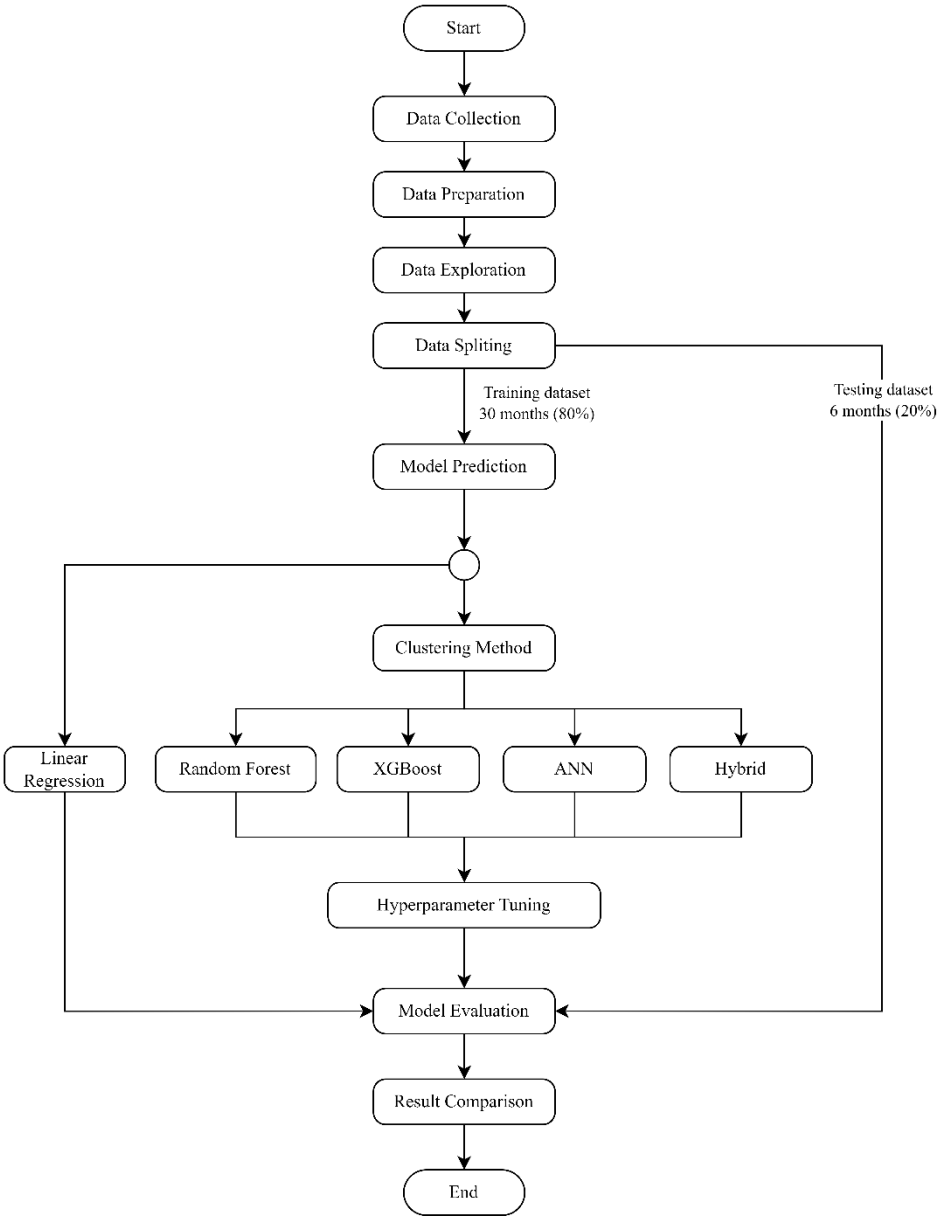


Figure 14 The overall processes

3.1 Data collection

The sales dataset used in this work is monthly sales of beauty products sold by the case study retail company for the period January 2020 to December 2022 (36 months). According to the total revenue in seven categories of beauty products, as shown in Figure 3, the facial moisturizer category is focused due to its high revenue, which is about 30% of the total revenue. This work considers the top ten best-selling beauty products in the category. The price and promotion, store and welfare history are also provided by the company. The COVID-19 new cases in Thailand are obtained from the ministry of public health's department of disease control. The consolidated data is shown in Table 4.

Table 4 Data description

Column	**Description**
Categories	Product category
Sub-category	Product sub-category
GP	Product gross profit group including 3 levels (high, medium, low) defined by the company.
Product name	Product name
Promotion mechanic	Types of promotions
Normal price	Regular price or price when not promotion
Sold prices	Selling price
Period	Month
Stores	Number of total active stores in the period
Subsidies and welfare programs	The Thai's government subsidies and welfare programs including the 50:50 co-payment scheme, the We Win scheme or others.
COVID-19 new cases	Number of total COVID-19 new cases in the period
Sales quantity	Amount of the product sold in the period

3.2 Data Preparation

To prepare the dataset, it is initially cleaned by locating and removing missing and duplicate rows. Then the data types were checked to ensure that they were correct.

The factors or independent variables considered in the study include the following:

- The price factor is defined as the selling price of promotions and non-promotions.
- The promotion characteristics factors are discount percentage, promotion period or 1-, 2-lag promotion period. This study does not take into account promotion types as a factor since most products are only promoted through one type of promotion.
- The period factor is defined as month.
- The store factor is defined as the number of total active stores in a month.
- Subsidies and welfare programs are defined as the Thai government's subsidies and welfare programs, including the 50:50 co-payment scheme, the We Win scheme or others.
- COVID-19 new cases are defined as the number of monthly total COVID-19 new cases in Thailand.

To define the target variable or dependent variable, the sales quantity is used in the work.

3.3 Data exploration

In this step, the dataset is explored and summarized to discover some insight, pattern, relationships, or useful information using summary statistics including the mean, standard deviation, minimum, or maximum value, or using data visualization such as a histogram or line chart. This approach helps to understand the dataset structure and distribution as well as detect some abnormal data or outliers.

3.4 Clustering method

According to the previous research, clustering and constructing sales prediction models for all the products in the same cluster was found to improve performance and provide more insight compared to using one prediction model for a single product or all the products (Yoon et al., 2023). Moreover, product

cannibalization may occur when similar products are doing a promotion or price discount. Thus, this work also includes price and promotion characteristic factors of other similar products as one of the independent variables to predict sales quantity. The clustering method in this work is conducted and clustered by three types: category, sub-category of beauty products assigned by the company and K-means method considering price, sales quantity and promotion to group similar products and then construct the prediction models, then taking exogenous factors such as the selling price, discount percentage, promotion period and lagged promotion period of the other products in the same group into account to predict sales and compare performance, as shown in Figure 15.

Steps of model construction as follows:

(i) Import the dataset table

(ii) Divide the entire dataset into two datasets including training dataset for 30 months (January 2020 – June 2022) and test dataset for 6 months (July 2022 – December 2022)

(iii) Perform feature scaling of the numeric factors on training dataset then fit on test dataset

(iv) Cluster the similar products into the same group

   a. Clustering by category: considering all products in the facial moisturizers category
   b. Clustering by subcategory: considering the products assigned into the same subcategory by the company
   c. Clustering by K-means method considering price, sales quantity and promotion

(v) Include factors of the products in the same group which are selling price, discount percentage and promotion period, as one of the independent variables

(vi) Use training dataset to construct the machine learning model with grid search and 10-fold cross validation method to tune hyperparameters.

(vii) Apply the optimal setting of hyperparameters that provides the lowest average error measurement or highest average accuracy during the cross-validation process to the test dataset

(viii) Validate and record the model performance by measurement metric on the test dataset

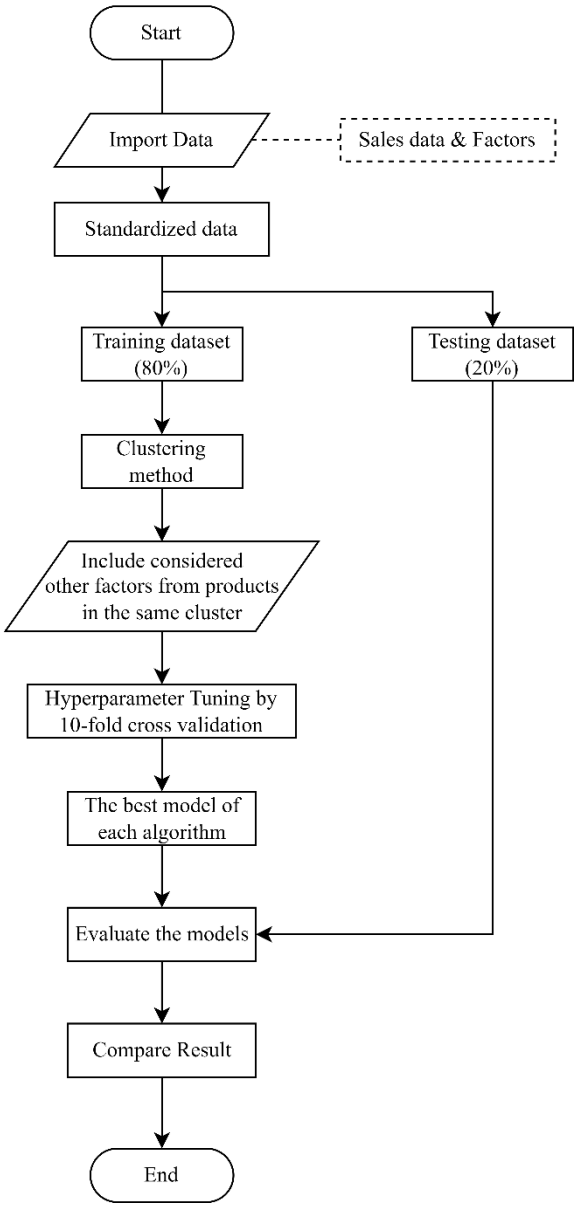


Figure 15 The methodology of machine learning techniques with clustering method

3.5 Prediction models

The prediction models that are considered in this work include as follows:

3.5.1 Linear regression

The linear regression is used to predict sales as the traditional method, identified important factors and compared the performance of the model using the Minitab program. The factors, including the monthly period, promotion period and the subsidies and welfare programs, are transformed into a binary dummy variable with 1 and 0 options.

The significant factors are selected by stepwise method with the entered alpha is 0.05 and removed alpha is 0.05. Under assumptions of linearity between variables and independent and identically normal distributed residuals, the model is then fitted, calculated predictions and compared model performance.

Steps of model construction as follows:

(i) Import the dataset table

(ii) Divide the entire dataset into two datasets including training dataset for 30 months (January 2020 – June 2022) and test dataset for 6 months (July 2022 – December 2022)

(iii) Assign factors and target variable of the training dataset and the test dataset

(iv) Use training dataset to fit stepwise regression with alpha to enter = 0.05, alpha to remove = 0.05 by Minitab

(v) Predict the training dataset and testing dataset by the regression equation

(vi) Validate and record the model performance by measurement metric on the test dataset

3.5.2 Machine learning algorithm

For machine learning methods, this work considers random forest, XGBoost and ANN to predict sales of the selected beauty products and constructs the machine learning models using Python. Factors using for training the models are either all the

factors (see in 3.2) or the significant factors from stepwise regression method and with or without considering factors from other products by clustering methods (see in 3.4), as shown in Table 5.

Table 5 The total cases of the features studied in this work

	Factor	
Cases	**Each product**	**Other products**
1	All factors	Not consider other products's factors
2	All factors	Consider factors of other products in the same group by category
3	All factors	Consider factors of other products in the same group by subcategory
4	All factors	Consider factors of other products in the same group by K-means
5	Significant factors	Not consider other products's factors
6	Significant factors	Consider factors of other products in the same group by category
7	Significant factors	Consider factors of other products in the same group by subcategory
8	Significant factors	Consider factors of other products in the same group by K-means

The numeric factors are standardized and scaled by using 'StandardScaler' in the Scikit-Learn library. Moreover, to find the best hyperparameters and validate the performance of the machine learning model, grid search and the 10-fold cross validation technique are used. Then, the model with the best hyperparameters, which results in the lowest mean squared errors, is applied to the test set and evaluated for predictive accuracy. The methodology for machine learning algorithms shows in Figure 16.

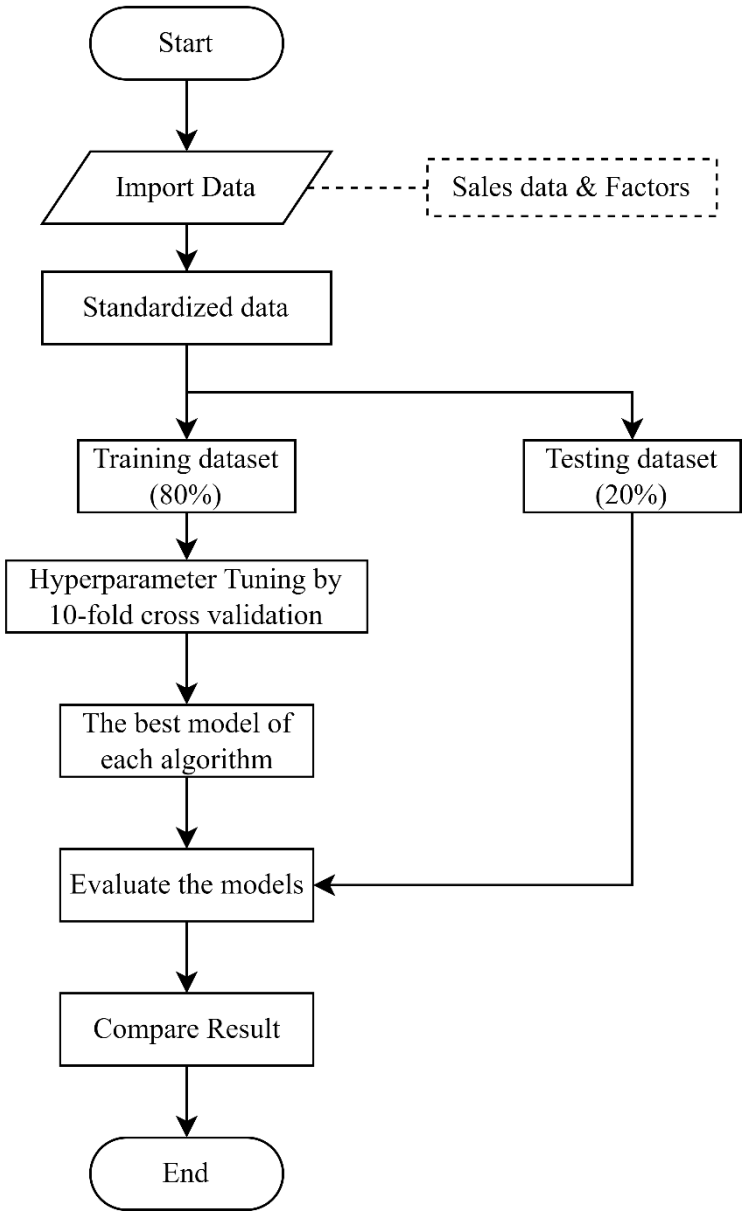


Figure 16 The methodology of the machine learning methods

3.5.2.1 Random forest

To construct random forest model, the 'RandomForestRegressor' in scikit-learn library is used. According to the previous research, the considered hyperparameters of random forest regression and their values or ranges is shown in Table 6.

Table 6 Hyperparameters search grid of random forest model

Hyperparameter	Value or Range
n_estimators	[10, 15, 20]
max_depth	[3, 5, 7, 9]
min_samples_split	[2, 5, 10]
min_samples_leaf	[2, 5, 10]
max_leaf_nodes	[3, 5, 10]

- Hyperparameters of random forest model studied in this work include as followed:
  - n_estimators is the number of trees. Normally, as the more trees, the better performance.
  - max_depth is the maximum depth or split of each tree can take which the deeper tree increase performance over training data but it also may overfit.
  - min_samples_split is the minimum number of samples placed in a node before the node is split. By increasing this hyperparameter, it can prevent the model overfitting.
  - min_samples_leaf is the minimum number of samples allowed in a leaf node.
  - max_leaf_nodes is the number that controls the growth of each tree. As each tree splitting into nodes, this hyperparameter is used to specify how many divisions of nodes should be done.

- Steps of model construction as follows:
  (i) Import required libraries and load the dataset table

  (ii) Divide the entire dataset into two datasets including training dataset for 30 months (January 2020 – June 2022) and test dataset for 6 months (July 2022 – December 2022)

  (iii) Perform feature scaling of the numeric factors on training dataset then fit on test dataset

(iv) Use training dataset to construct the machine learning model with grid search and 10-fold cross validation method to tune hyperparameters.

(v) Apply the optimal setting of hyperparameters that provides the lowest average error measurement or highest average accuracy during the cross-validation process to the test dataset

(vi) Validate and record the model performance by measurement metric on the test dataset

3.5.2.2 XGBoost

For XGBoost model construction, 'XGBRegressor' in xgboost library is used. From the previous research, the considered hyperparameters of the model and their values or ranges is shown in Table 7.

Table 7 Hyperparameters search grid of XGBoost model

Hyperparameter	**Value or Range**
n_estimators	[10, 15, 20]
max_depth	[3, 5]
min_child_weight	[1, 2, 3, 5]
eta	[0.01, 0.03, 0.1, 0.3]
subsample	[0.5, 0.7, 1.0]
gamma	[0, 0.5, 1]
reg_lambda	[1, 2, 3, 5]

- Hyperparameters of XGBoost model studied in this work include as followed:
  - n_estimators is the number of trees in the model. By increase this value, the performance of the model generally improves but it can lead to overfitting.

- max_depth is the maximum depth or split of each tree can take which the deeper tree increase performance over training data but it also may overfit.
- min_child_weight is the minimum amount of weight required in a child.
- eta is the learning rate which used to weight each model, often set to small values such as 0.3, 0.1, 0.01, or smaller (Auppakorn & Phumchusri, 2022).
- subsample is the fraction of samples used in each tree. A small value leads to less complex models, which can help prevent overfitting. This value is commonly set between 0.5 and 1 (Mustika et al., 2019).
- gamma is the minimum loss reduction required to make a split. Higher values increase the regularization, reduce overfitting.
- reg_lambda is L2 regularization term on weights which can be used to reduce overfitting.

- Steps of model construction as follows:
  (i) Import required libraries and load the dataset table
  (ii) Divide the entire dataset into two datasets including training dataset for 30 months (January 2020 – June 2022) and test dataset for 6 months (July 2022 – December 2022)
  (iii) Perform feature scaling of the numeric factors on training dataset then fit on test dataset
  (iv) Use training dataset to construct the machine learning model with grid search and 10-fold cross validation method to tune hyperparameters.
  (v) Apply the optimal setting of hyperparameters that provides the lowest average error measurement or highest average accuracy during the cross-validation process to the test dataset
  (vi) Validate and record the model performance by measurement metric on the test dataset

3.5.2.3 Artificial Neural Network (ANN)

To construct ANN model, python library including keras.model, keras.layer, keras.wrappers.scikit_learn and 'KerasRegressor' command are used. As reviewed papers, the hyperparameters and their values or ranges that considered are shown in Table 8. In this work, the seed value, a number used to randomly initialize the weights of the network, is fixed by using tensorflow.random.set_seed command so that every model constructions have the same initial random weights and obtain the same result (Auppakorn & Phumchusri, 2022; Samang, 2020).

Table 8 Hyperparameters search grid of ANN model

Hyperparameter	**Value or Range**
batch_size	[4, 8, 16]
epoch	100
dropout	[0, 0.2, 0.5]
learning_rate	[0.01, 0.03, 0.1]
num_layers	[1, 2]
num_units	[20, 50, 100]
activation	['relu']

- Hyperparameters of ANN studied in this work include as followed:
  - batch_size is the number of sub samples processed within each epoch before the weights are updated.
  - epoch is the number of times the entire training dataset is passed through the network. One epoch means that the training dataset is passed forward and backward through the neural network once. A smaller epoch results in underfitting, while a larger epoch leads to overfitting and is time-consuming. The number of epochs is typically 10, 100, 500, 1,000 or larger (Brownlee, 2018). This work uses 100 epochs.
  - dropout is a probability of reducing the number of nodes in each layer of the model to prevent overfitting.

- learning_rate is determining the step size at which the network updates its weights during training.
- num_layer is the number of hidden layers.
- num_unit is the number of neurons in hidden layers.
- activation is an activation function applied at hidden layer(s) and output layer. The activation function uses in this work is rectified linear activation function or ReLU due to non-negative output values and common use (Parhi & Nowak, 2020; Pratama & Kang, 2021).

- Steps of model construction as follows:
  (i) Import required libraries and load the dataset table
  (ii) Divide the entire dataset into two datasets including training dataset for 30 months (January 2020 – June 2022) and test dataset for 6 months (July 2022 – December 2022)
  (iii) Perform feature scaling of the numeric factors on training dataset then fit on test dataset
  (iv) Use training dataset to construct the machine learning model with grid search and 10-fold cross validation method to tune hyperparameters.
  (v) Apply the optimal setting of hyperparameters that provides the lowest average error measurement or highest average accuracy during the cross-validation process to the test dataset
  (vi) Validate and record the model performance by measurement metric on the test dataset

3.5.3 Hybrid model

Hybrid models can improve the model's forecasting or prediction performance, as demonstrated in several previous studies. This work performs two types of hybrid structures, including parallel hybrid structure to reduce the possibility of using an improper model and series hybrid structure to enhance

forecasting accuracy because of comprehensive pattern detection and modeling.

In the hybrid modelling, the best model that has the highest accuracy (lowest error measurement) is selected from each of the three algorithms, including random forest, XGBoost and ANN; the total is three single models. Then, the hybrid model is construct from the pairs of models which means each time the model is constructed, two of the three models are selected for modeling and changed until all combinations are completed.

The parallel hybrid model is performed as shown in Figure 17. After training and predicting by the selected two single models, the predictions of these models are passed as input to train the linear regression model. Subsequently, the final predictions are predicted, then evaluated and compare model performance.

Figure 18 shows the series hybrid model construction. Firstly, one of the two selected model, called the first model, are trained, predicted and calculated residuals. The residuals then are passed as input to another model, called second model, then the second model are trained and predicted. The final predictions are calculated by summing the predictions from the two models and then evaluated and compare model performance.

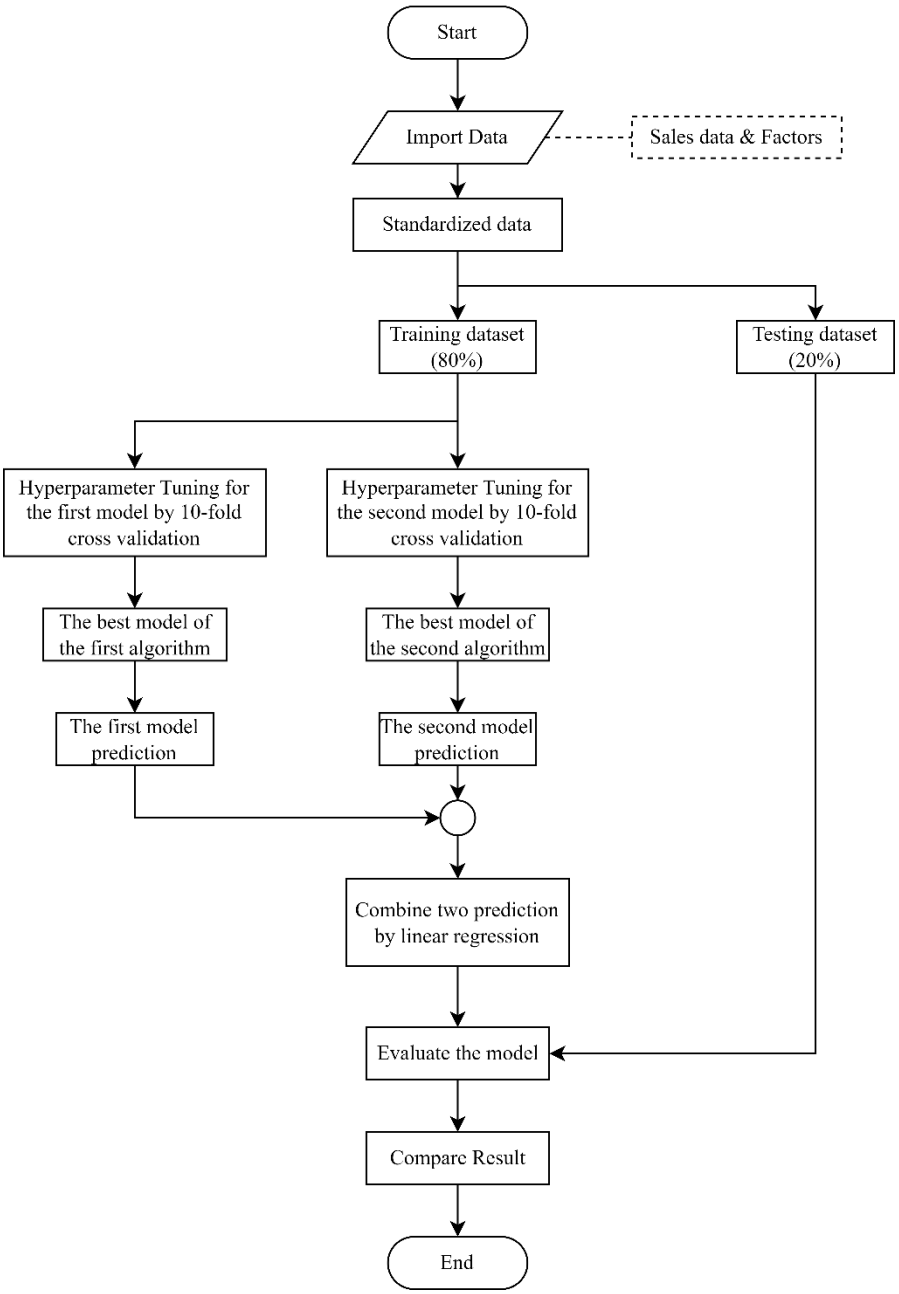


Figure 17 The methodology of parallel hybrid model

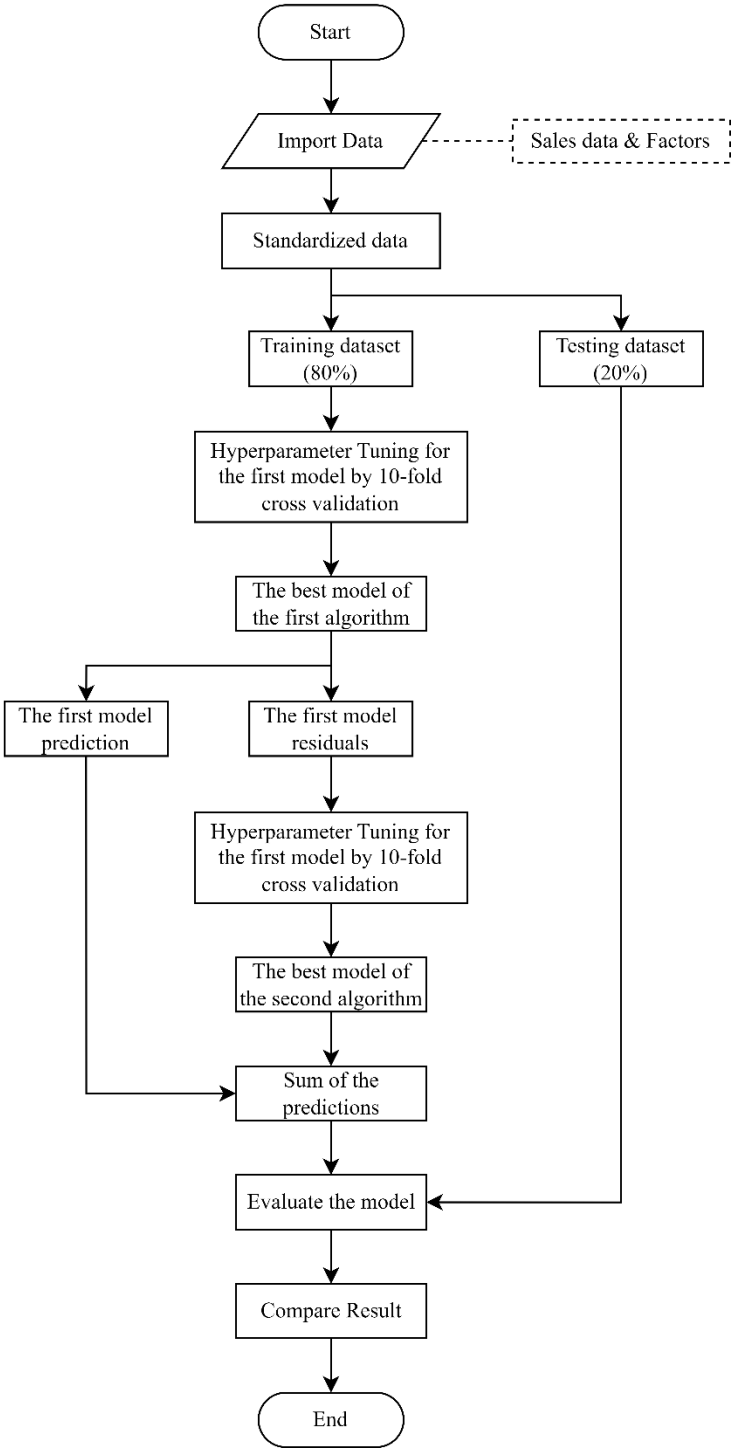


Figure 18 The methodology of series hybrid model

3.6 Result comparison

3.6.1 Model evaluation and selection

The performance of the models is measured using R^2 and MAPE. The R^2 of the linear regression model is used to compare how well the models fit. Moreover, the MAPE was conducted to evaluate the accuracy of the models. The lower the MAPE is, the more accurate the model. The weighted MAPE (WMAPE), using weight from sales revenue of the products, was calculated to compare overall performance of prediction models as given in Equation. (7). The model with the highest accuracy or the lowest WMAPE on the test set is selected to be the best model for sales prediction of beauty products.

$$WMAPE = \frac{\sum_{i=1}^{n} w_i MAPE_i}{\sum_{i=1}^{n} w_i} \quad (7)$$

where w_i is the weights of product i

$MAPE_i$ is the MAPE of product i

n is the number of products

3.6.2 Factor analysis

For the factors that affect sales of beauty products, these can be indicated and analyzed by stepwise linear regression and SHAP value. The SHAP value is applied to interpret the impact of factors on the predictions of the model that provide insights of relationships between the factors and the dependent variable.

Chapter 4 Results and Discussion

This chapter provides the results of data exploration, model prediction, and their performances.

4.1 Result of data exploration

According to the pareto chart shown in Figure 3, this study focused on the top 10 best sellers of the best-selling category, facial moisturizer, to predict sales quantity. The case-study retail company provides data on beauty products from January 2020 to December 2022 (a period of 36 months). After cleaning the dataset, the data on selecting products and factors is explored. Table 9 demonstrate the unit sales data summary of the 10 beauty products of the facial moisturizer category. The table includes the mean, standard deviation (std), min and max values, as well as the selling price. As shown in the table, the sales quantity of all the products are on a different scale and show high variation. The selling price of each product has varied values depending on price reduction promotions, and the highest value is the regular price or non-promotion price. Additionally, from the histogram of the 10 product's sales quantity in the facial moisturizer category, as displayed in Figure 19, most of the products have a right-skew distribution.

Table 9 Summary of the unit sales - Facial moisturizer category

Product name	Subcategory	GP	mean	std	min	max	Selling Price
SKU01	Basic skin care	Low	16,432.83	13,962.80	4,719	48,542	45, 47.5, 49
SKU02	Anti-aging	Low	11,272.69	9,301.54	2,915	34,878	29, 29.5, 39
SKU03	Anti-aging	Low	712.22	717.25	28	2,970	359, 419, 459, 499, 599
SKU04	Whitening	Medium	9,460.08	3,364.72	4,140	16,752	32, 34, 34.5, 35, 37.5, 39
SKU05	Men	Low	26,167.06	18,768.04	9,922	81,237	12.5, 15
SKU06	Whitening	Medium	25,451.44	22,256.88	6,352	95,791	10, 15
SKU07	UV protection	Medium	18,140.47	13,553.12	6,434	49,027	12.5, 15
SKU08	Anti-aging	Medium	14,812.92	11,325.26	3,630	36,992	15, 20
SKU09	Whitening	Medium	13,828.61	8,894.18	4,094	36,717	15, 20
SKU10	Whitening	Low	12,033.28	7,140.07	4,822	28,936	17.5, 20

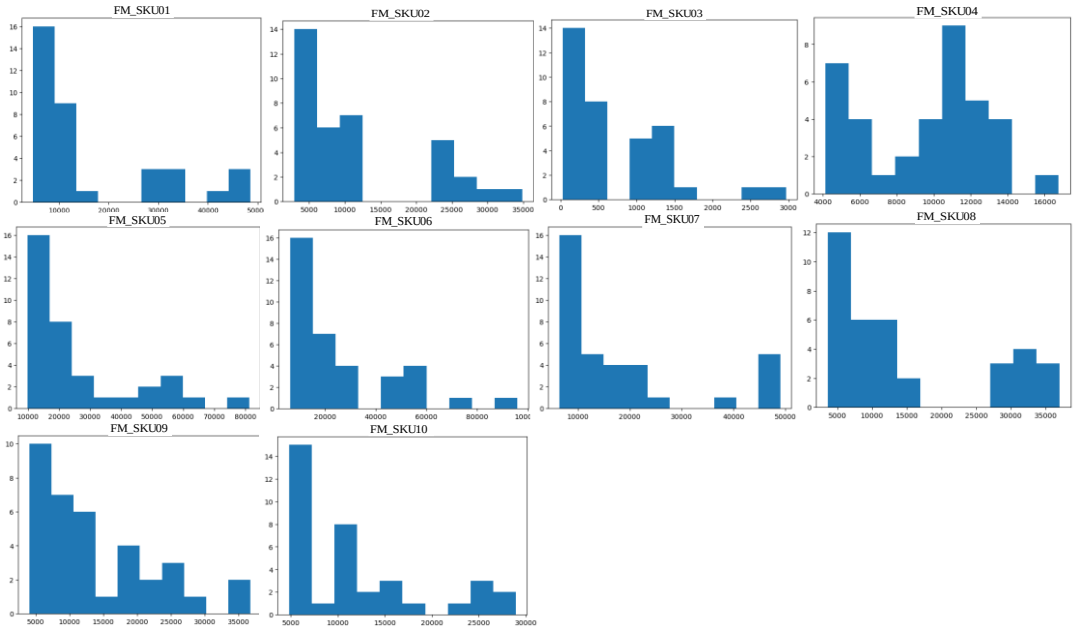


Figure 19 Histogram of facial moisturizer product's unit sales

Moreover, the number of COVID-19 new cases in Thailand is illustrated in Figure 20. There are two peaks of COVID-19 new cases during April 2021–September 2021 and February 2022–May 2022. From the subsidies and welfare programs offered by Thailand's government to stimulate the economy in October 2020–March 2021, May 2021, July 2021–December 2021, February 2022–April 2022, and September 2022–October 2022. These external factors could affect sales.

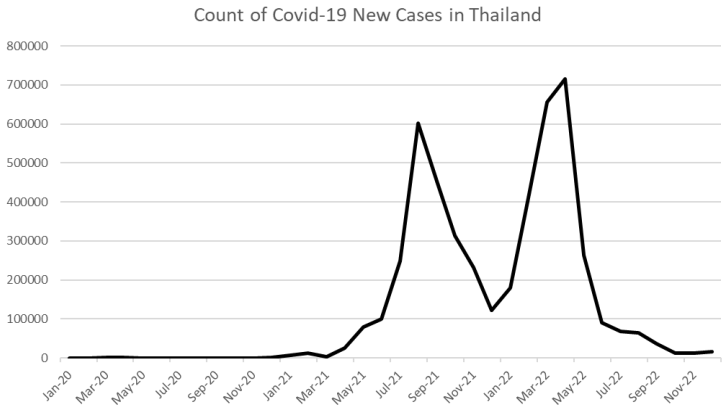


Figure 20 The number of COVID-19 new cases

4.2 Result of clustering methods

In this research, to compare model with and without considering factors of other products in the same group, beauty products in the facial moisturizer category were clustered into groups by three types including by category, by subcategory or by K-means method.

4.2.1 Clustering by category

According to the training dataset, Table 10 demonstrates the data summary of all the 79 beauty products of the facial moisturizer category.

Table 10 The data summary of all the facial moisturizer products

	Sales Quantity	Regular price	Selling price	Discount percentage	Promotion period	lag 1 promotion period	lag 2 promotion period
count	2370	2370	2370	2370	2370	2370	2370
mean	5386.92	119.82	113.54	4.43	0.22	0.21	0.20
std	9709.33	135.90	128.72	9.82	0.42	0.41	0.40
min	3.00	10.00	7.50	-23.08	0	0	0
max	95791.00	599.00	599.00	82.70	1	1	1
Count if event occurred					632	590	564

4.2.2 Clustering by subcategory

In the dataset, the company classifies all the beauty products of the facial moisturizer category into 9 groups according the subcategory type including Anti-acne, Anti-aging, Basic skin care, For eyes, Men, Melasma treatment, Toner, UV protection and Whitening. The 10 products that focused in this study are from Anti-aging, Basic skin care, Men, UV protection and Whitening subcategories. The data summary tables of the 5 subcategories are presented in Tables 11 – 15, respectively.

Table 11 The data summary of products in the Anti-aging subcategory

	Sales Quantity	Regular price	Selling price	Discount percentage	Promotion period	lag 1 promotion period	lag 2 promotion period
count	480	480	480	480	480	480	480
mean	3544.45	206.00	195.68	4.69	0.23	0.21	0.20
std	6384.13	195.46	185.64	9.45	0.42	0.41	0.40
min	9.00	20.00	15.00	0.00	0	0	0
max	35647.00	599.00	599.00	40.07	1	1	1
Count if event occurred					108	101	97

Table 12 The data summary of products in the Basic skin care subcategory

	Sales Quantity	Regular price	Selling price	Discount percentage	Promotion period	lag 1 promotion period	lag 2 promotion period
count	330	330	330	330	330	330	330
mean	3171.54	74.27	72.81	2.35	0.14	0.14	0.13
std	5260.45	43.13	43.22	7.96	0.35	0.35	0.34
min	61.00	25.00	12.50	0.00	0	0	0
max	43957.00	149.00	149.00	50.63	1	1	1
Count if event occurred					46	46	44

Table 13 The data summary of products in the Men subcategory

	Sales Quantity	Regular price	Selling price	Discount percentage	Promotion period	lag 1 promotion period	lag 2 promotion period
count	210	210	210	210	210	210	210
mean	9692.20	66.43	63.69	4.30	0.20	0.20	0.18
std	13040.92	90.34	86.58	9.27	0.40	0.40	0.38
min	113.00	12.00	10.00	0.00	0	0	0
max	81237.00	269.00	269.00	37.50	1	1	1
Count if event occurred					42	41	37

Table 14 The data summary of products in the UV protection subcategory

	Sales Quantity	Regular price	Selling price	Discount percentage	Promotion period	lag 1 promotion period	lag 2 promotion period
count	510	510	510	510	510	510	510
mean	5639.39	140.47	131.30	5.00	0.24	0.23	0.22
std	10484.64	131.87	124.17	9.98	0.43	0.42	0.41
min	15.00	10.00	7.50	0.00	0	0	0
max	95366.00	479.00	479.00	50.00	1	1	1
Count if event occurred					122	117	112

Table 15 The data summary of products in the Whitening subcategory

	Sales Quantity	Regular price	Selling price	Discount percentage	Promotion period	lag 1 promotion period	lag 2 promotion period
count	510	510	510	510	510	510	510
mean	8227.57	71.76	66.88	4.54	0.20	0.20	0.19
std	10848.70	103.84	95.95	10.39	0.40	0.40	0.39
min	11.00	10.00	7.50	0.00	0	0	0
max	80850.00	399.00	399.00	50.00	1	1	1
Count if event occurred					102	101	95

4.2.3 Clustering by K-means method

To cluster products into groups by K-means method, first the training dataset with sales quantity and price factor were scaled by StandardScaler. The model was constructed by ‘Kmeans’ in scikit-learn library to fit K-means method and using yellowbrick package to find the optimal number of cluster or K values. After fitting the model, an elbow plot is presented in Figure 21 which it suggests that the optimal number of clusters is 5 (K = 5). However, the 10 products that focused in this study are clustered into 3 groups as shown in Table 16. the summary data of the 3 groups are shown in Tables 17-19, respectively.

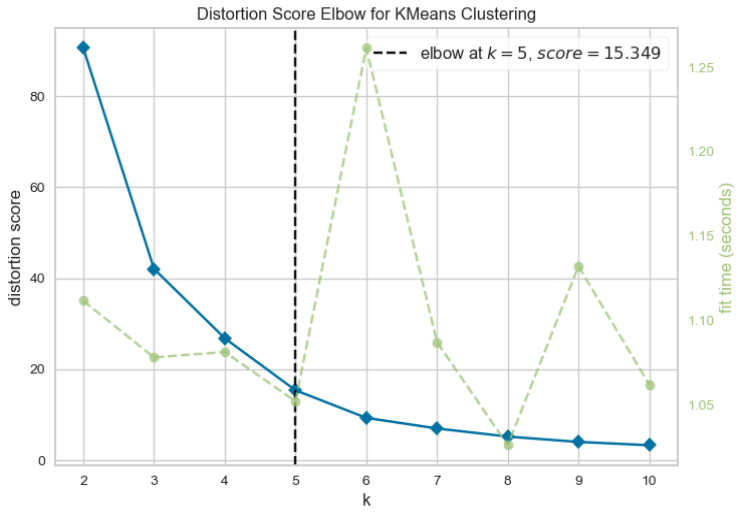


Figure 21 The elbow plot

Table 16 The assigned group of products using K-means method

Group	Total	SKU
1	21	1, 2, 4, 8-10, 13, 15, 28, 37, 43, 44, 46, 55, 61, 65, 67, 70, 72, 74, 78
2	4	3, 14, 24, 52
3	6	5-7, 47, 54, 69
4	34	11, 12, 16, 17, 21, 22, 25-27, 29-35, 38, 40-42, 45, 49, 50, 53, 56, 59, 63, 64, 66, 68, 75-77, 79
5	14	18-20, 23, 36, 39, 48, 51, 57, 58, 60, 62, 71, 73

Table 17 The data summary of products in group 1

	Sales Quantity	Regular price	Selling price	Discount percentage	Promotion period	lag 1 promotion period	lag 2 promotion period
count	630	630	630	630	630	630	630
mean	9366.45	25.62	24.60	3.89	0.19	0.18	0.17
std	7768.06	13.85	13.59	9.04	0.39	0.39	0.38
min	1571.00	10.00	7.50	0.00	0	0	0
max	53343.00	49.00	49.00	50.00	1	1	1
Count if event occurred					118	114	108

Table 18 The data summary of products in group 2

	Sales Quantity	Regular price	Selling price	Discount percentage	Promotion period	lag 1 promotion period	lag 2 promotion period
count	120	120	120	120	120	120	120
mean	381.04	551.50	517.59	6.13	0.31	0.29	0.28
std	481.34	50.90	72.38	9.82	0.46	0.46	0.45
min	9.00	479.00	359.00	0.00	0	0	0
max	2970.00	599.00	599.00	40.07	1	1	1
Count if event occurred					37	35	34

Table 19 The data summary of products in group 3

	Sales Quantity	Regular price	Selling price	Discount percentage	Promotion period	lag 1 promotion period	lag 2 promotion period
count	180	180	180	180	180	180	180
mean	24802.83	19.83	19.23	3.57	0.17	0.17	0.16
std	17167.76	8.79	9.08	8.62	0.38	0.38	0.36
min	6352.00	15.00	10.00	0.00	0	0	0
max	95366.00	39.00	39.00	37.50	1	1	1
Count if event occurred					31	31	28

4.3 Result of model prediction

4.3.1 Linear regression

By fitting the linear regression model on the training dataset with a stepwise method at alpha = 0.05 using the Minitab program to predict the sales quantity of the 10 beauty products in the facial moisturizer category and studying the important factors, the results are presented in Figures A1–A10 (see in Appendix). The finding found that the R^2 of all products was in the range of 42.56% and 96.13%, whose average value was 83.50%, as shown in Table 20. All products, except SKU04, had R^2 values higher than 80% indicating that more than 80% of the variation in the dependent variable can be explained by the independent variables. The models are good fits. The influencing factors of the products are shown in Table 21, where the number in the table means the coefficient of the regression equation, and the highlight color of green and red represents that the factor affects sales quantity positively and negatively to sales quantity.

Table 20 The R^2 of the 10 beauty products

Product name	Subcategory	R-sq(adj)
SKU01	Basic skin caere	83.84%
SKU02	Anti-aging	94.43%
SKU03	Anti-aging	96.13%
SKU04	Whitening	42.56%
SKU05	Men	84.32%
SKU06	Whitening	92.40%
SKU07	UV protection	83.22%
SKU08	Anti-aging	93.03%
SKU09	Whitening	83.25%
SKU10	Whitening	81.86%
	Average	83.50%

Table 21 The influencing factors to sales quantity of the 10 beauty products from stepwise method

	Product name									
Factors	SKU01	SKU02	SKU03	SKU04	SKU05	SKU06	SKU07	SKU08	SKU09	SKU10
Price	-4187	-2126								
Month_1										
Month_2		3042								
Month_3			326.7							
Month_4										
Month_5										
Month_6										
Month_7										
Month_8										
Month_9								4522		
Month_10										
Month_11										
COVID-19					-0.01616	-0.02949		-0.01534		-0.01217
Store	27.98	-23	-0.852	-13.04					-17.31	
Welfare			121.1			7300		2732	3757	2831
%Discount			65.07				1940		534.8	
Promotion period			-345		41330	35846		25640		16591
Promotion period Lag1										
Promotion period Lag2	-6450			-3466						

Note: the number is coefficient and the green and red color mean positive and negative effect

From Table 21, the result found that the selected factors were different for each product. However, some variables had a similar effect on sales quantity. For the price and promotion characteristic factors, most of the products had either a price factor, a discount percentage factor, or a promotion period factor that affected sales, excluding SKU03, for which both the discount percentage and promotion period were significant. According to the results, the price factor in SKU01 and SKU02 has a negative effect on sales quantity, which means a higher price leads to a lower sales quantity. The promotion characteristic factors, including discount percentage and promotion period, have a positive effect on sales quantity, except SKU03. The higher the discount percentage, the greater the increase in sales quantity. Promotion also increases sales quantity. The monthly period factor may not influence sales for most products, as the findings show that it is insignificant. For the external variables, including subsidies and welfare programs and COVID-19 new cases, they had positive and negative effects on sales quantity, respectively. Having subsidies or welfare programs can help increase sales. On the other hand, when the COVID-19 virus is widespread, leading to an increase in COVID-19 new cases, sales will decrease.

After the model was fit to the training dataset and then applied to the testing dataset, Table 22 shows the MAPE and WMAPE results. The results revealed that the WMAPE, weighted by product revenue, of the training dataset and the testing dataset were 32.68% and 43.92%, respectively. Additionally, the predictions for the products as compared to actual sales are presented in Figures 22–31.

Table 22 The MAPE and WMAPE on training and testing dataset of the 10 beauty products

Product name	Subcategory	Weight	MAPE	
			Train	Test
SKU01	Basic skin caere	0.22	29.62%	24.16%
SKU02	Anti-aging	0.11	23.31%	74.26%
SKU03	Anti-aging	0.10	85.52%	52.09%
SKU04	Whitening	0.10	26.95%	43.87%
SKU05	Men	0.10	23.84%	36.18%
SKU06	Whitening	0.09	26.36%	47.70%
SKU07	UV protection	0.07	35.66%	22.51%
SKU08	Anti-aging	0.07	29.78%	63.77%
SKU09	Whitening	0.07	21.00%	51.81%
SKU10	Whitening	0.07	22.89%	48.71%
		WMAPE	32.68%	43.92%

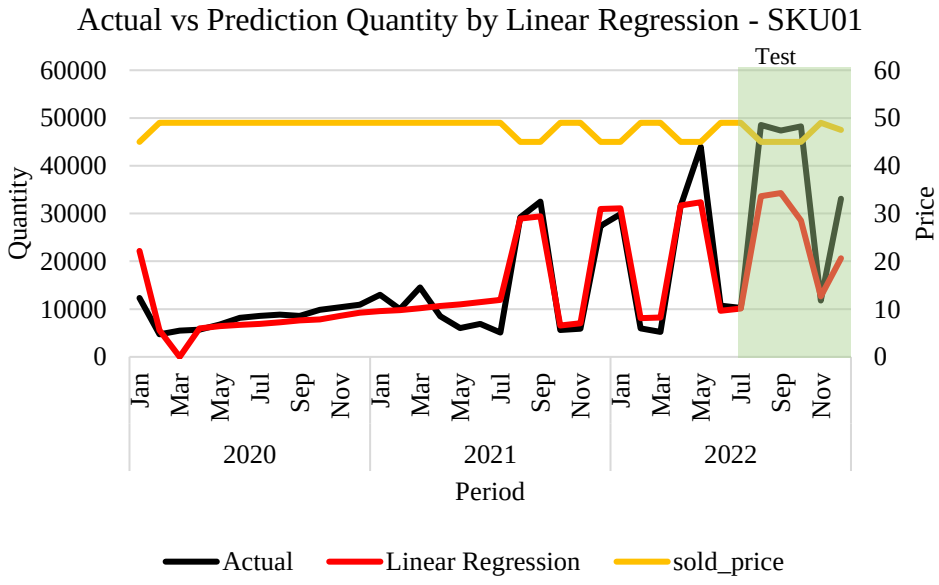


Figure 22 The prediction of SKU01 by linear regression

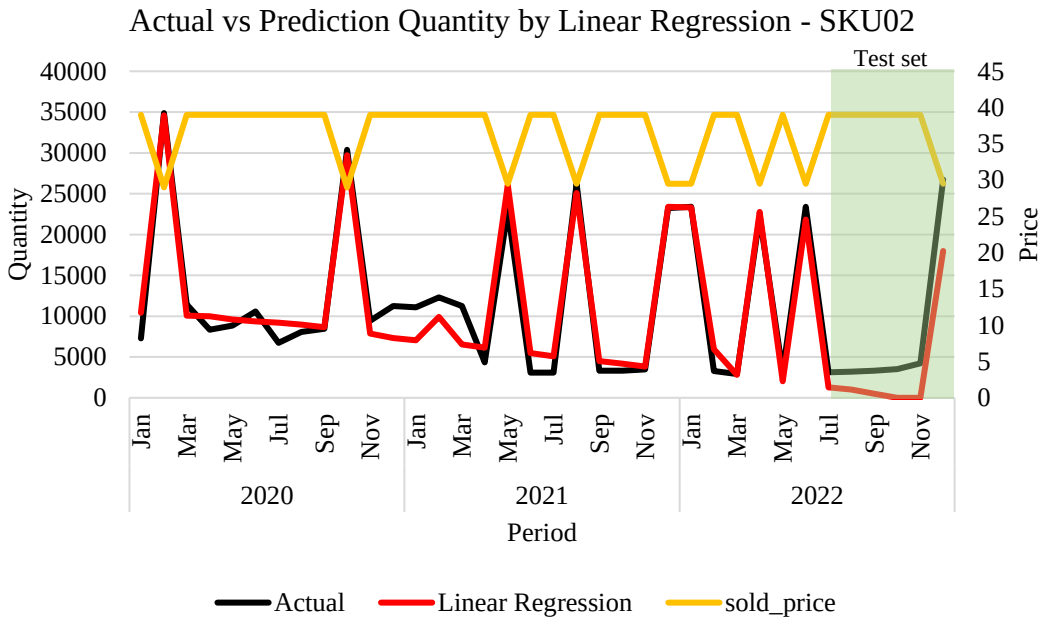


Figure 23 The prediction of SKU02 by linear regression

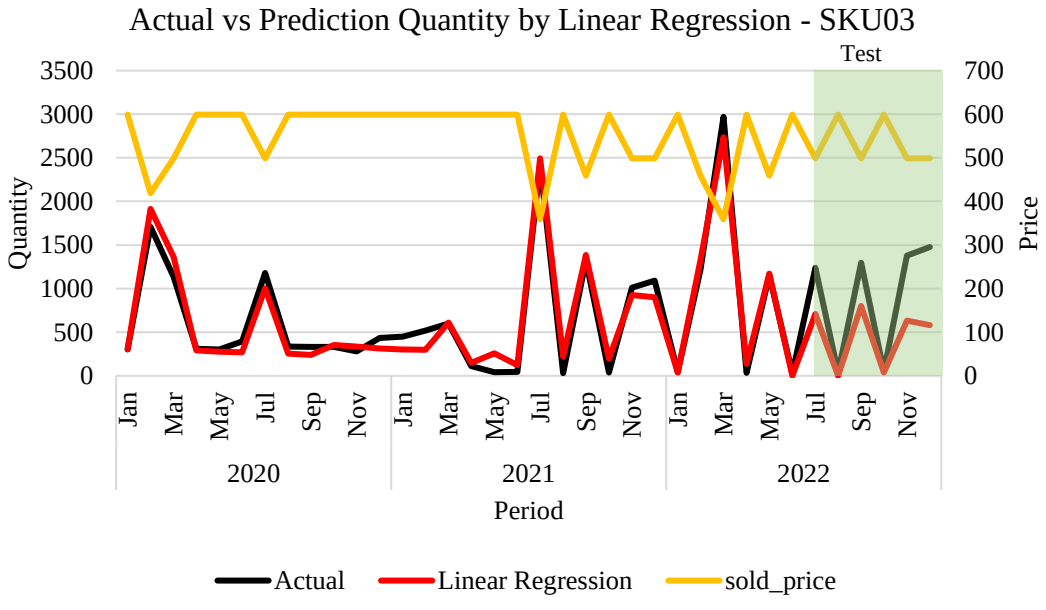


Figure 24 The prediction of SKU03 by linear regression

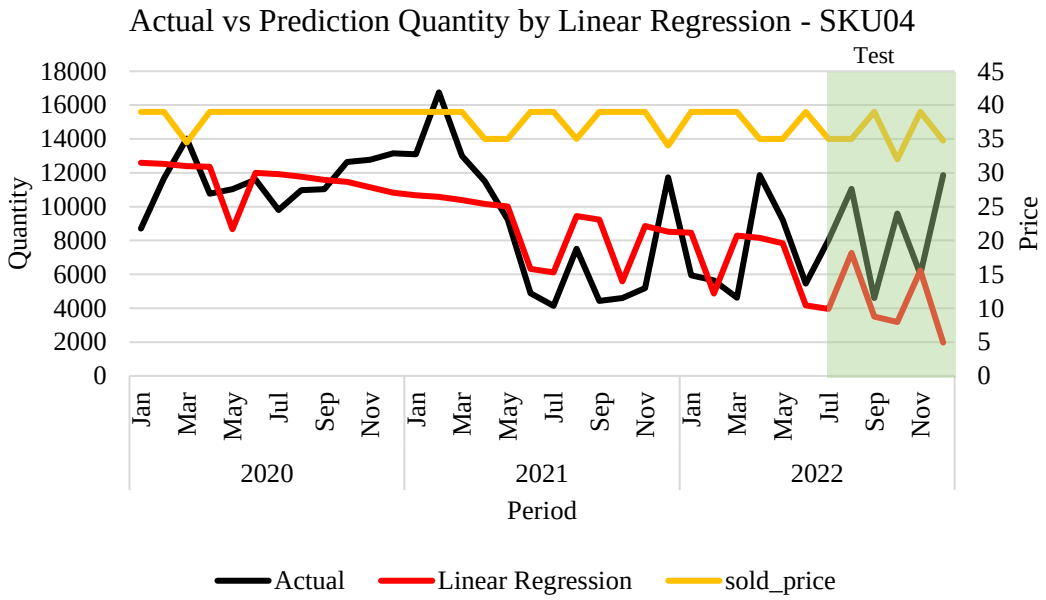


Figure 25 The prediction of SKU04 by linear regression

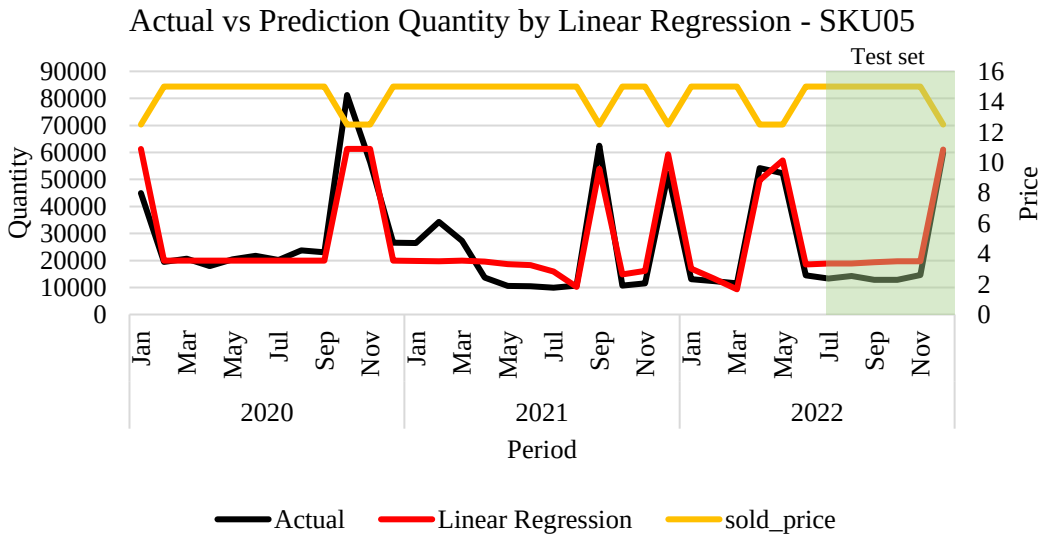


Figure 26 The prediction of SKU05 by linear regression

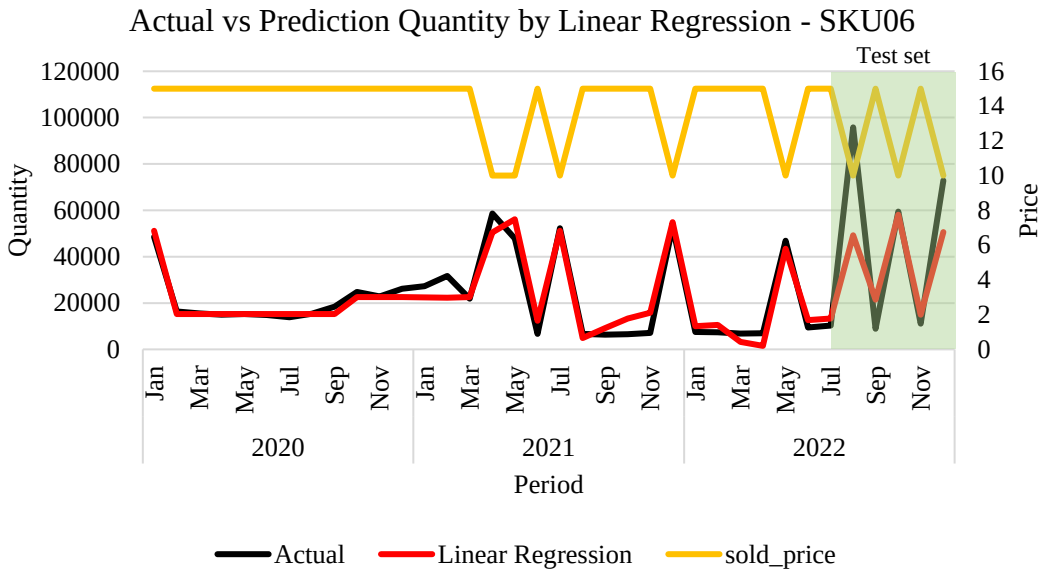


Figure 27 The prediction of SKU06 by linear regression

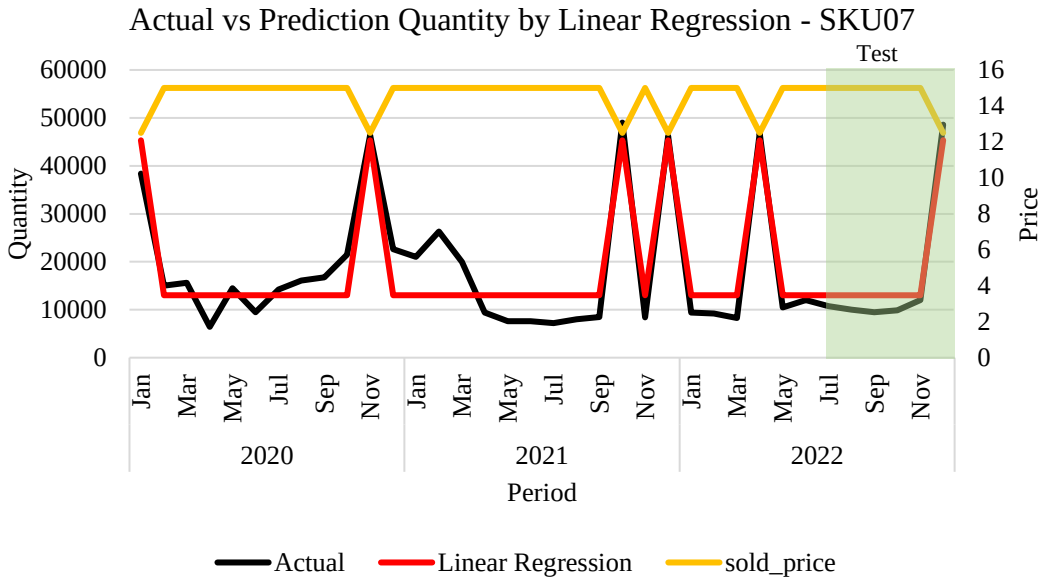


Figure 28 The prediction of SKU07 by linear regression

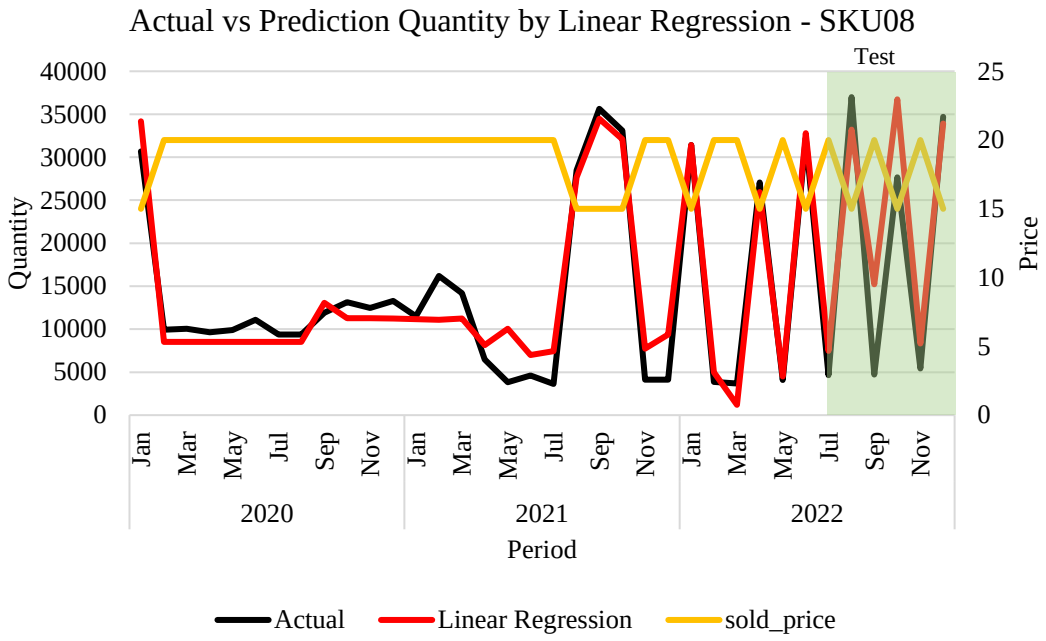


Figure 29 The prediction of SKU08 by linear regression

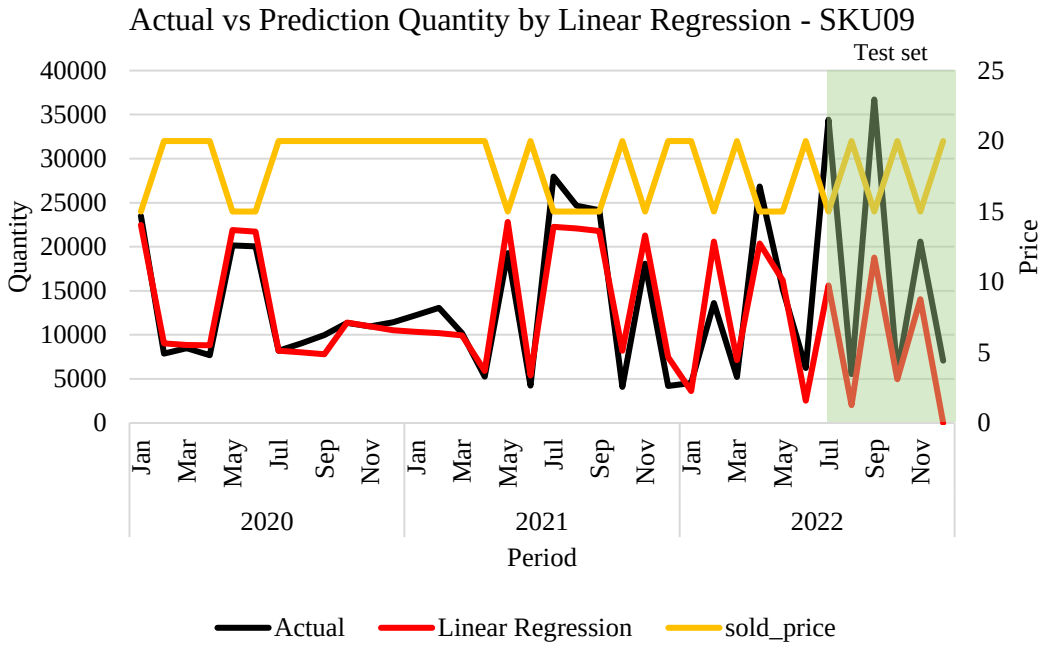


Figure 30 The prediction of SKU09 by linear regression

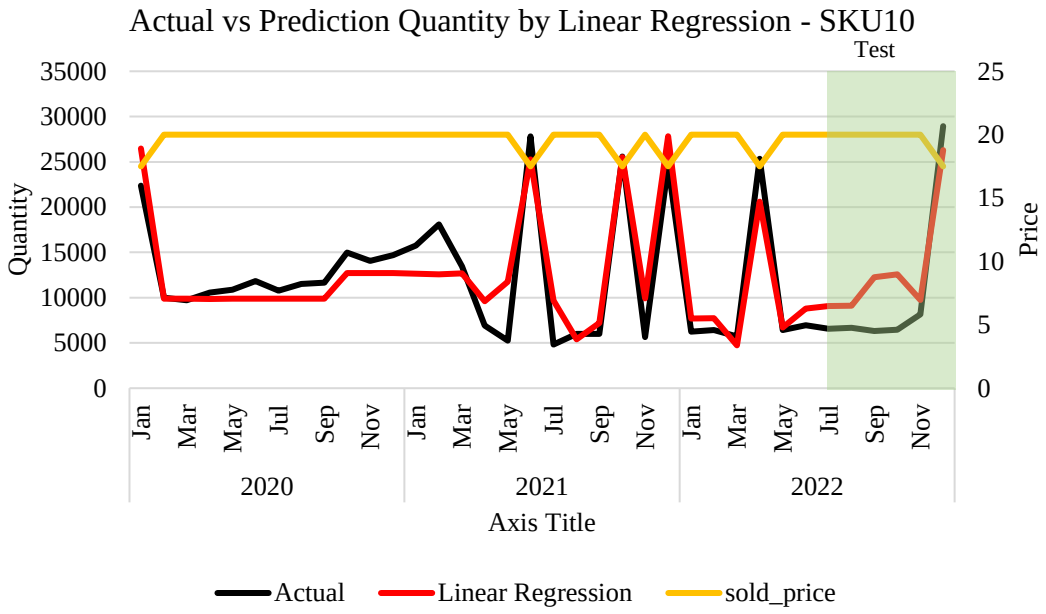


Figure 31 The prediction of SKU10 by linear regression

4.3.2 The machine learning model

The machine learning techniques, including random forest, XGBoost and ANN, were studied and constructed to predict sales of the 10 products. To train the models, all the independent variables listed in 3.2 and the significant factors from stepwise regression with and without clustering other products (see in 3.4) were used as the input in the models. The numeric features were scaled, then grid search and the 10-fold cross-validation method were used to tune the hyperparameters of each model until they gave the lowest mean squared errors. Then, the model is evaluated and compared results.

4.3.2.1 Random forest result

After creating models and tuning hyperparameters, WMAPEs, measuring overall prediction performance, were calculated and summarized, as shown in Table 23. The result revealed that using only significant factors for each product and not considering exogenous products' factors gave the lowest WMAPE on the testing dataset, which was 28.15%. It seems that using significant factors may be better than using all the factors as the lower WMAPE in most cases, and clustering products by

subcategory or by the K-means method provided a lower WMAPE compared to by category.

Table 23 Summarized results of WMAPE with different factors using random forest model

	Factor		WMAPE		
Cases	**Each product**	**Other products**	Train	CV	Test
1	All factors	Not consider other products's factors	11.52%	19.31%	39.08%
2	All factors	Consider factors of other products in the same group by category	18.83%	33.69%	45.75%
3	All factors	Consider factors of other products in the same group by subcategory	14.62%	25.60%	31.80%
4	All factors	Consider factors of other products in the same group by K-means	12.54%	22.25%	34.58%
5	Significant factors	Not consider other products's factors	13.77%	20.91%	28.15%
6	Significant factors	Consider factors of other products in the same group by category	20.79%	37.33%	39.47%
7	Significant factors	Consider factors of other products in the same group by subcategory	14.73%	24.84%	33.54%
8	Significant factors	Consider factors of other products in the same group by K-means	15.33%	24.75%	30.26%

According to Table 23, the overall lowest WMAPE or the best case were analyzed. The selected hyperparameters of all the products and the summarized results are shown in Tables 24 and 25, respectively.

Table 24 Selected hyperparameters of the random forest model of each product

Hyperparameters	**Value or Range**	**Product name**									
		SKU01	SKU02	SKU03	SKU04	SKU05	SKU06	SKU07	SKU08	SKU09	SKU10
n_estimators	[10, 15, 20]	20	20	20	10	20	15	20	20	20	20
max_depth	[3, 5, 7, 9]	7	3	7	5	3	5	3	5	5	5
min_samples_split	[2, 5, 10]	2	2	2	5	10	5	2	5	2	5
min_samples_leaf	[2, 5, 10]	2	2	2	2	2	2	2	2	2	2
max_leaf_nodes	[3, 5, 10]	10	10	10	5	3	5	3	5	10	10

Table 25 Results of the 10 beauty products using random forest model

Product name	**Subcategory**	**Weight**	**Runtime (sec)**	**MAPE**			**MSECV/MSETrain**
				Train	**CV**	**Test**	
SKU01	Basic skin caere	0.22	173.33	15.07%	25.15%	33.99%	1.04
SKU02	Anti-aging	0.11	297.78	9.02%	16.03%	8.52%	1.23
SKU03	Anti-aging	0.10	317.03	15.64%	29.02%	15.36%	1.24
SKU04	Whitening	0.10	351.36	14.89%	22.85%	31.81%	1.30
SKU05	Men	0.10	481.75	12.79%	15.80%	29.09%	0.97
SKU06	Whitening	0.09	496.60	8.15%	11.79%	55.22%	1.21
SKU07	UV protection	0.07	528.71	35.84%	39.23%	19.93%	1.12
SKU08	Anti-aging	0.07	506.28	7.85%	11.66%	27.55%	1.24
SKU09	Whitening	0.07	459.42	10.89%	18.36%	20.85%	1.31
SKU10	Whitening	0.07	433.80	7.01%	12.64%	33.48%	1.53
			WMAPE	13.77%	20.91%	28.15%	

From Table 25, the runtime of the random forest model for each product was around 2 to 8 minutes, which was an average of 6 minutes. The fractions of the

average squared error of the cross validation and training set of the products were approximately 1.0 to 1.5, which is not overfitting. The predictions for the products are presented as compared to actual sales in Figures 32-41.

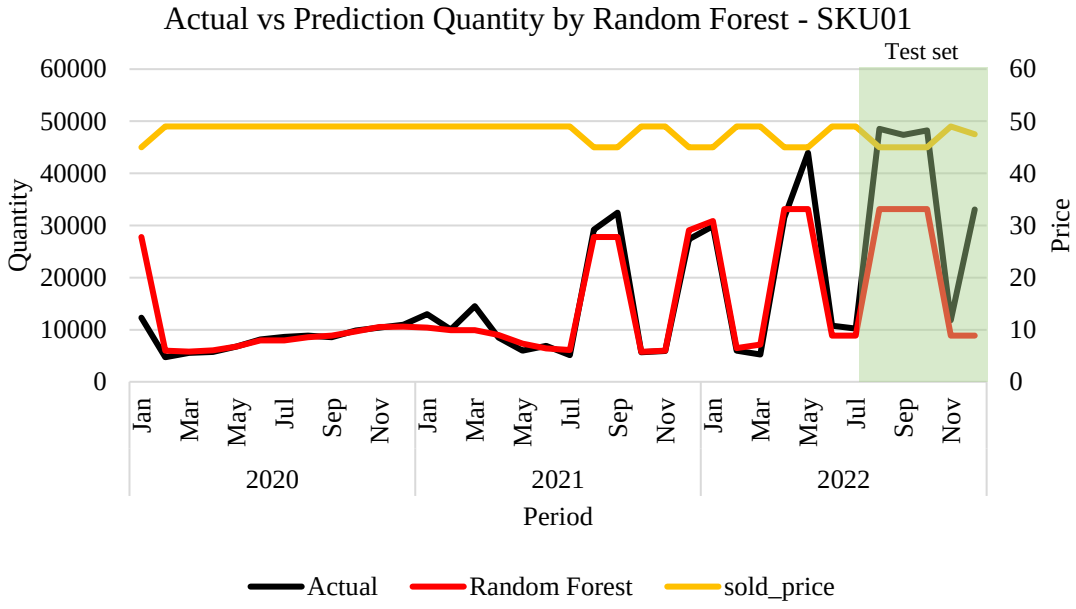


Figure 32 The prediction of SKU01 by random forest model

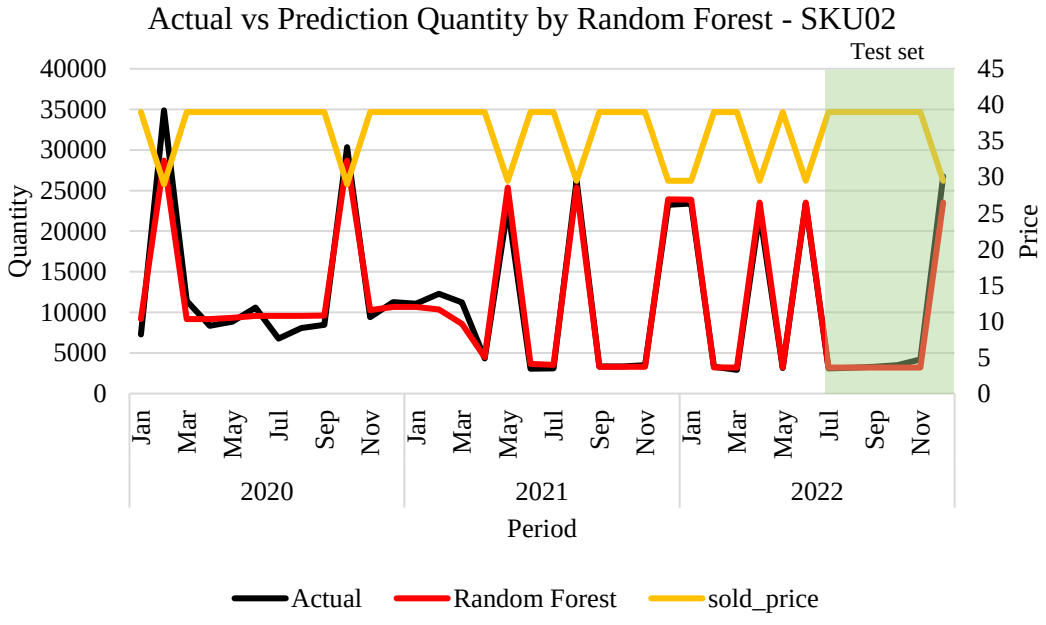


Figure 33 The prediction of SKU02 by random forest model

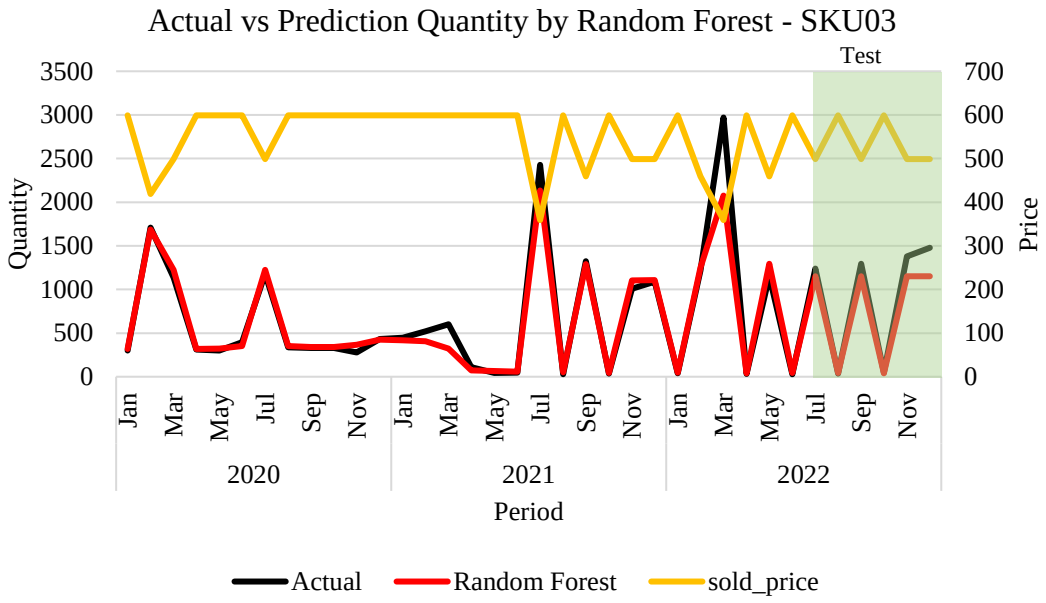


Figure 34 The prediction of SKU03 by random forest model

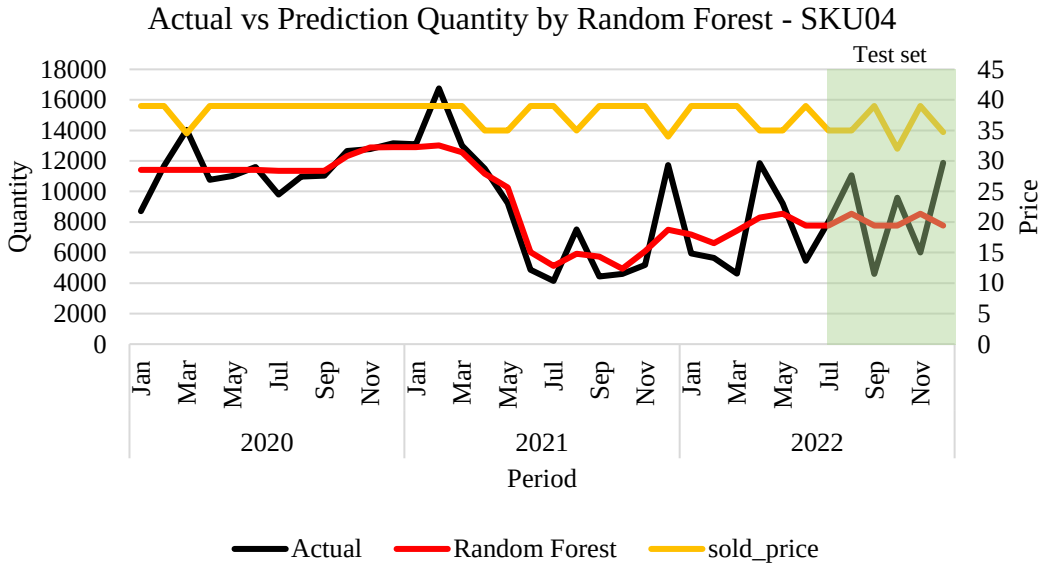


Figure 35 The prediction of SKU04 by random forest model

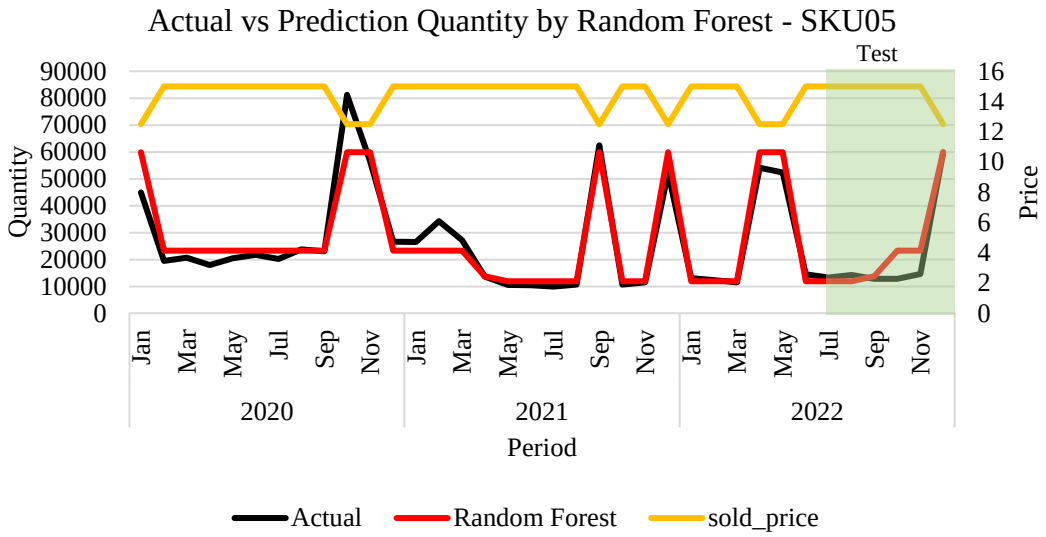


Figure 36 The prediction of SKU05 by random forest model

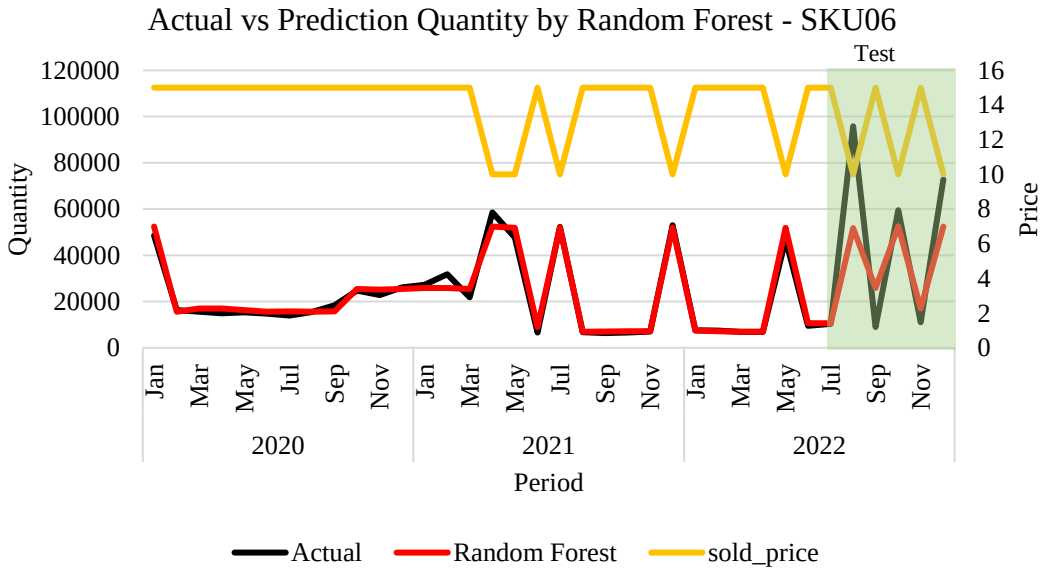


Figure 37 The prediction of SKU06 by random forest model

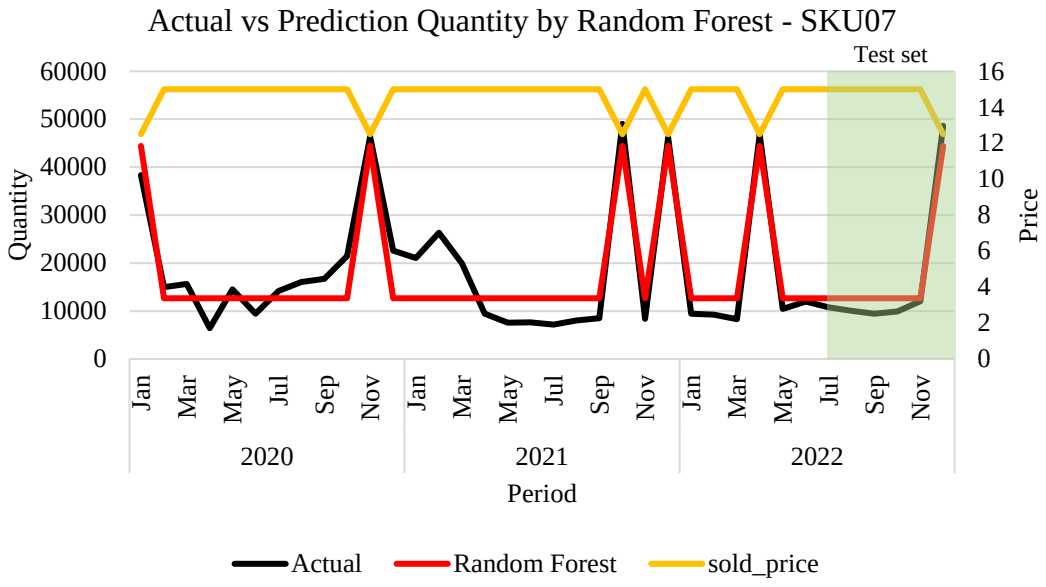


Figure 38 The prediction of SKU07 by random forest model

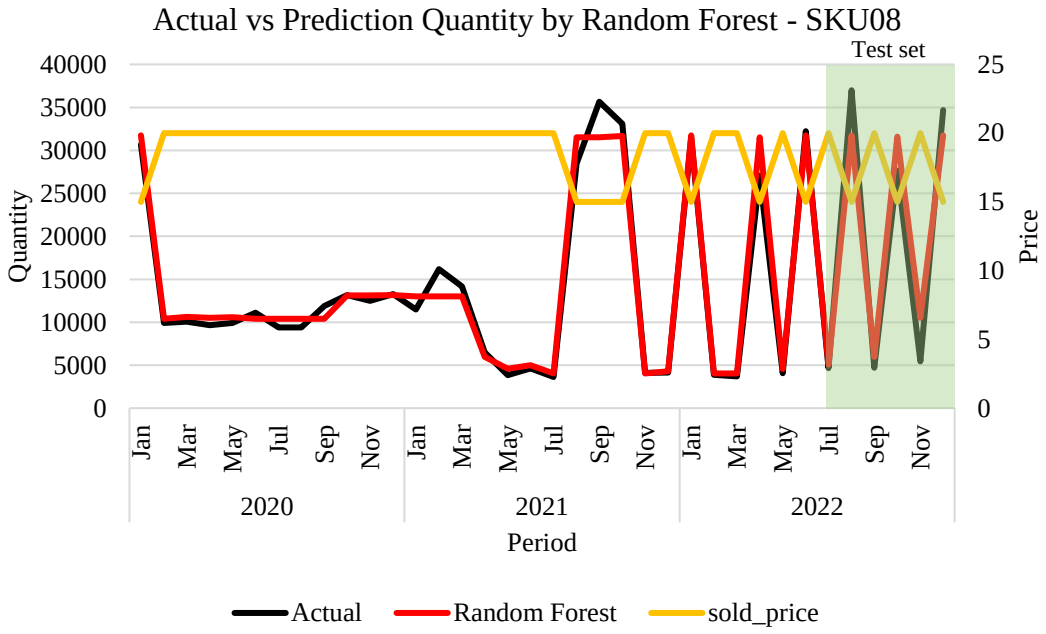


Figure 39 The prediction of SKU08 by random forest model

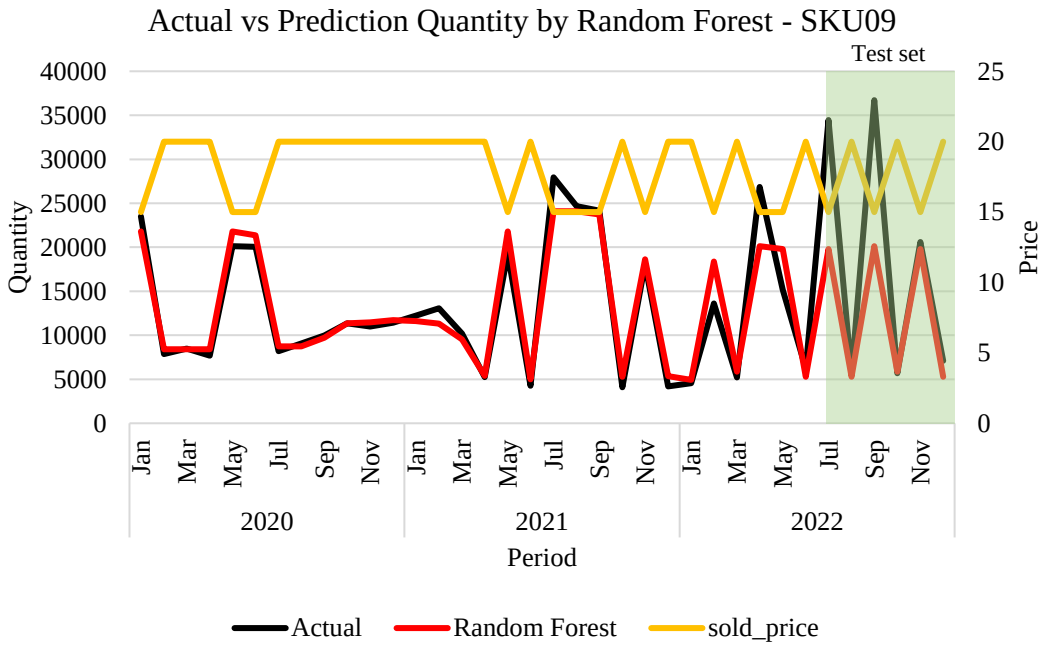


Figure 40 The prediction of SKU09 by random forest model

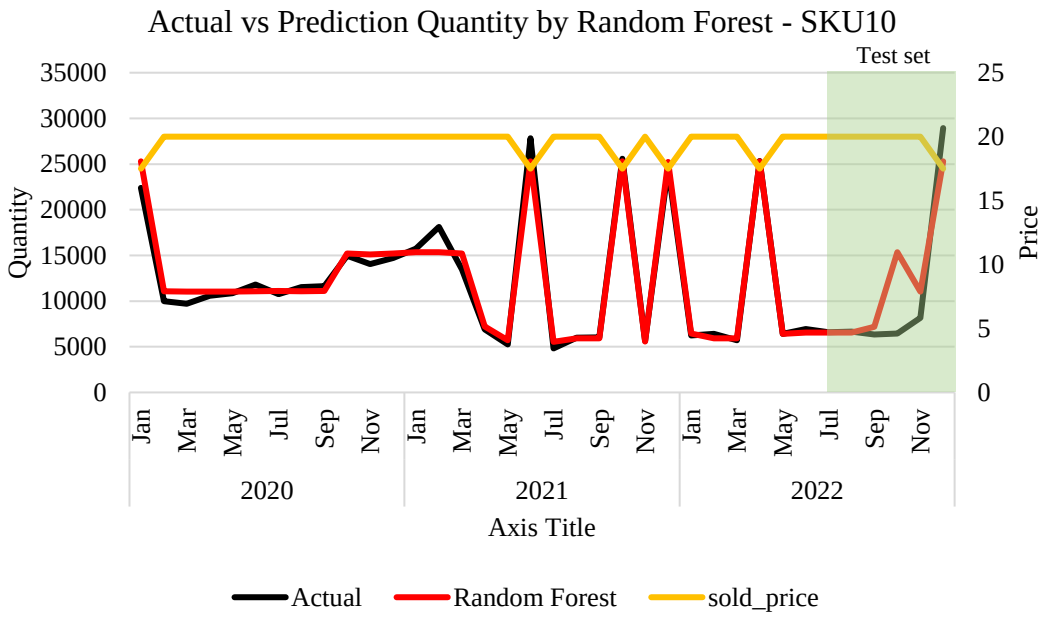


Figure 41 The prediction of SKU10 by random forest model

4.3.2.2 XGBoost result

Table 26 shows the summary of the WMAPE of the XGBoost models of different cases. The model that used significant factors from stepwise regression had a lower WMAPE on the testing dataset than the model that used all of the factors. It looks like using significant factors might be better than using all of them. For the highest overall prediction accuracy among the XGBoost models, use significant factors from stepwise regression and consider factors from other products in the same subcategory with a WMAPE of 32.60%.

Table 26 Summarized results of WMAPE with different factors using XGBoost model

	Factor		WMAPE		
Cases	**Each product**	**Other products**	Train	CV	Test
1	All factors	Not consider other products's factors	4.74%	24.38%	57.43%
2	All factors	Consider factors of other products in the same group by category	7.29%	35.57%	56.02%
3	All factors	Consider factors of other products in the same group by subcategory	5.27%	31.97%	51.24%
4	All factors	Consider factors of other products in the same group by K-means	6.32%	29.15%	53.47%
5	Significant factors	Not consider other products's factors	10.02%	18.29%	33.98%
6	Significant factors	Consider factors of other products in the same group by category	4.67%	32.08%	34.76%
7	Significant factors	Consider factors of other products in the same group by subcategory	6.10%	25.66%	32.60%
8	Significant factors	Consider factors of other products in the same group by K-means	8.06%	21.09%	37.54%

According to Table 26, the overall lowest WMAPE or the best case were analyzed. The selected hyperparameters of all the products and the summarized results are shown in Tables 27 and 28, respectively.

Table 27 Selected hyperparameters of the XGBoost model of each product

Hyperparameters	**Value or Range**	**Product name**									
		SKU01	SKU02	SKU03	SKU04	SKU05	SKU06	SKU07	SKU08	SKU09	SKU10
n_estimators	[10, 15, 20]	20	20	20	20	15	20	20	15	20	15
max_depth	[3, 5]	3	5	5	5	3	3	5	3	3	3
min_child_weight	[1, 2, 3, 5]	1	1	1	5	3	3	3	1	5	1
eta	[0.01, 0.03, 0.1, 0.3]	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
subsample	[0.5, 0.7, 1.0]	0.5	0.7	1	1	1	1	1	1	0.7	1
gamma	[0, 0.5, 1]	0	0	0	0	0	0	0	0	0	0
reg_lambda	[1, 2, 3, 5]	1	1	1	5	3	1	2	3	1	1

Table 28 Results of the 10 beauty products using XGBoost model

Product name	Subcategory	Weight	Runtime (sec)	MAPE			MSECV/MSETrain
				Train	CV	Test	
SKU01	Basic skin caere	0.22	1,669.48	9.46%	25.27%	31.01%	1.89
SKU02	Anti-aging	0.11	4,068.64	2.72%	20.99%	26.73%	2.77
SKU03	Anti-aging	0.10	1,942.89	1.80%	67.67%	29.95%	9.32
SKU04	Whitening	0.10	3,372.78	10.13%	26.11%	30.45%	1.84
SKU05	Men	0.10	1,717.51	7.04%	16.40%	28.92%	1.26
SKU06	Whitening	0.09	3,907.37	2.67%	12.79%	58.85%	3.00
SKU07	UV protection	0.07	1,940.80	3.45%	30.76%	16.22%	5.20
SKU08	Anti-aging	0.07	3,974.63	4.12%	13.87%	34.54%	1.61
SKU09	Whitening	0.07	3,434.19	11.97%	23.86%	35.00%	1.69
SKU10	Whitening	0.07	3,325.83	2.96%	12.78%	38.71%	2.56
			WMAPE	6.10%	25.66%	32.60%	

From Table 28, the runtime of the XGBoost model for each product was around 30 to 60 minutes, which was an average of 40 minutes. Most of the products had a fraction of the average squared error of the cross validation and training set of approximately 1.0 to 3.0, which is not overfitting. However, SKU03 and SKU07 had a high proportion of average error between the cross validation and the training dataset, so the models seemed to be overfit on the training dataset. The predictions for the products are presented in Figures 42–51.

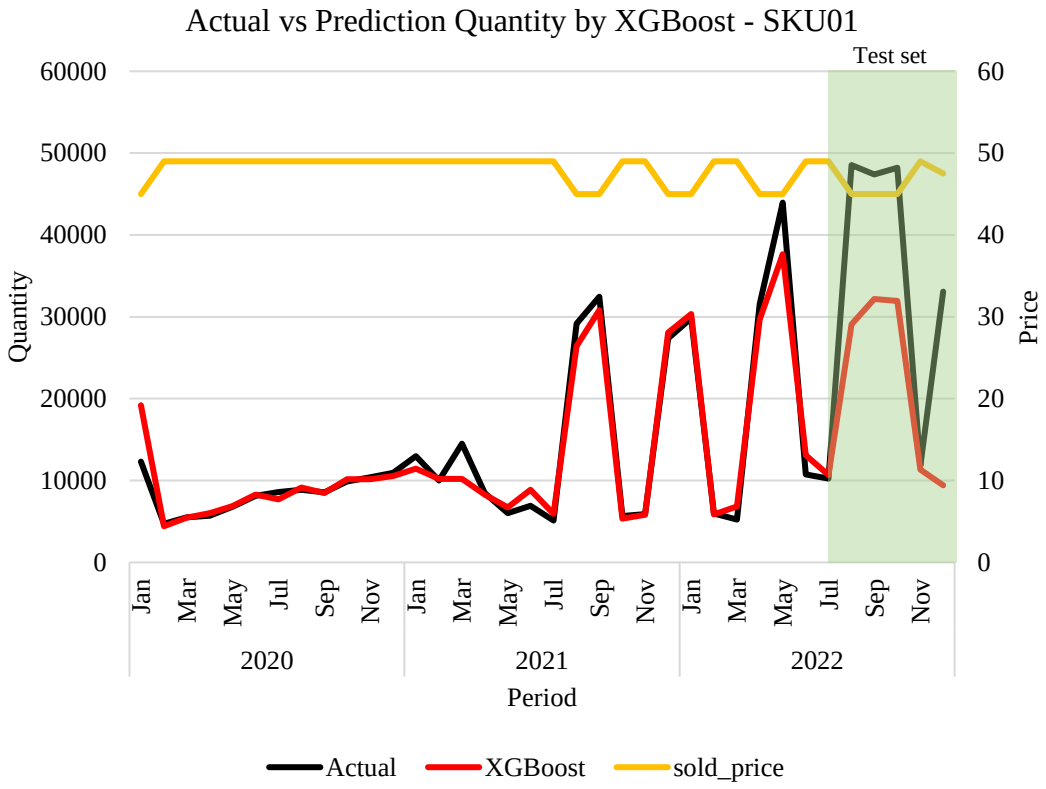


Figure 42 The prediction of SKU01 by XGBoost model

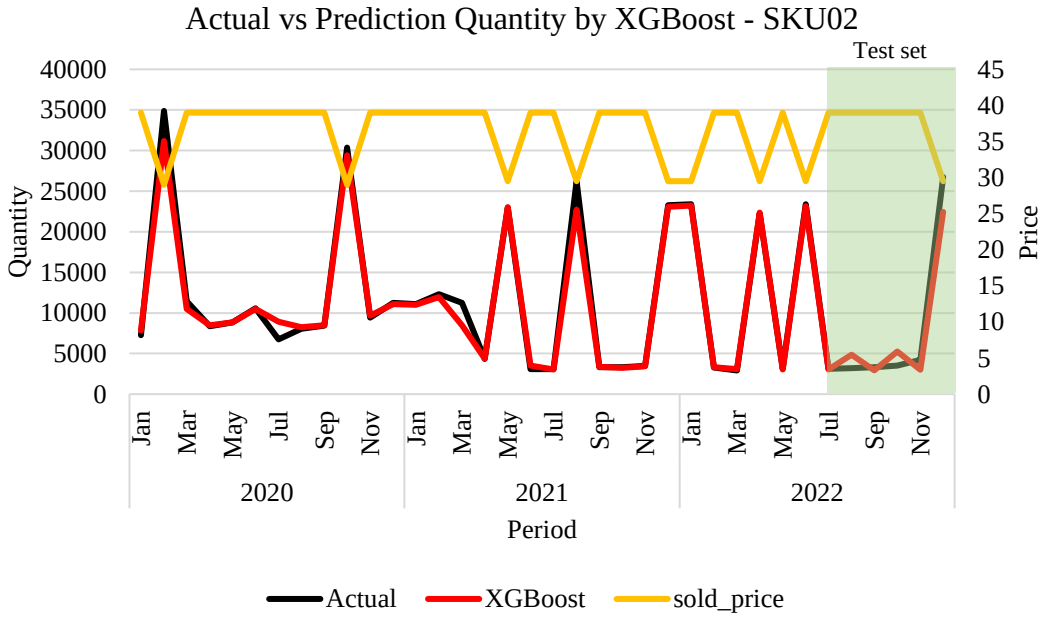


Figure 43 The prediction of SKU02 by XGBoost model

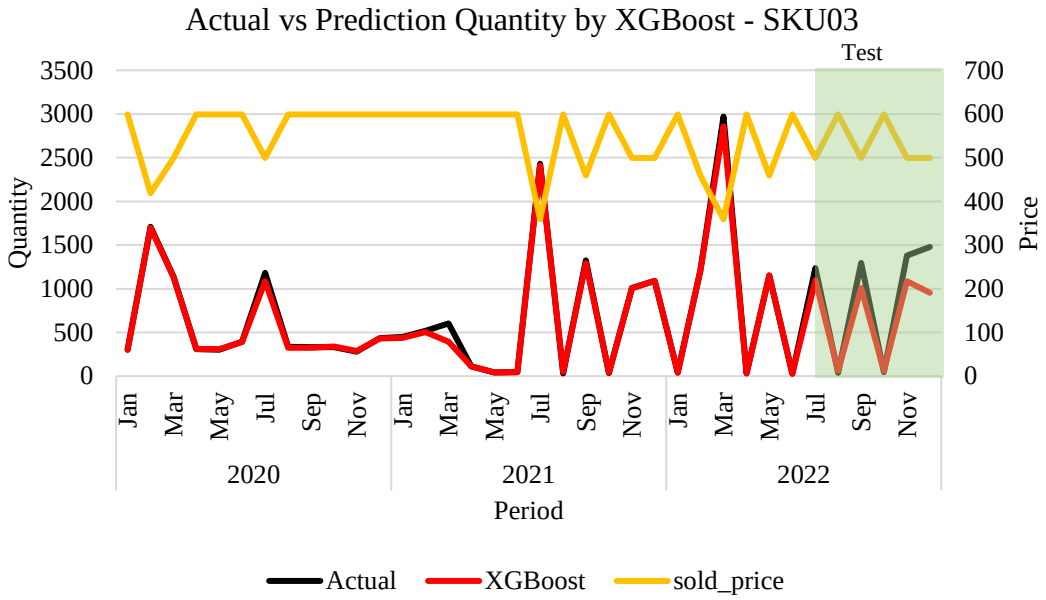


Figure 44 The prediction of SKU03 by XGBoost model

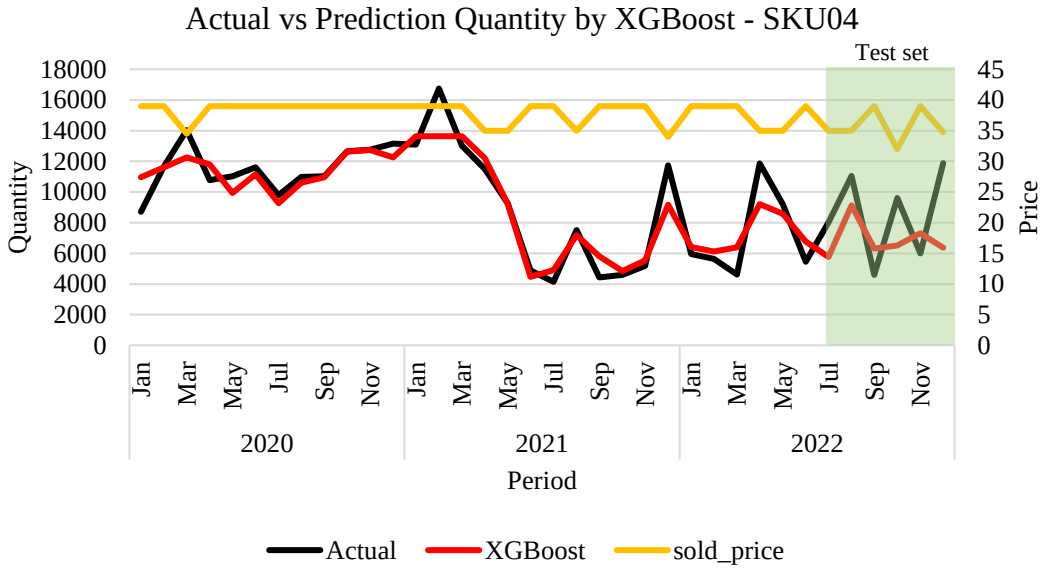


Figure 45 The prediction of SKU04 by XGBoost model

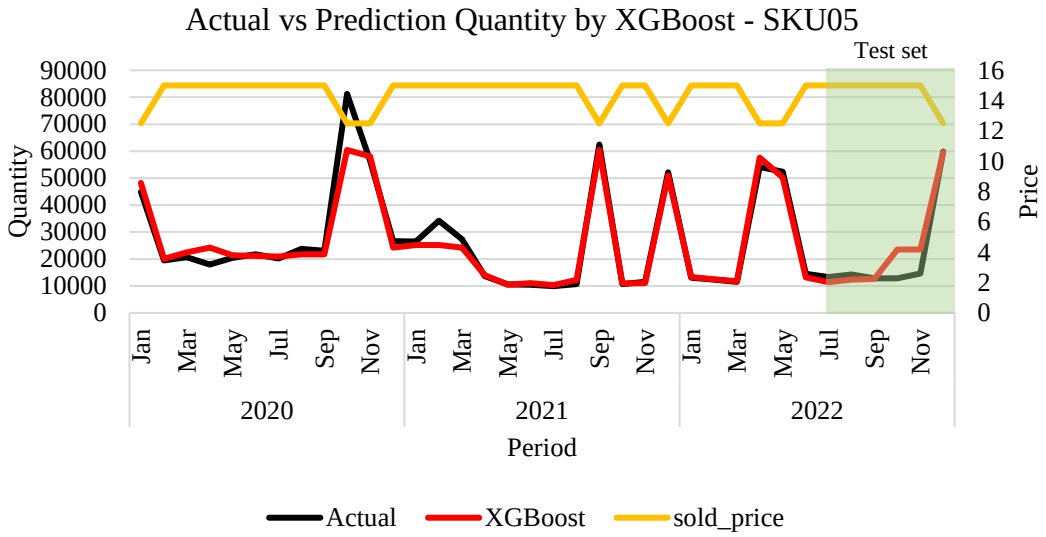


Figure 46 The prediction of SKU05 by XGBoost model

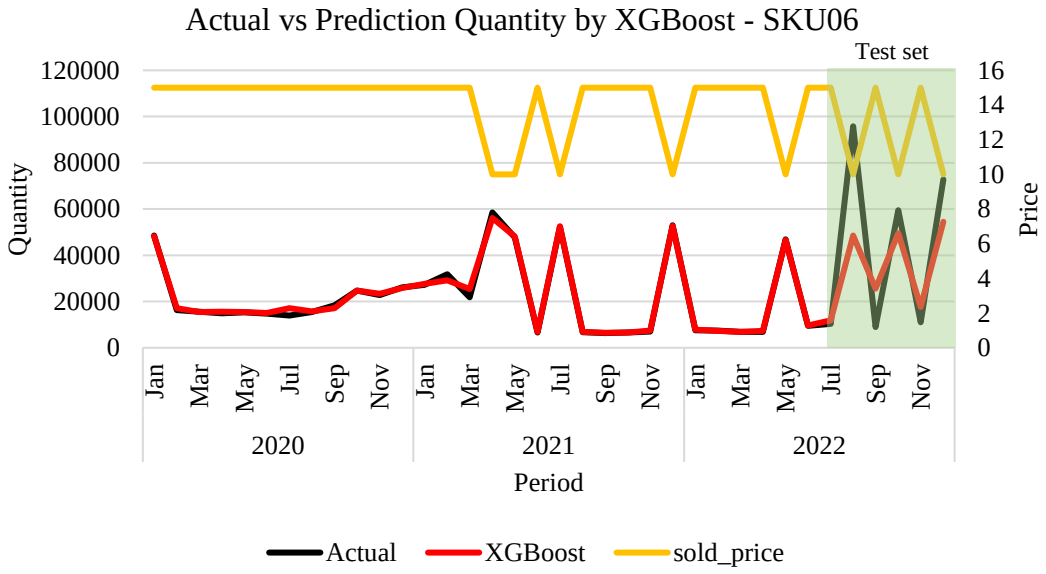


Figure 47 The prediction of SKU06 by XGBoost model

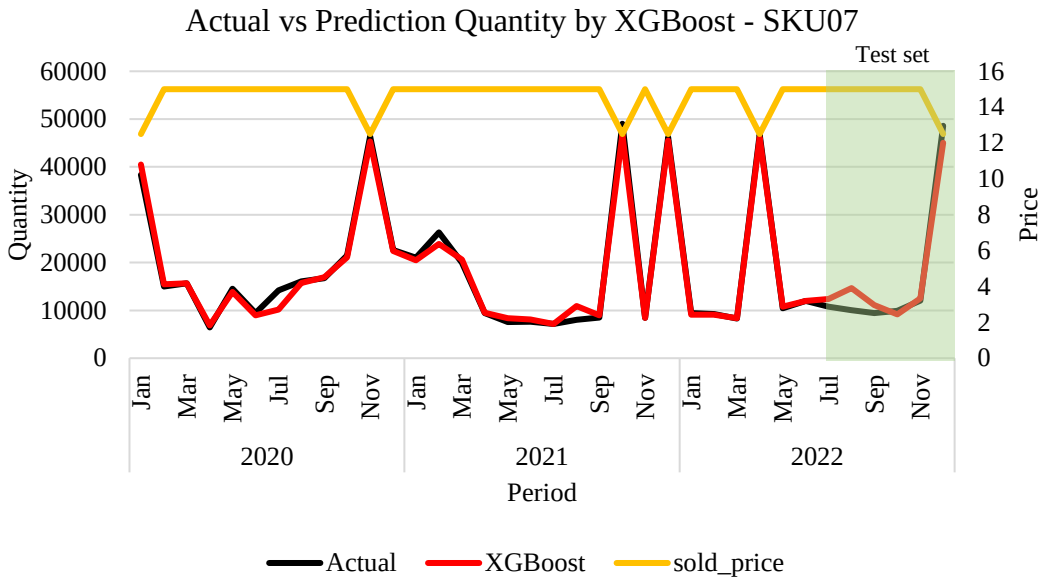


Figure 48 The prediction of SKU07 by XGBoost model

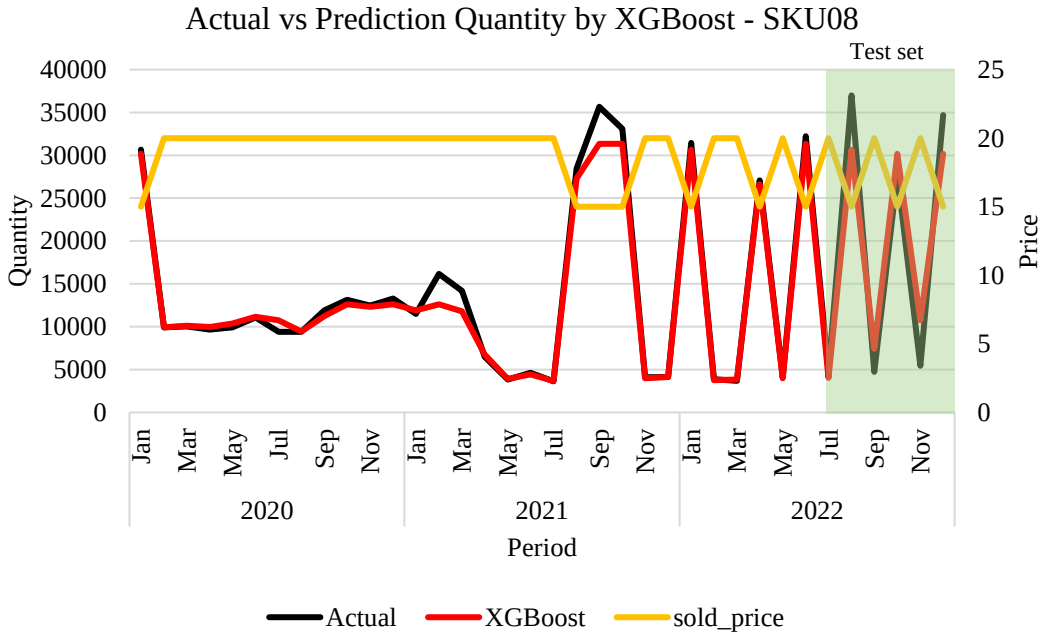


Figure 49 The prediction of SKU08 by XGBoost model

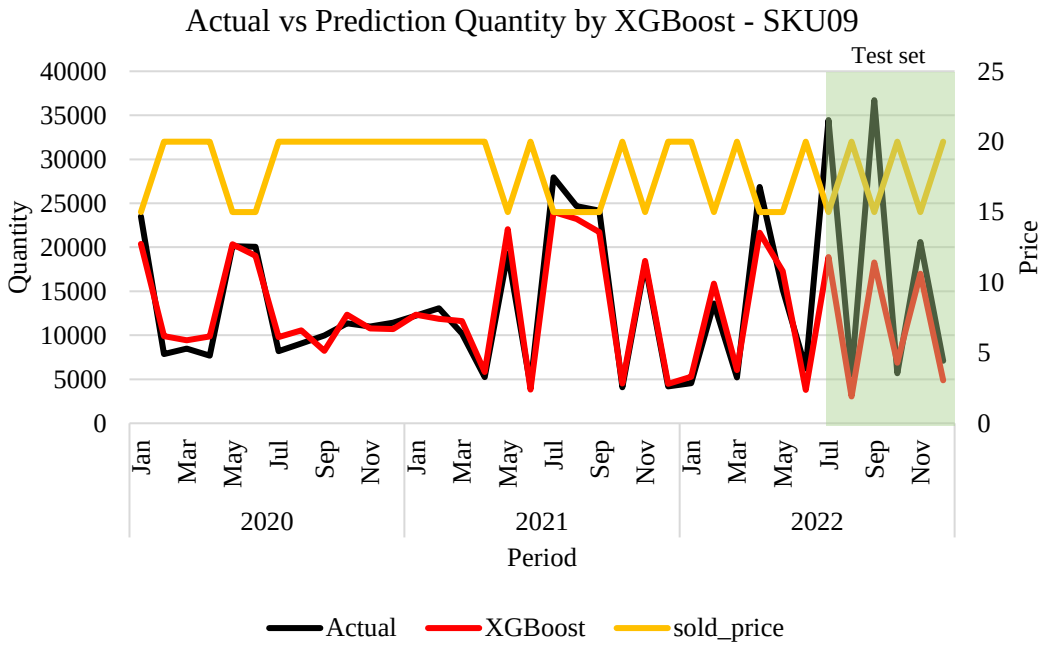


Figure 50 The prediction of SKU09 by XGBoost model

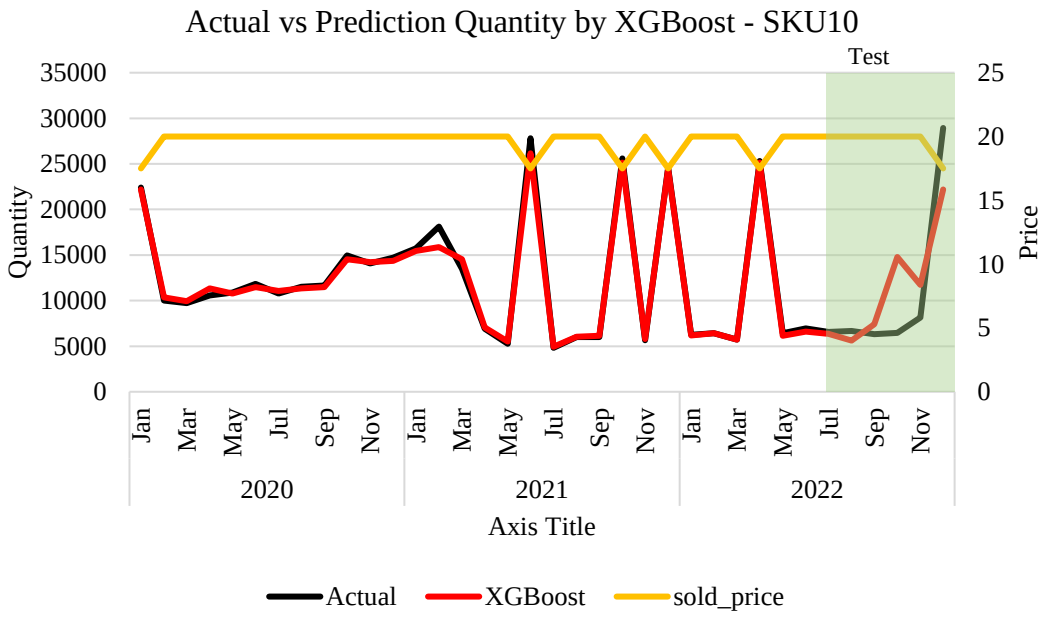


Figure 51 The prediction of SKU10 by XGBoost model

4.3.2.3 ANN result

ANN models were constructed and evaluated for their performance through 10-fold cross-validation to tune hyperparameters. The overall MAPE of the models is weighted by product revenue and summarized as presented in Table 29. The result revealed that using all factors of each product and not considering other products' factors gave the lowest WMAPE on the testing dataset, which was 38.32%. The finding found that using all factors cases had a lower WMAPE than using factors from stepwise regression. Moreover, without considering the factors of other products, the prediction accuracy was higher than with considering the other products' factors. For clustering methods, using K-means techniques provided a lower WMAPE than other methods. However, most of the cases had poor overall prediction accuracy, which may be due to the small training dataset and its easy overfitting.

Table 29 Summarized results of WMAPE with different factors using ANN model

	Factor		WMAPE		
Cases	**Each product**	**Other products**	Train	CV	Test
1	All factors	Not consider other products's factors	37.09%	52.88%	38.32%
2	All factors	Consider factors of other products in the same group by category	61.80%	95.53%	100.58%
3	All factors	Consider factors of other products in the same group by subcategory	57.07%	84.80%	83.19%
4	All factors	Consider factors of other products in the same group by K-means	62.50%	93.63%	65.24%
5	Significant factors	Not consider other products's factors	48.73%	52.64%	44.06%
6	Significant factors	Consider factors of other products in the same group by category	70.20%	97.45%	105.16%
7	Significant factors	Consider factors of other products in the same group by subcategory	71.04%	95.11%	94.90%
8	Significant factors	Consider factors of other products in the same group by K-means	77.49%	101.53%	73.49%

According to Table 29, the overall lowest WMAPE or the best case were analyzed. The selected hyperparameters of all the products and the summarized results are shown in Tables 30 and 31, respectively.

Table 30 Selected hyperparameters of the ANN model of each product

Hyperparameters	**Value or Range**	**Product name**									
		SKU01	SKU02	SKU03	SKU04	SKU05	SKU06	SKU07	SKU08	SKU09	SKU10
batch_size	[4, 8, 16]	4	4	4	4	4	4	4	4	4	4
epoch	100	100	100	100	100	100	100	100	100	100	100
dropout	[0, 0.2, 0.5]	0	0	0.2	0	0	0	0	0	0	0
learning_rate	[0.01, 0.03, 0.1]	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
num_layers	[1, 2]	2	2	2	2	2	2	2	2	2	2
num_units	[20, 50, 100]	100	100	100	100	100	100	100	100	100	100
activation	['relu']	relu	relu	relu	relu	relu	relu	relu	relu	relu	relu

Table 31 Results of the 10 beauty products using ANN model

Product name	Subcategory	Weight	Runtime (sec)	MAPE			MSECV/MSETrain
				Train	CV	Test	
SKU01	Basic skin caere	0.22	3,097.26	36.35%	47.09%	37.32%	1.11
SKU02	Anti-aging	0.11	3,079.68	35.14%	46.19%	35.62%	1.30
SKU03	Anti-aging	0.10	3,019.83	66.50%	148.31%	39.04%	1.76
SKU04	Whitening	0.10	3,020.93	20.56%	32.92%	36.09%	1.54
SKU05	Men	0.10	3,044.36	34.69%	39.92%	31.23%	1.01
SKU06	Whitening	0.09	3,162.89	36.08%	40.63%	50.52%	0.95
SKU07	UV protection	0.07	3,350.61	32.60%	37.59%	39.73%	1.08
SKU08	Anti-aging	0.07	3,219.23	49.85%	57.99%	35.84%	1.24
SKU09	Whitening	0.07	3,254.03	30.64%	37.42%	40.48%	1.32
SKU10	Whitening	0.07	3,714.69	27.05%	36.08%	42.07%	1.37
			WMAPE	37.09%	52.88%	38.32%	

From Table 31, the runtime of the ANN model for each product was around 50 to 60 minutes, which was an average of 53 minutes. Besides, the models may not have overfitting issues since the values of the error proportion of the cross validation and training set of the products were around 1.0 to 1.8. The predictions for the products are presented in Figures 52 – 61.

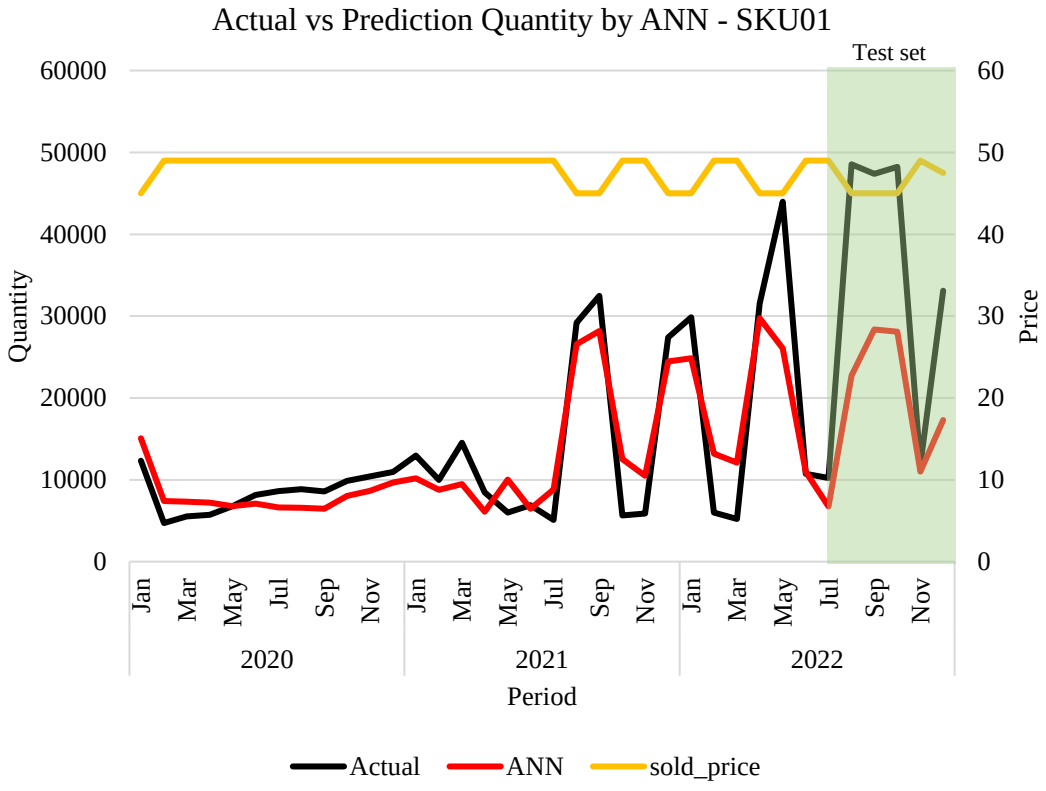


Figure 52 The prediction of SKU01 by ANN model

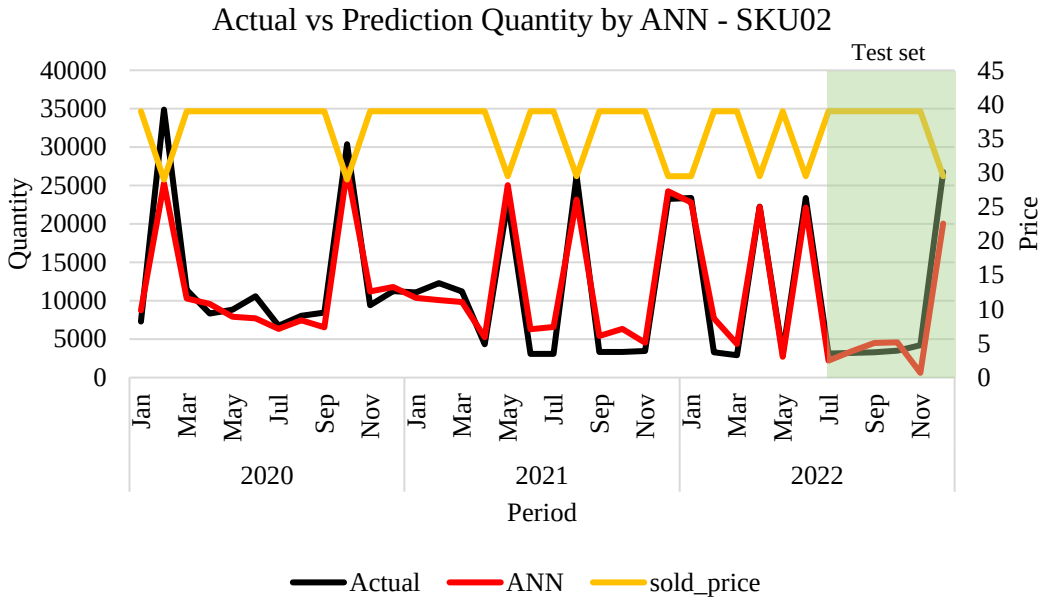


Figure 53 The prediction of SKU02 by ANN model

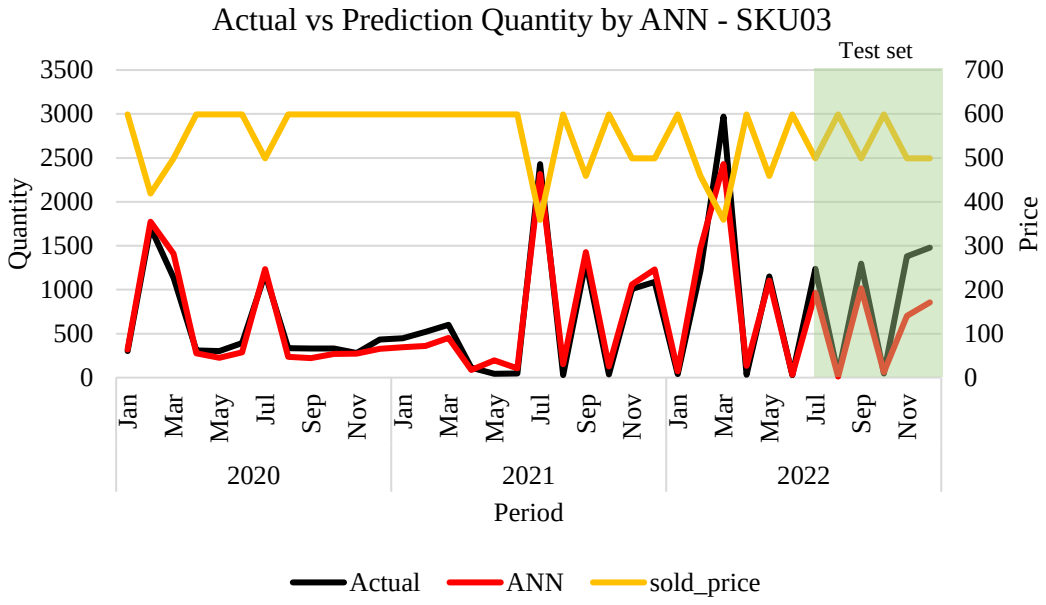


Figure 54 The prediction of SKU03 by ANN model

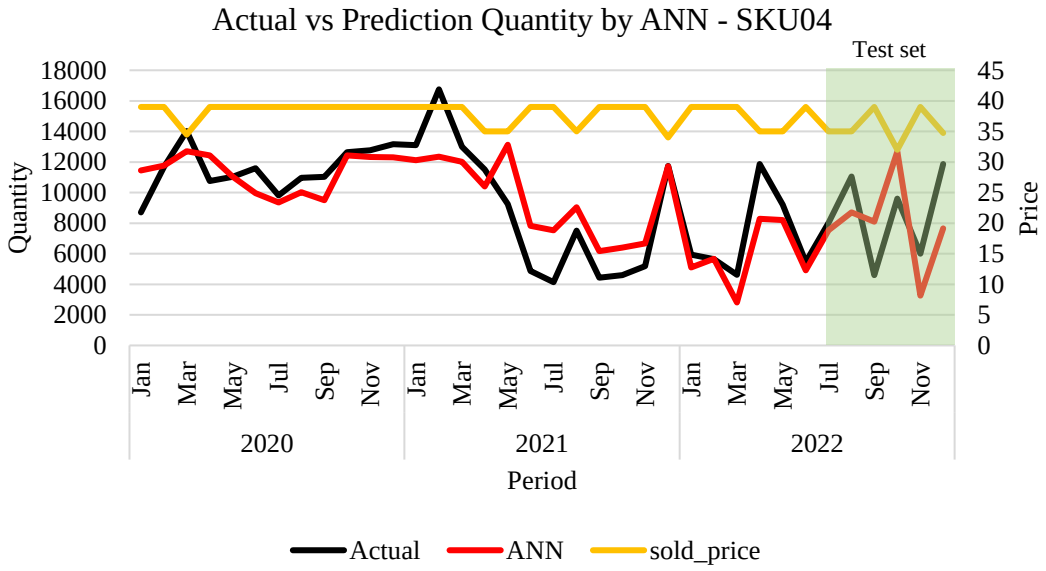


Figure 55 The prediction of SKU04 by ANN model

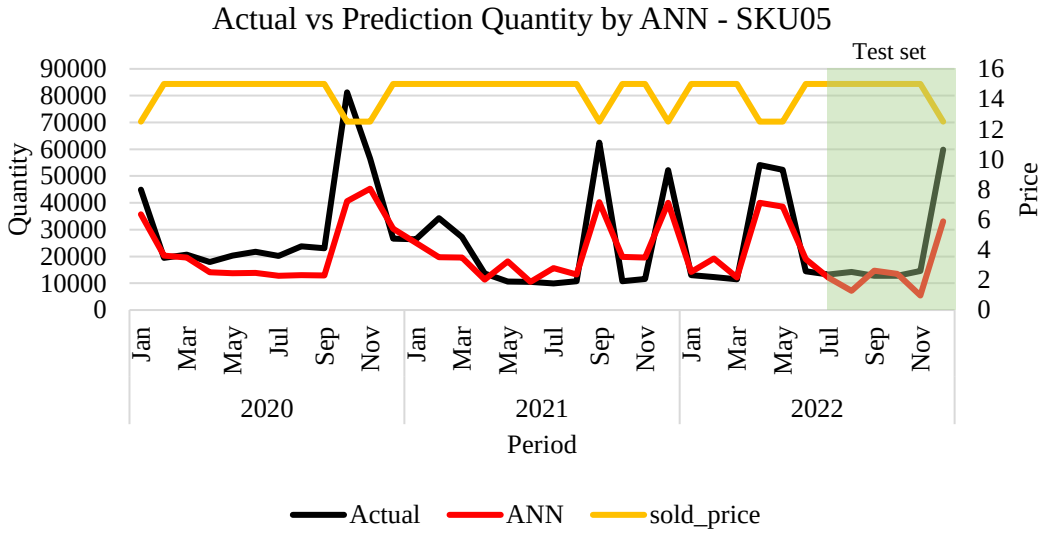


Figure 56 The prediction of SKU05 by ANN model

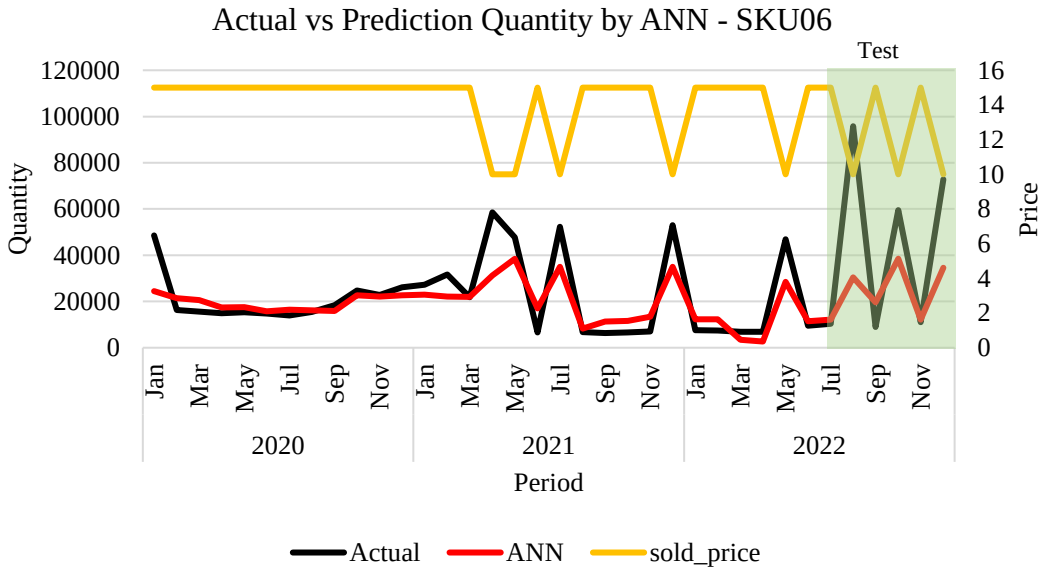


Figure 57 The prediction of SKU06 by ANN model

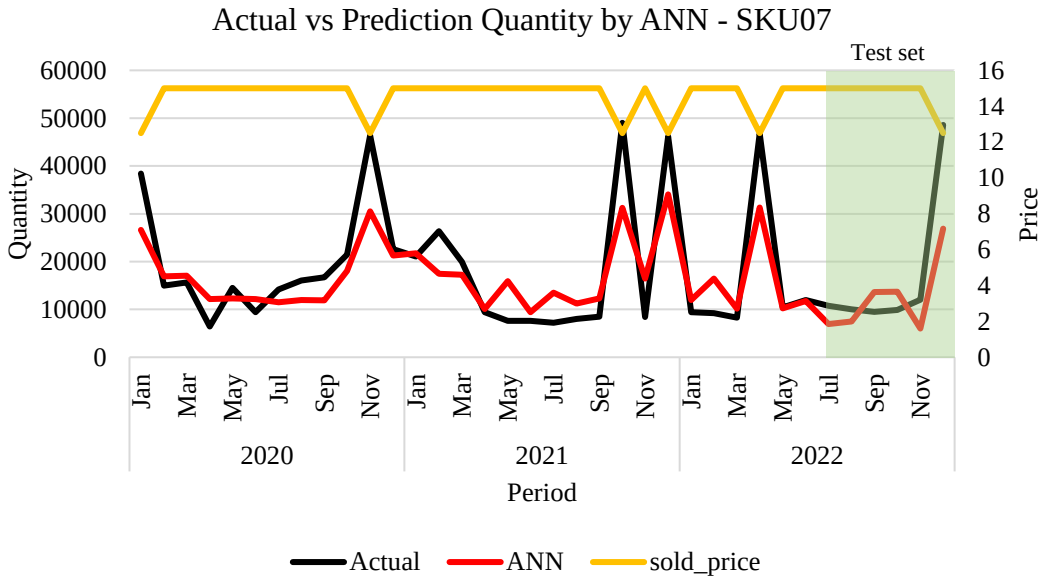


Figure 58 The prediction of SKU07 by ANN model

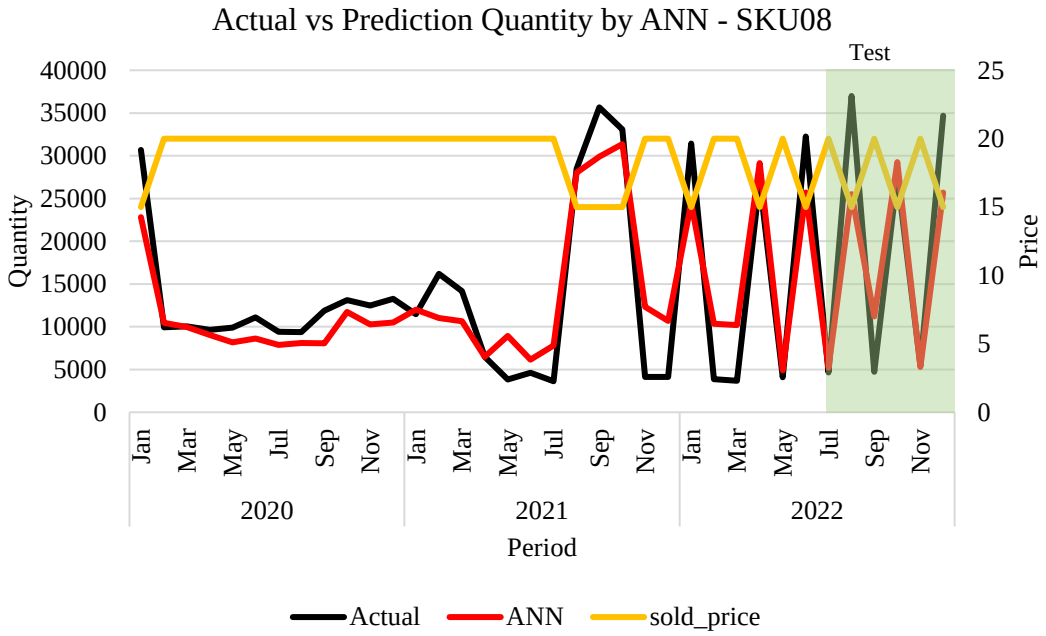


Figure 59 The prediction of SKU08 by ANN model

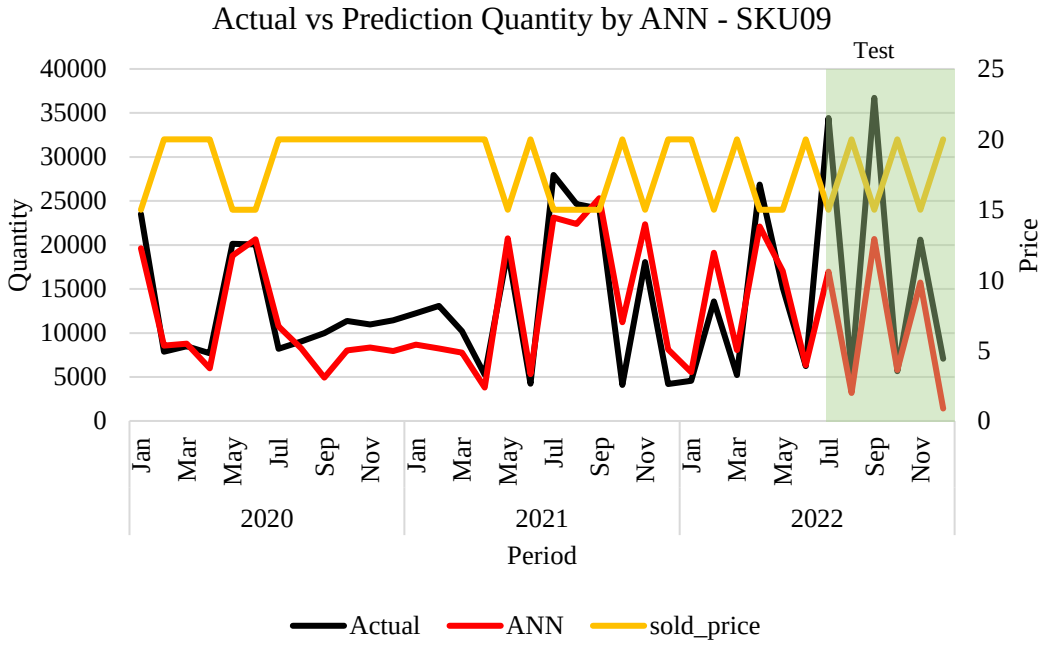


Figure 60 The prediction of SKU09 by ANN model

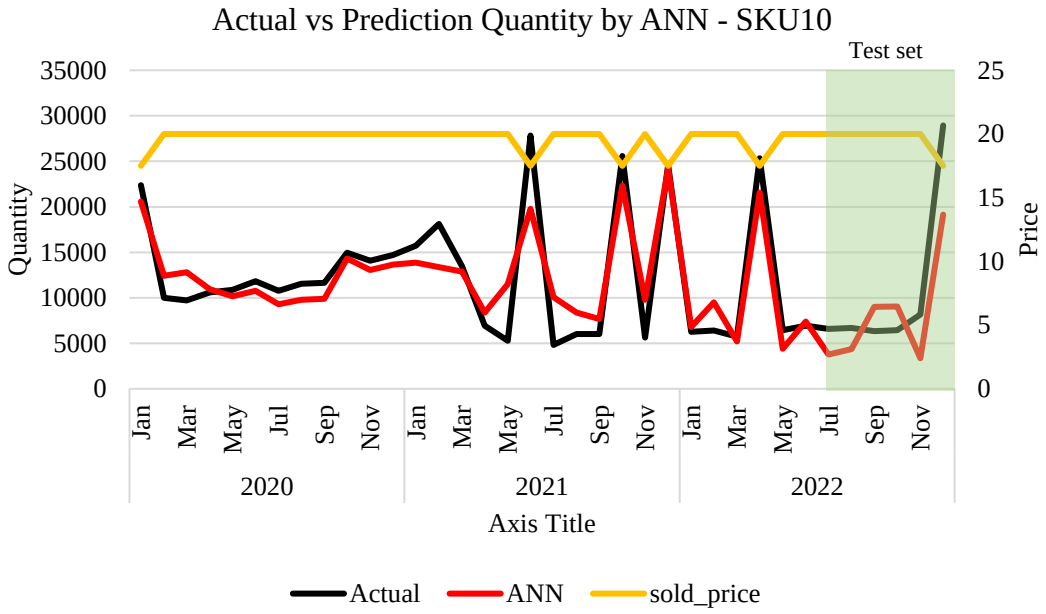


Figure 61 The prediction of SKU10 by ANN model

4.3.3 Hybrid model result

The three machine learning models that had the best performance of each technique were selected and constructed to be hybrid models, ordered by lowest WMAPE, including: (i) a random forest model using significant factors from the stepwise method and not considering factors of other products with WMAPE of 28.15% (Case5 in Table 23); (ii) an XGBoost model using significant factors from the stepwise method and considering factors of other products in the same subcategory with WMAPE of 32.60% (Case7 in Table 26); and (iii) an ANN model using all factors and not considering factors of other products with WMAPE of 38.32% (Case1 in Table 29).

4.3.3.1 Parallel hybrid model

The predictions of the two out of three models are selected and combined by passing them as input or independent variables in linear regression. A total of three combinations of models were constructed and evaluated, as shown in Table 32.

Table 32 WMAPE Results of the parallel hybrid models

Model		WMAPE	
Model 1	**Model 2**	**Train**	**Test**
Random forest	XGBoost	8.02%	35.16%
Random forest	ANN	16.55%	31.56%
XGBoost	ANN	7.29%	32.81%

From the result, the hybrid model of random forest and ANN had the lowest WMAPE. The result of all products is shown in Table 33 and the predictions for the products are presented in Figures 62 – 71.

Table 33 Results of the 10 beauty products using random forest and ANN model

Product name	Subcategory	Weight	MAPE	
			Train	**Test**
SKU01	Basic skin caere	0.22	21.74%	33.46%
SKU02	Anti-aging	0.11	10.31%	17.67%
SKU03	Anti-aging	0.10	41.52%	39.48%
SKU04	Whitening	0.10	10.88%	23.69%
SKU05	Men	0.10	12.72%	24.50%
SKU06	Whitening	0.09	7.88%	53.46%
SKU07	UV protection	0.07	25.57%	38.72%
SKU08	Anti-aging	0.07	7.43%	26.38%
SKU09	Whitening	0.07	9.93%	22.77%
SKU10	Whitening	0.07	5.68%	36.67%
		WMAPE	16.55%	31.56%

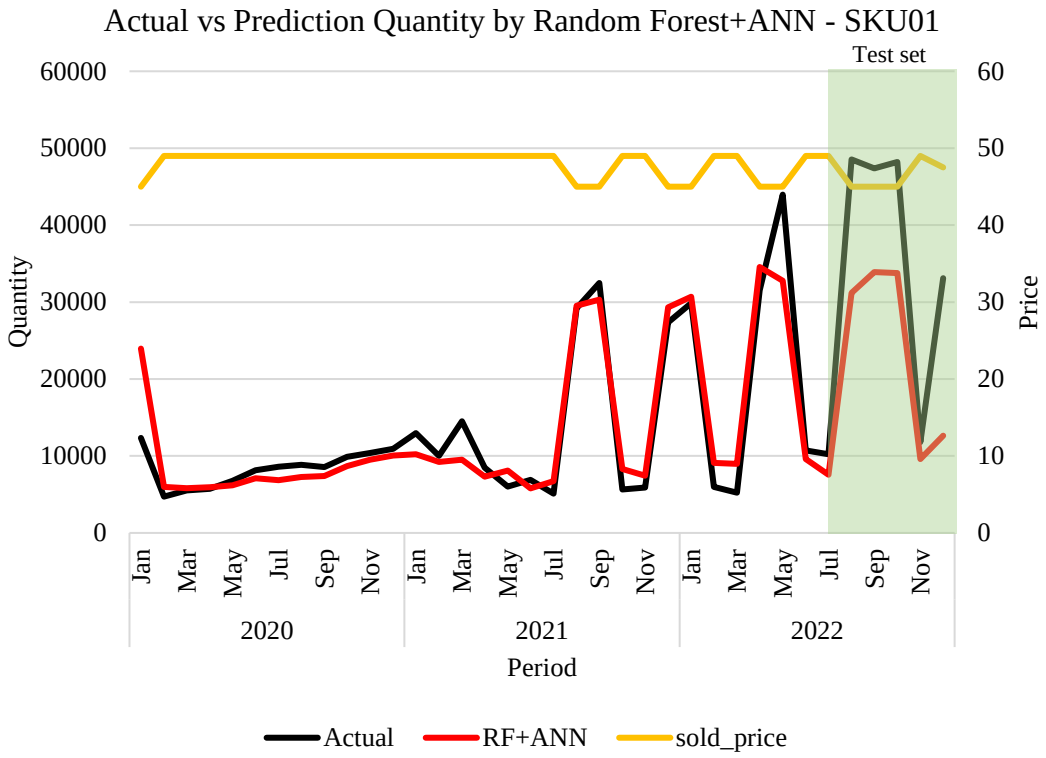


Figure 62 The prediction of SKU01 by parallel hybrid model

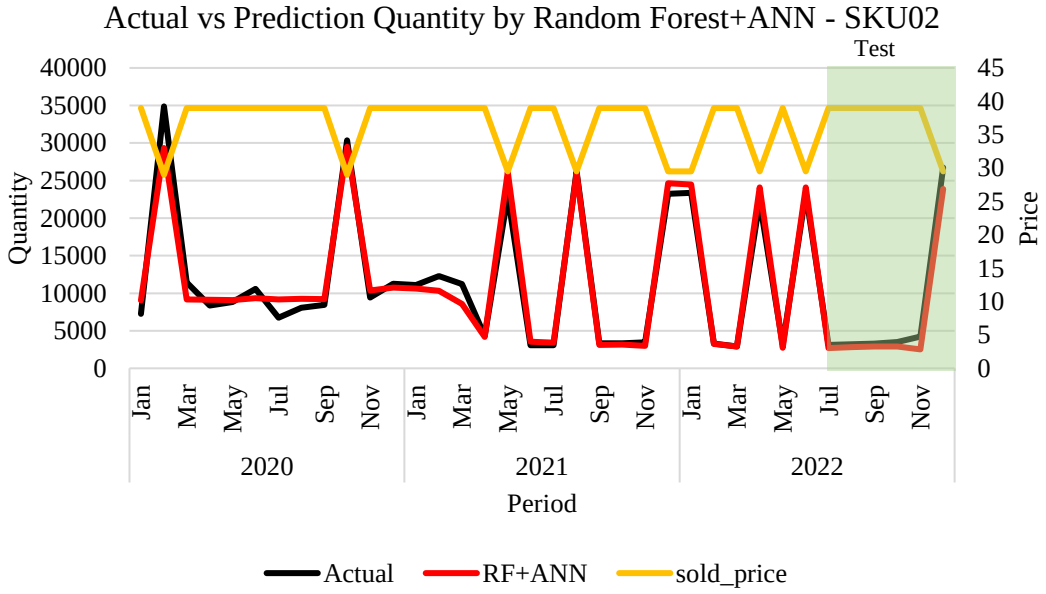


Figure 63 The prediction of SKU02 by parallel hybrid model

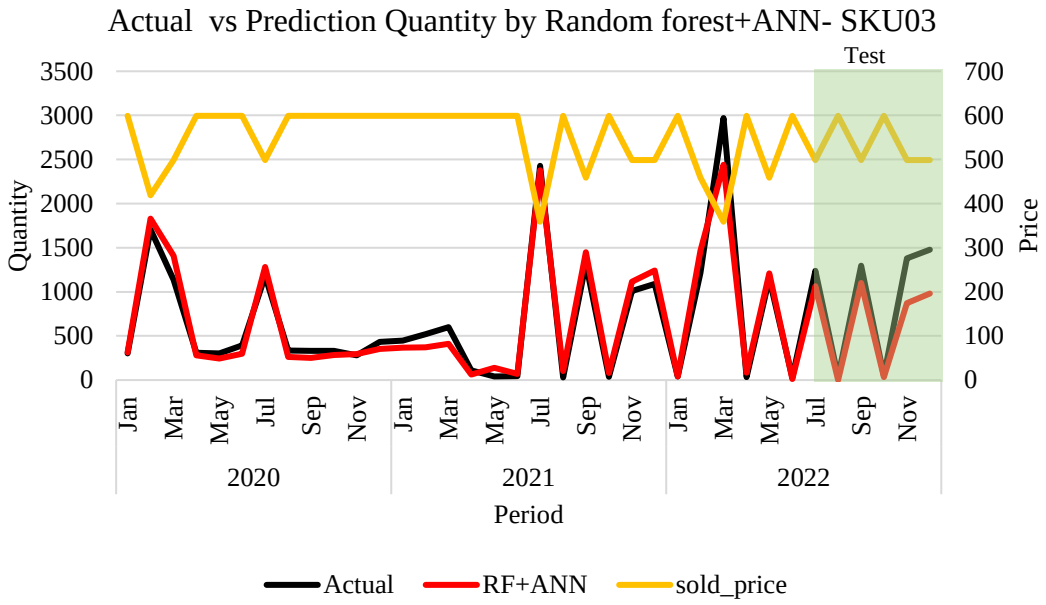


Figure 64 The prediction of SKU03 by parallel hybrid model

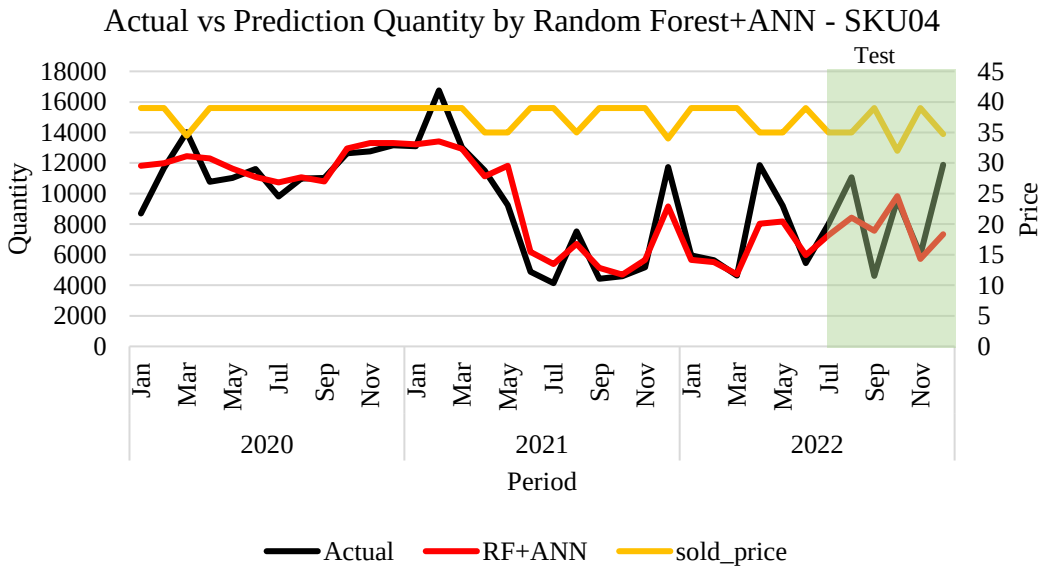


Figure 65 The prediction of SKU04 by parallel hybrid model

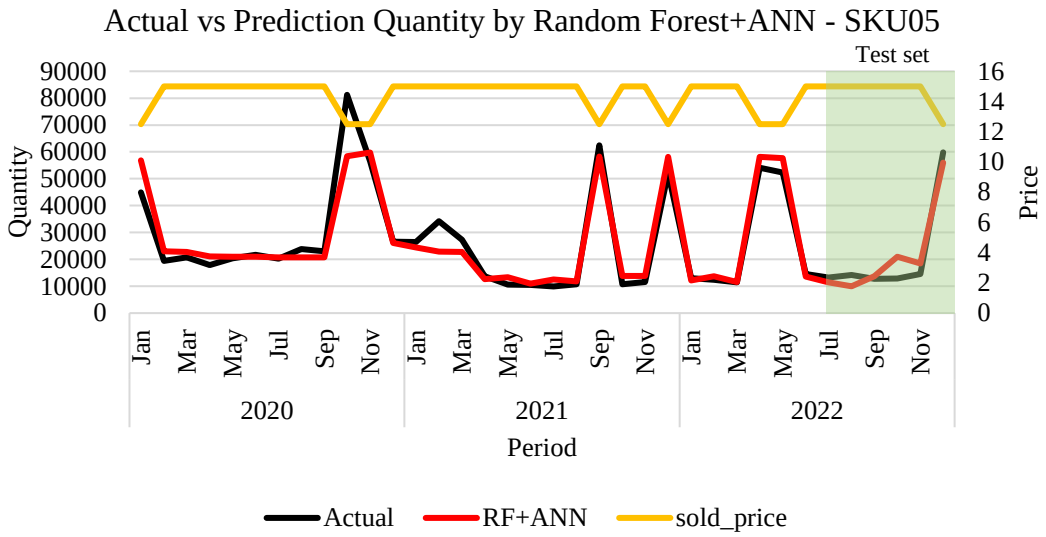


Figure 66 The prediction of SKU05 by parallel hybrid model

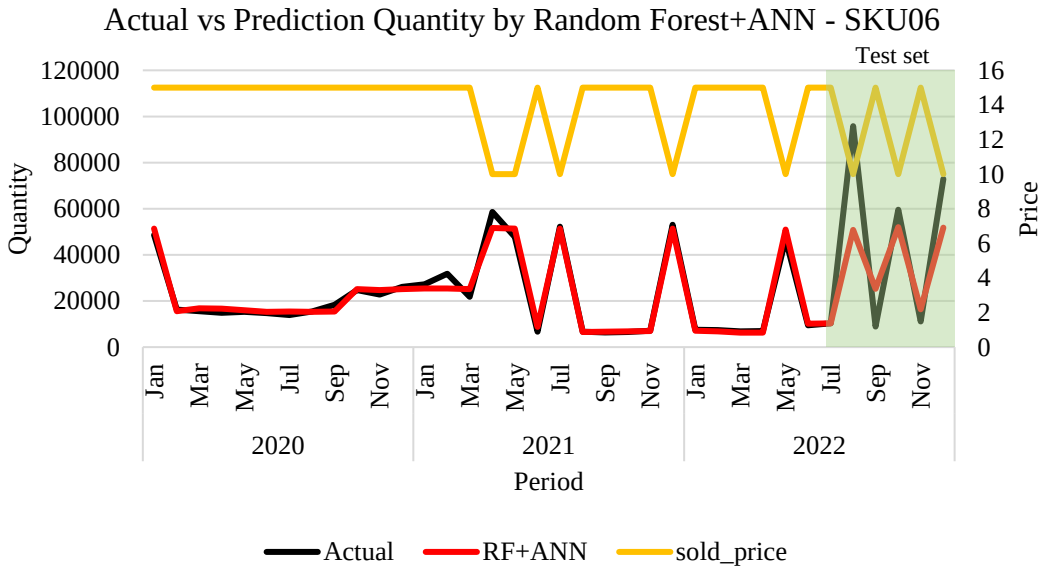


Figure 67 The prediction of SKU06 by parallel hybrid model

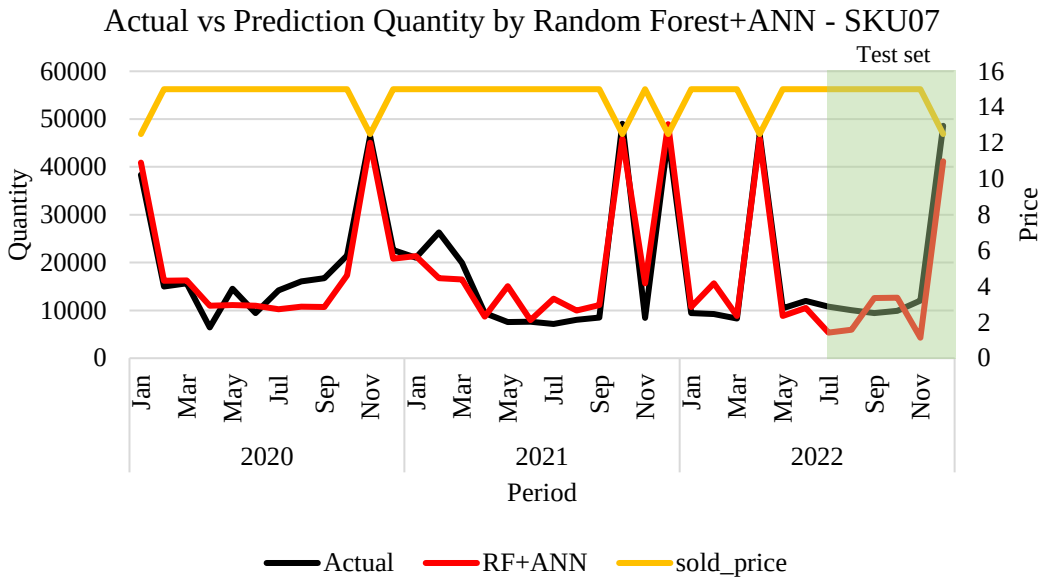


Figure 68 The prediction of SKU07 by parallel hybrid model

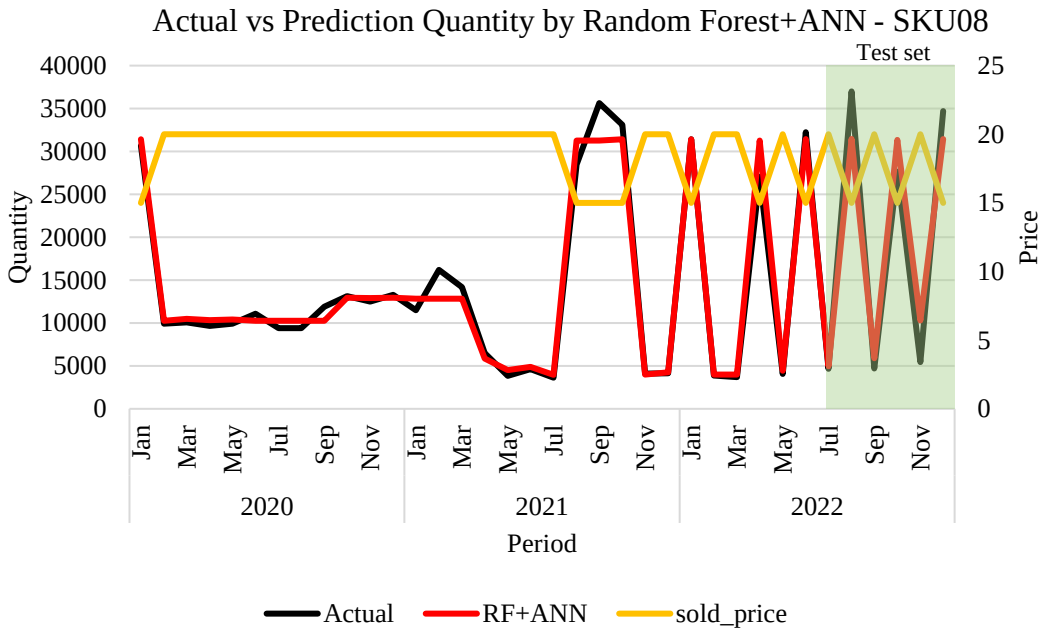


Figure 69 The prediction of SKU08 by parallel hybrid model

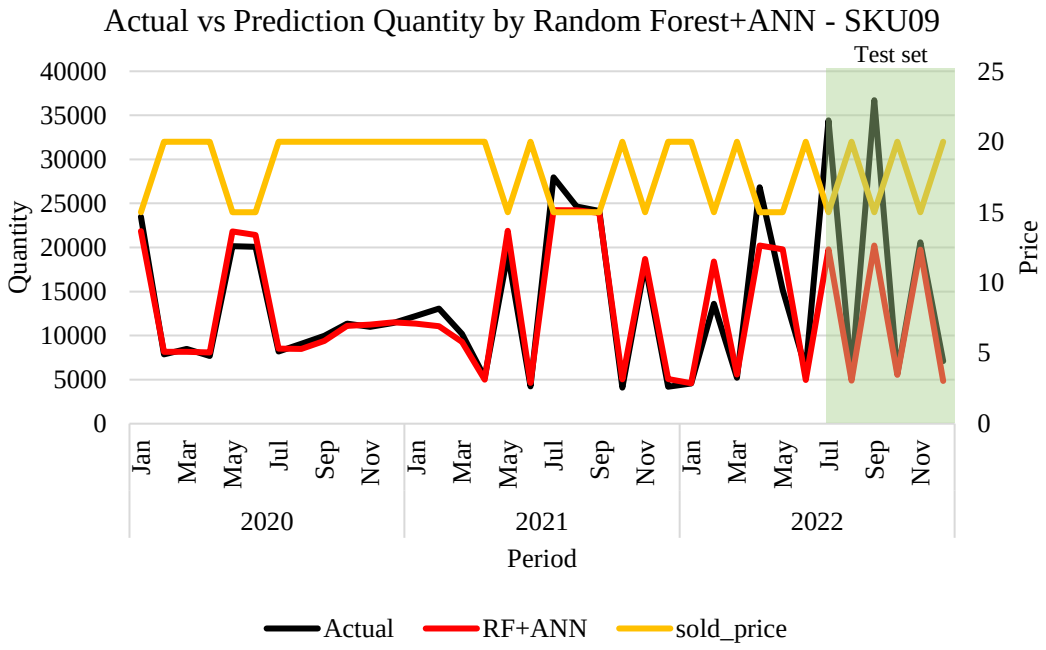


Figure 70 The prediction of SKU09 by parallel hybrid model

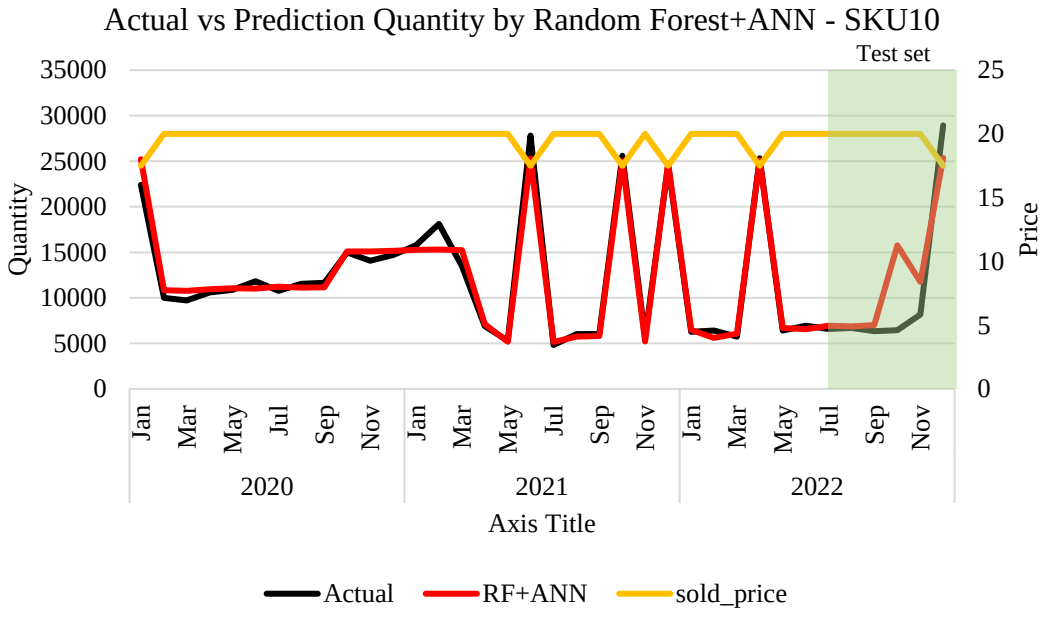


Figure 71 The prediction of SKU10 by parallel hybrid model

4.3.3.2 Series hybrid model

After selecting two out of three models, a total of six combinations of models were constructed and evaluated by WMAPE, as shown in Table 34. To construct the series hybrid models, the first model was fitted, predicted the output and calculated residuals. Then, the second model was trained by using the factors depended on the best case of the second model, and the residuals from the first model were passed as dependent variable. The hyperparameters setting of the second model used to tune were the same values or ranges as the single model, except for the activation function of the ANN hyperparameter at the hidden layer(s) and output layer, which were hyperbolic tangent ('tanh') and linear function ('linear'), respectively, to be able to support both positive and negative output. After applying the selected hyperparameters to the model, the predictions of the second model were calculated. The final prediction was calculated by summing the first predictions and the second predictions, and then the performances were evaluated and compared.

Table 34 WMAPE Results of the series hybrid models

Model		**Runtime (sec)**	**WMAPE**		
Model 1	**Model 2**		**Train**	**CV**	**Test**
Random forest	XGBoost	2058.43	10.01%	15.48%	27.65%
Random forest	ANN	3552.95	13.43%	13.64%	28.04%
XGBoost	ANN	3313.73	7.06%	8.00%	32.61%
XGBoost	Random forest	121.31	7.64%	8.61%	33.45%
ANN	Random forest	161.74	21.68%	27.91%	39.35%
ANN	XGBoost	2022.29	9.15%	28.50%	40.35%

According to Table 34, it found that the hybrid model of random forest and XGBoost was outperformed other series hybrid models as its lowest WMAPE of 27.65%. The selected hyperparameters of all the products and the summarized results are shown in Tables 35 and 36, respectively.

Table 35 Selected hyperparameters of the random forest and XGBoost model of each product

Hyperparameters	Value or Range	Product name									
		SKU01	SKU02	SKU03	SKU04	SKU05	SKU06	SKU07	SKU08	SKU09	SKU10
n_estimators	[10, 15, 20]	20	15	15	10	10	10	10	10	15	10
max_depth	[3, 5]	5	3	5	3	5	5	3	3	3	3
min_child_weight	[1, 2, 3, 5]	2	1	1	5	3	5	2	2	5	3
eta	[0.01, 0.03, 0.1, 0.3]	0.03	0.3	0.3	0.01	0.01	0.1	0.1	0.01	0.3	0.3
subsample	[0.5, 0.7, 1.0]	1	0.7	0.7	1	0.7	1	1	1	0.5	0.7
gamma	[0, 0.5, 1]	0	0	1	0	0	0	0	0	0	0
reg_lambda	[1, 2, 3, 5]	3	2	2	1	3	1	2	1	1	5

Table 36 Results of the 10 beauty products using random forest and XGBoost model

Product name	Subcategory	Weight	Runtime (sec)	MAPE		
				Train	CV	Test
SKU01	Basic skin caere	0.22	1924.98	11.31%	13.62%	32.79%
SKU02	Anti-aging	0.11	2015.59	4.09%	13.34%	11.97%
SKU03	Anti-aging	0.10	2025.41	8.36%	37.67%	13.26%
SKU04	Whitening	0.10	2112.09	15.01%	15.39%	31.94%
SKU05	Men	0.10	1876.33	12.56%	12.90%	29.57%
SKU06	Whitening	0.09	2100.23	6.35%	8.65%	50.91%
SKU07	UV protection	0.07	2133.48	20.09%	26.03%	12.42%
SKU08	Anti-aging	0.07	2199.13	7.61%	7.86%	28.46%
SKU09	Whitening	0.07	2091.78	9.24%	12.41%	25.23%
SKU10	Whitening	0.07	2105.24	3.39%	5.61%	35.03%
			WMAPE	10.01%	15.48%	27.65%

From Table 36, the runtime of the hybrid model for each product was around 31 to 36 minutes, which was an average of 34 minutes. The predictions for the products are presented in Figures 72 – 81.

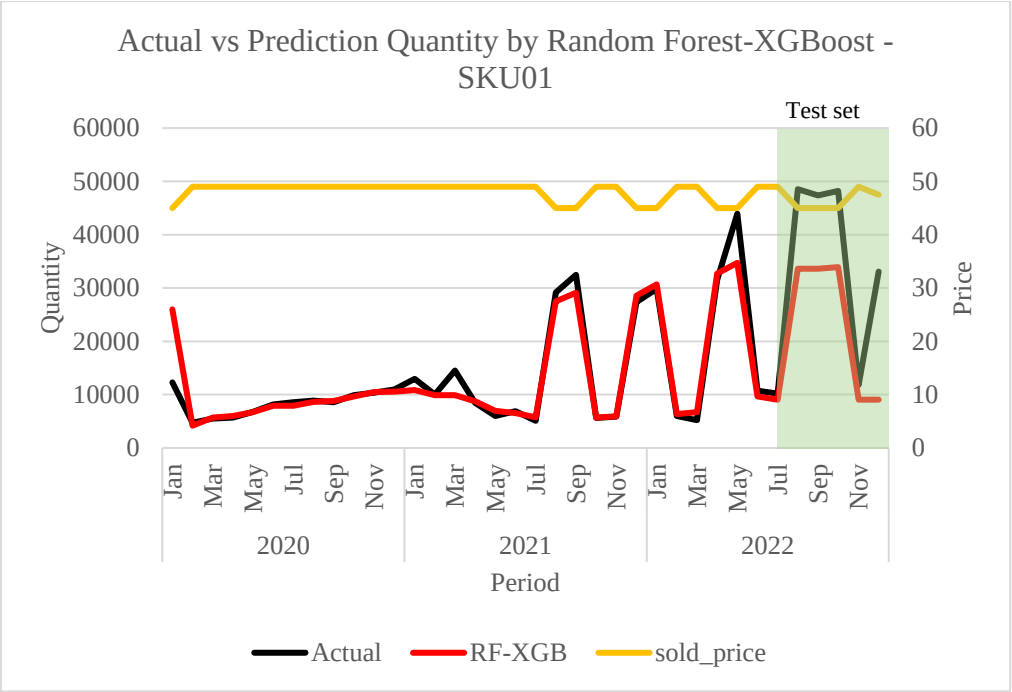


Figure 72 The prediction of SKU01 by series hybrid model

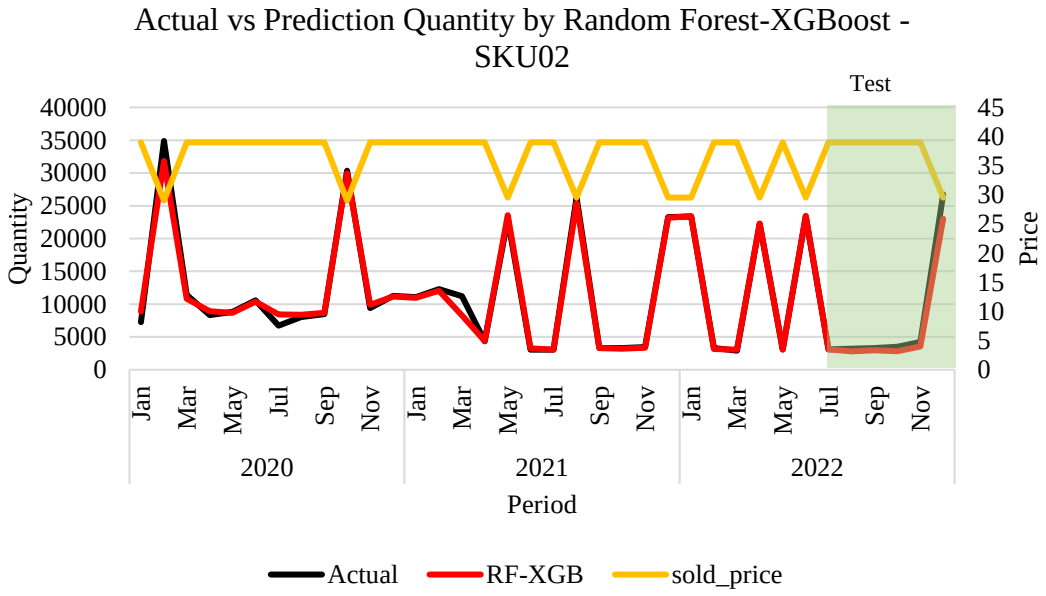


Figure 73 The prediction of SKU02 by series hybrid model

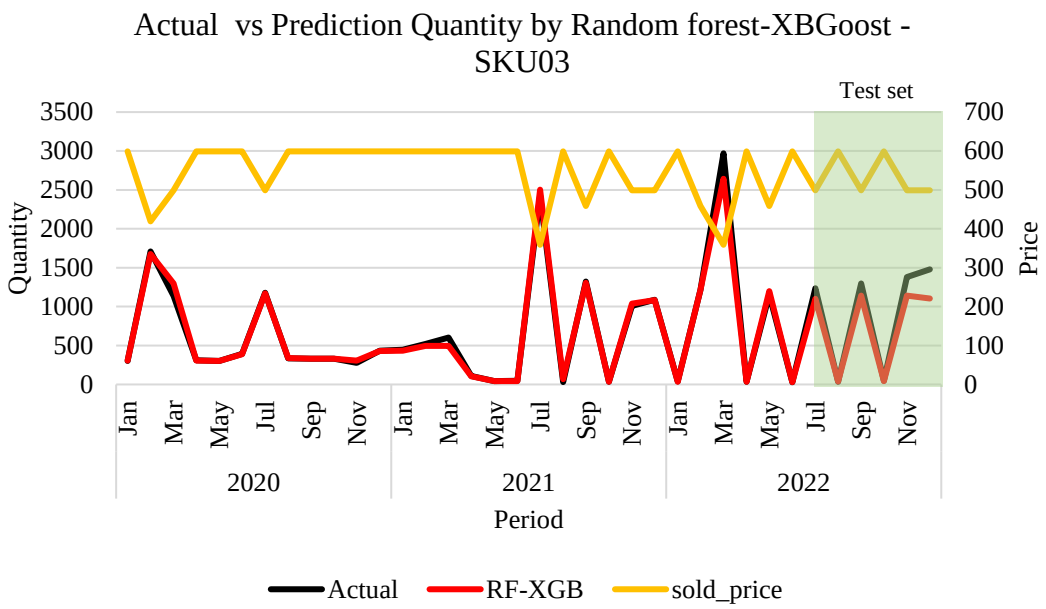


Figure 74 The prediction of SKU03 by series hybrid model

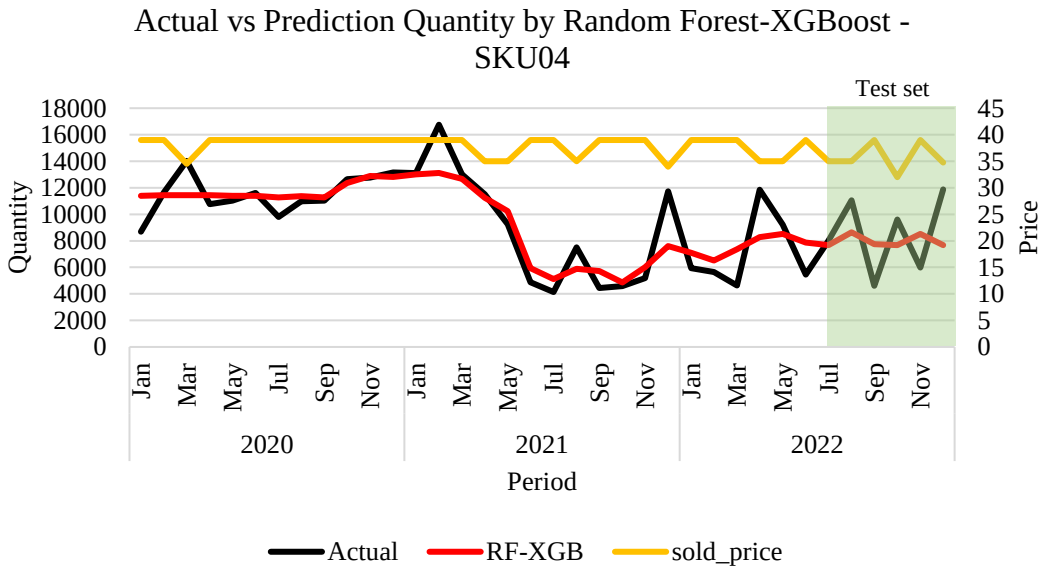


Figure 75 The prediction of SKU04 by series hybrid model

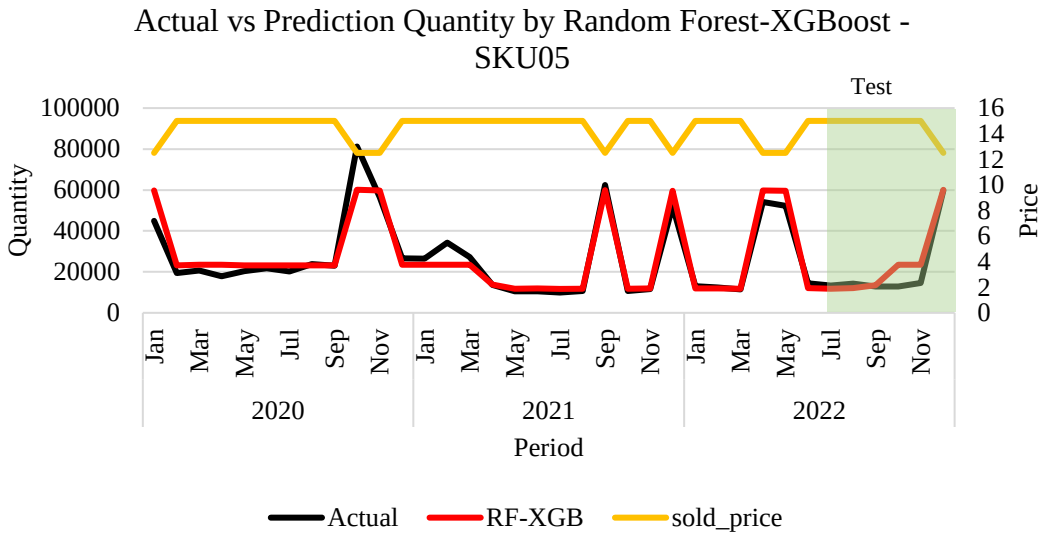


Figure 76 The prediction of SKU05 by series hybrid model

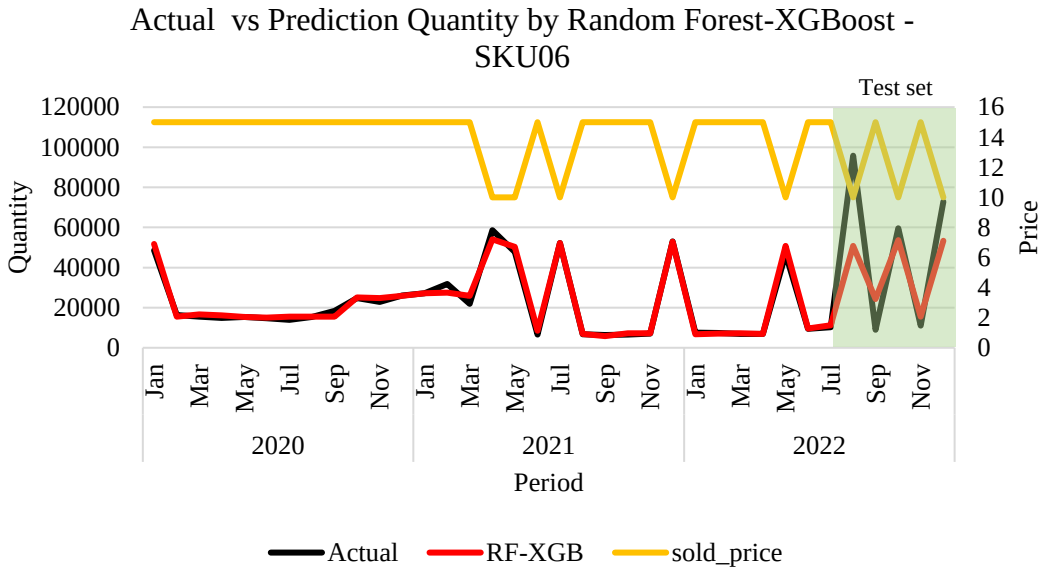


Figure 77 The prediction of SKU06 by series hybrid model

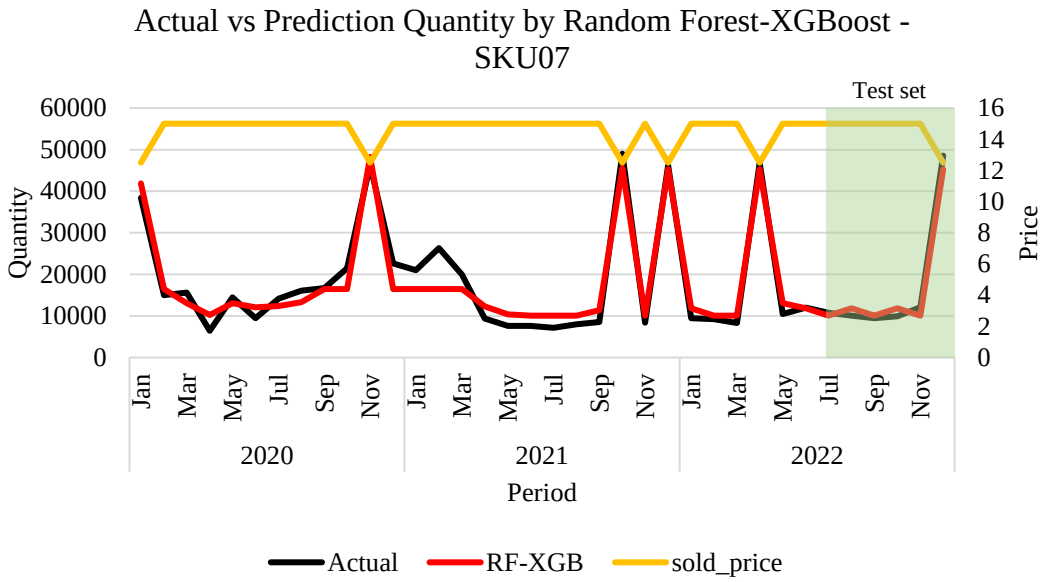


Figure 78 The prediction of SKU07 by series hybrid model

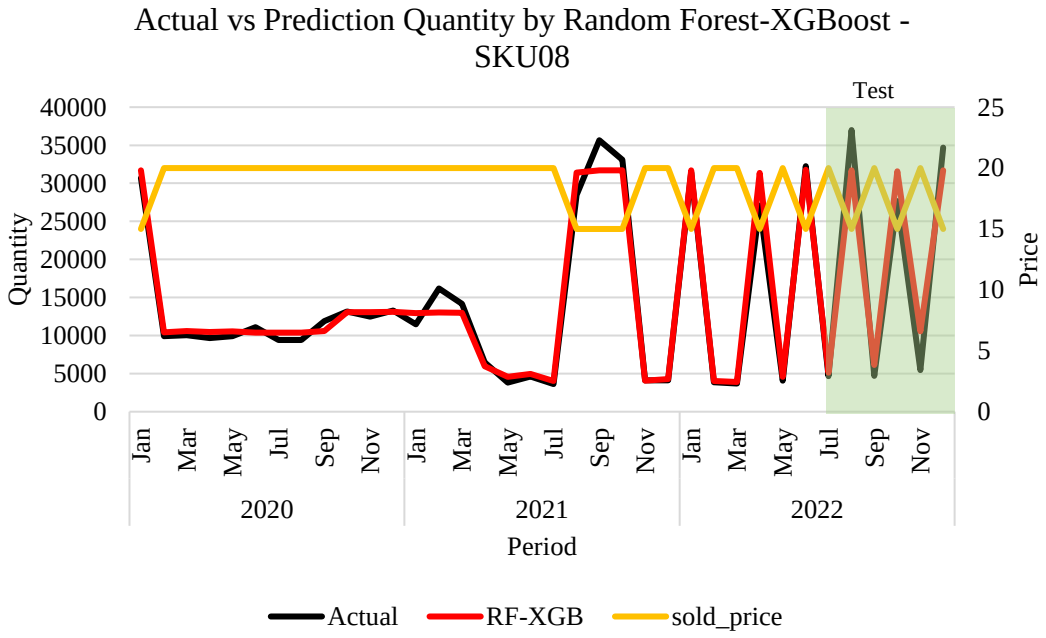


Figure 79 The prediction of SKU08 by series hybrid model

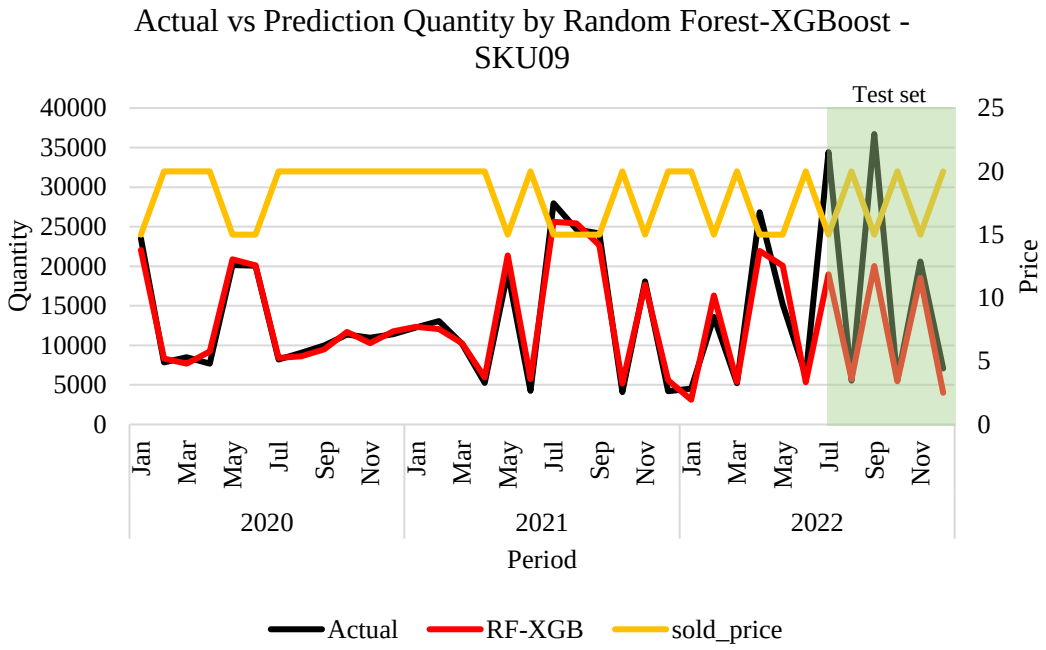


Figure 80 The prediction of SKU09 by series hybrid model

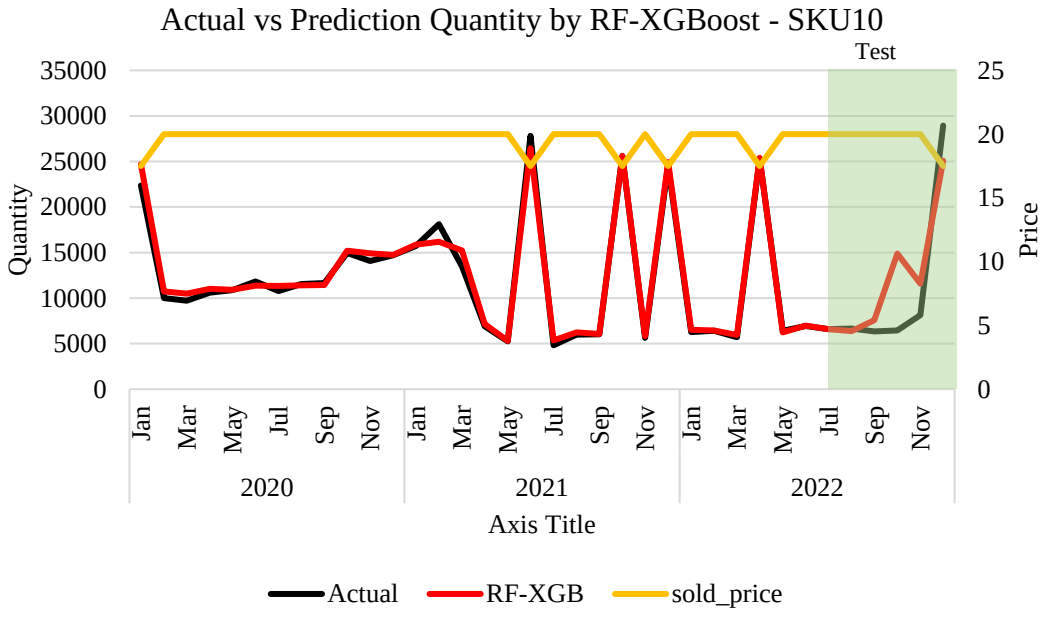


Figure 81 The prediction of SKU10 by series hybrid model

4.4 Result comparison

4.4.1 Overall performance

4.4.1.1 Model evaluation and selection

The independent variables or factors, including selling price, discount percentage, promotion period, lagged promotion period, monthly period, number of active stores, subsidies and welfare programs, and number of COVID-19 new cases, were used to fit the models and study their effects on the sales quantity of ten beauty products in facial moisturizer categories. The models studied in this work were linear regression, random forest, XGBoost, ANN, hybrid models with parallel and series structure. After tuning model hyperparameters by grid search and 10-fold cross validation method and fitting the models, the model performance was measured by MAPE. The weighted MAPE (WMAPE) was weighted by revenue of the 10 beauty products to compare overall performance of each model. The WMAPE on the testing dataset was then used to choose the suitable model for the products. Table 37 demonstrates the summary of the overall WMAPE of the models.

Table 37 WMAPE of the prediction models

	Model	Runtime (sec)	WMAPE Test set
	Linear Regression	-	43.92%
	Random forest	404.61	28.15%
	XGBoost	2,935.41	32.60%
	ANN	3,196.35	38.32%
Hybrid Parallel	Random forest+ANN	0.02	31.56%
Hybrid Series	Random forest-XGBoost	2,058.43	27.65%

Table 37 shows that all of the machine learning models and hybrid models had a lower WMAPE than the linear regression method. This means that using machine learning methods can make predictions more accurate than the traditional model.

The machine learning models with the highest overall prediction accuracy or lowest WMAPE were the random forest model using factors from stepwise regression with a WMAPE of 28.15%, followed by the XGBoost model using factors from stepwise regression and taking into account factors of other products in the same subcategory, and the ANN model using all factors with a WMAPE of 32.60% and 38.32%, respectively. Unlike the random forest model, XGBoost and ANN used about 6 times and 8 times longer running times to train the model compared to the random forest model, respectively, because XGBoost and ANN may have more complex architectures and require more iterations for convergence, and there were more hyperparameters to tune. Moreover, the ANN model had a low overall prediction accuracy, which may be due to the small dataset used to train the model.

For the hybrid models, the series structure of random forest and XGBoost models outperformed other models on the testing dataset, getting 27.65% of the WMAPE. The hybrid model can improve prediction performance. However, compared to the best single machine learning model, the random forest model, there is a slight difference in WMAPE, which is 0.5% better than single models. The model also uses more running times, which is five times the random forest use, or around 30 minutes. According to Singh et al. (2019) and Saha et al. (2022) studies about forecasting sales with promotions or events, the results found that the performance of their model measured by WMAPE was in the range of 10% to 38%, and in this work, the WMAPE of the random forest model is within that range. Therefore, it can be concluded that the random forest model is the most suitable model for this dataset to predict the sales quantity of beauty products.

4.4.1.2 Factor analysis

To interpret the important features, the result from the stepwise regression as shown in Table 21 and SHAP value are used. From the selected model, the random forest model using significant factors from the stepwise

regression, SHAP value was calculated to interpret the impact of factors on the predictions. In the SHAP value's plot, the x-axis is the value of the dependent variable, and the y-axis is factors or features, which are shown in the order of importance, with the first one being the most important and the last being the least important one. The dot color of red and blue represents the high value and low value, respectively. As displayed in Figure 82-91, SHAP values of the 10 products were summarized in Table 38.

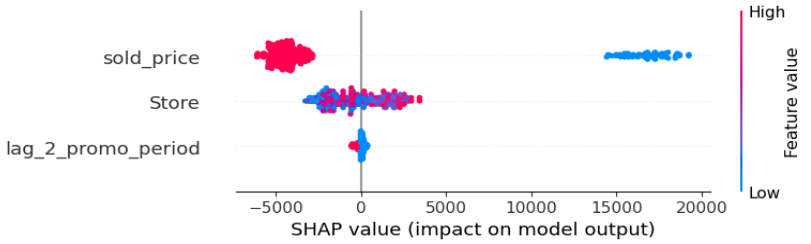


Figure 82 SHAP value of SKU01 using random forest model

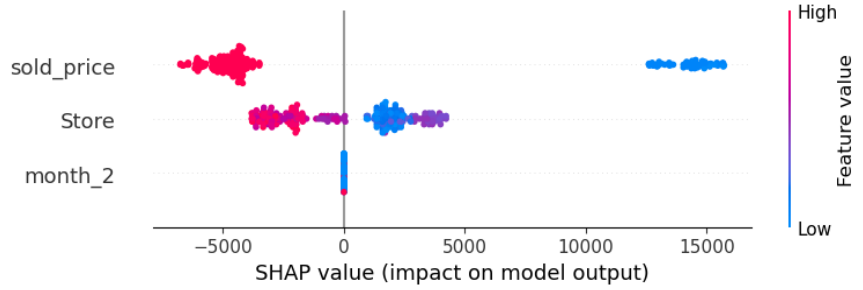


Figure 83 SHAP value of SKU02 using random forest model

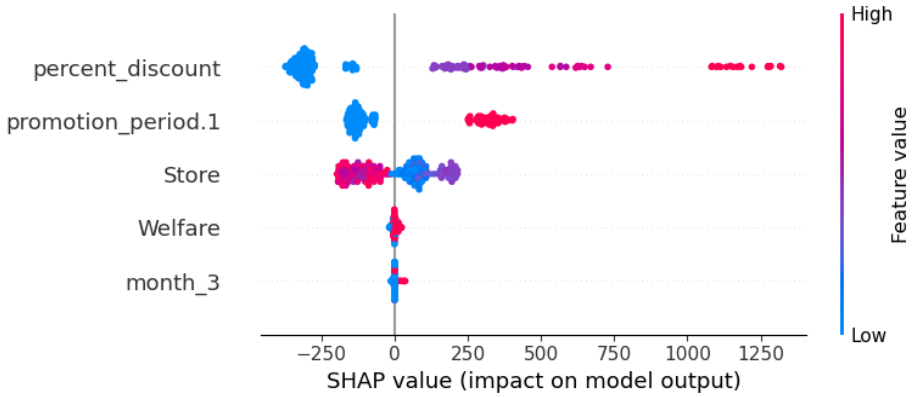


Figure 84 SHAP value of SKU03 using random forest model

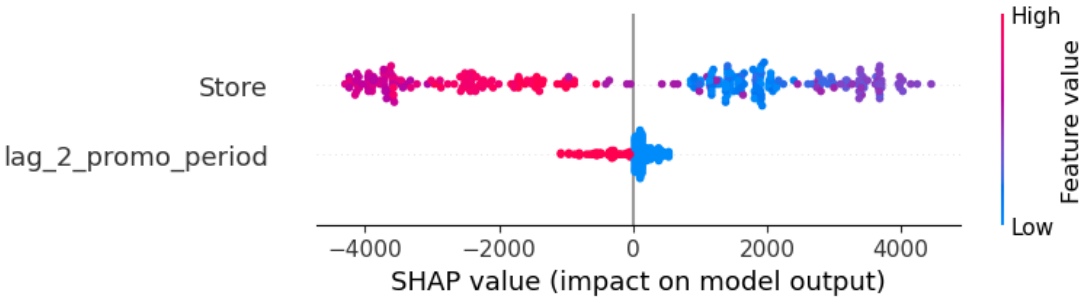


Figure 85 SHAP value of SKU04 using random forest model

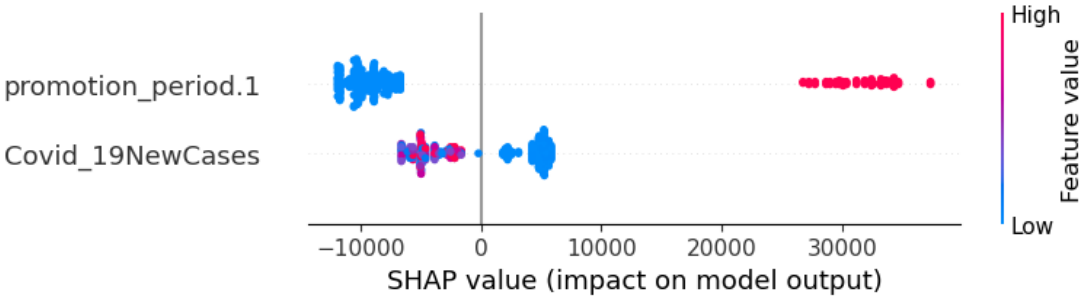


Figure 86 SHAP value of SKU05 using random forest model

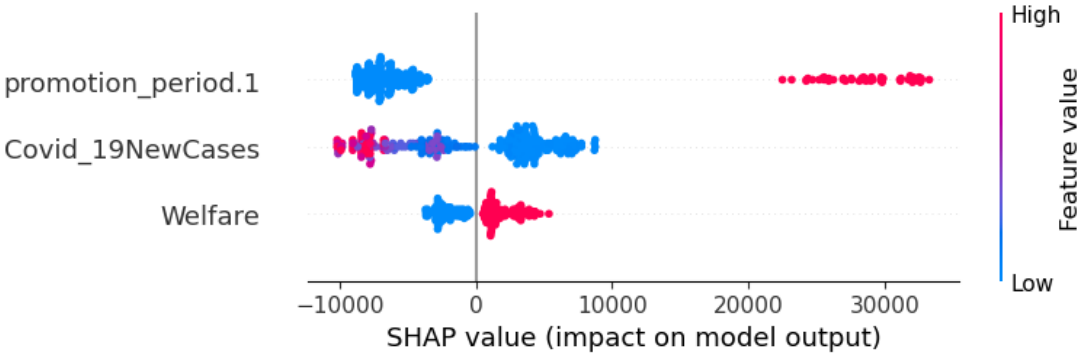


Figure 87 SHAP value of SKU06 using random forest model

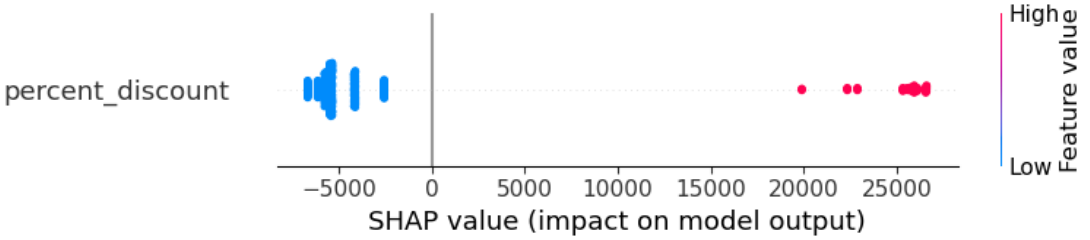


Figure 88 SHAP value of SKU07 using random forest model

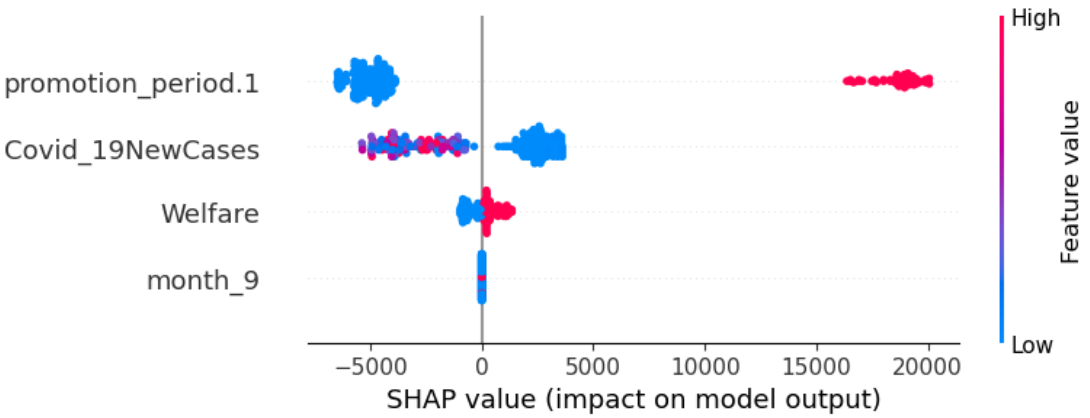


Figure 89 SHAP value of SKU08 using random forest model

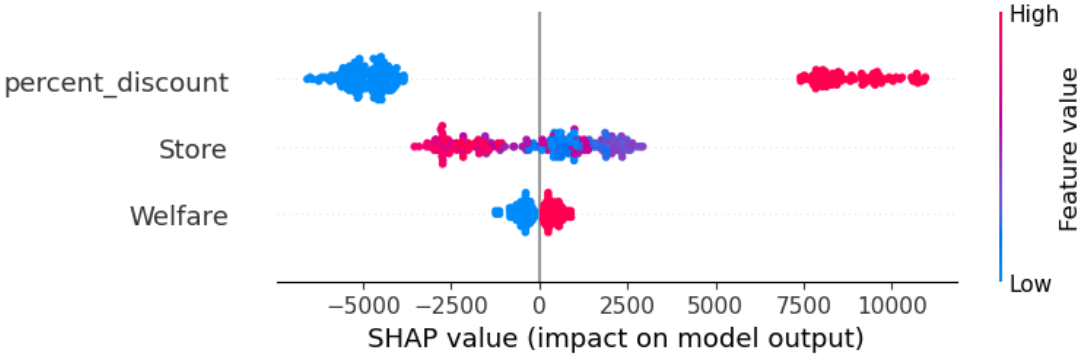


Figure 90 SHAP value of SKU09 using random forest model

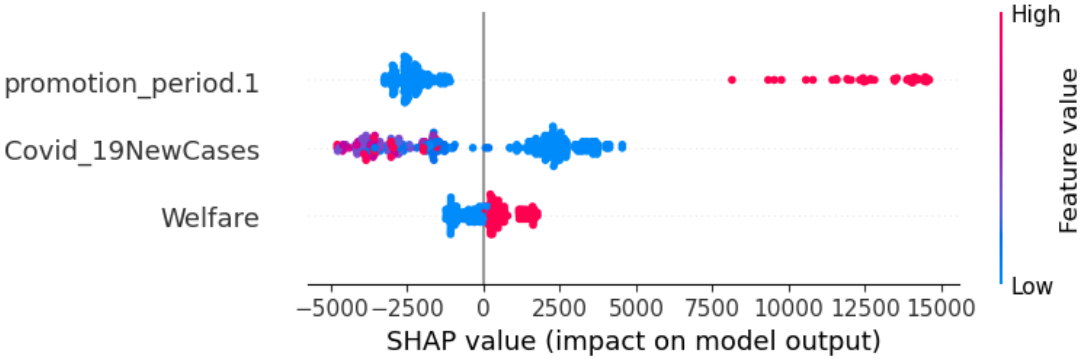


Figure 91 SHAP value of SKU10 using random forest model

Table 38 Summary of important factors from SHAP value using random forest model

	Product name									
Factors	SKU01	SKU02	SKU03	SKU04	SKU05	SKU06	SKU07	SKU08	SKU09	SKU10
Price	(-) 1	(-) 1								
Month_1										
Month_2										
Month_3										
Month_4										
Month_5										
Month_6										
Month_7										
Month_8										
Month_9										
Month_10										
Month_11										
COVID-19					(-) 2	(-) 2		(-) 2		(-) 2
Store	(+) 2	(-) 2	(-) 3	(-) 1					(-) 2	
Welfare			(+) 4			(+) 3		(+) 3	(+) 3	(+) 3
%Discount			(+) 1				(+) 1		(+) 1	
Promotion period			(+) 2		(+) 1	(+) 1		(+) 1		(+) 1
Promotion period Lag1										
Promotion period Lag2	(-) 3			(-) 2						

Note: (i) Highlight color: Red - negative impact to the output, Green - positive impact to the output, yellow – small impact or insignificant to output; (ii) The number shows the order of the important factors to the prediction (1 is the most important)

Comparing the influencing factors for each product between linear regression with stepwise method in Table 21 and the random forest model in Table 38, their results were mostly similar. The monthly period factor may not influence sales since the findings show that it is small impact or insignificant to sales quantity. Unlike the stepwise result, the promotion period factor of SKU03 from the SHAP value had a positive effect on sales, which corresponds to the actual quantity that will increase when doing a promotion, as shown in Figure 92.

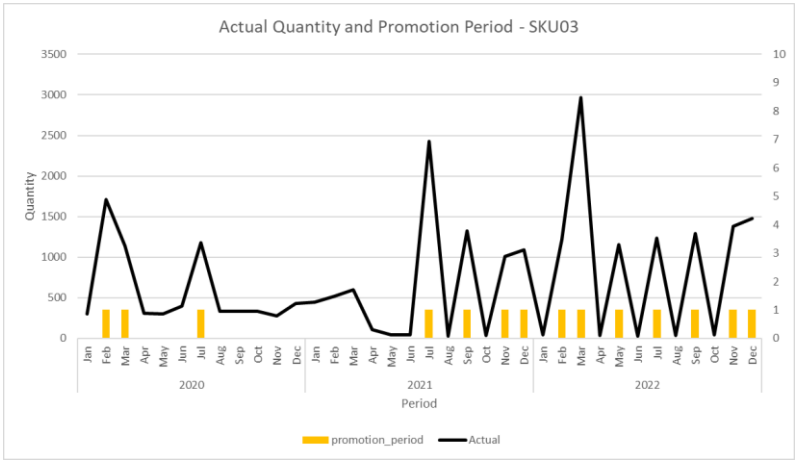
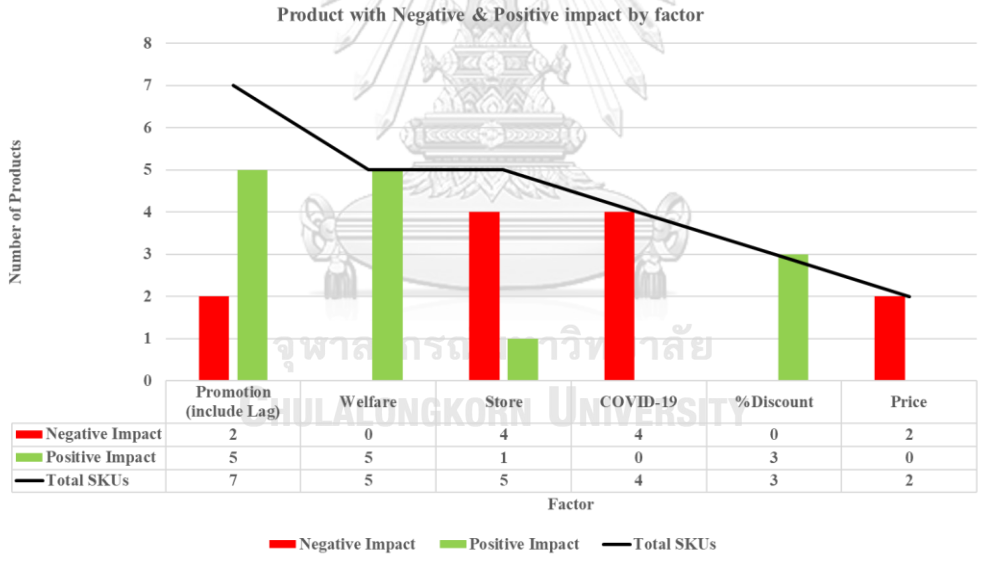


Figure 92 Quantity and promotion period of SKU03

Moreover, the SHAP value shows which factors are most important or have a high contribution to the prediction. From total 10 products, factors that impact negatively and positively to sales quantity are shown in Figure 93.



	Promotion (include Lag)	Welfare	Store	COVID-19	%Discount	Price
Negative Impact	2	0	4	4	0	2
Positive Impact	5	5	1	0	3	0
Total SKUs	7	5	5	4	3	2

Figure 93 Products with negative and positive impact by factors

The finding revealed that factors influencing sales were the promotion period, the subsidies and welfare programs, the number of stores, the number of COVID-19 new cases, discount percentage and selling price. Considering the factors that have a positive effect on sales, including the promotion period, the subsidies and welfare program, discount percentage and the number of stores, it was reasonable and respond to normal customer behaviors when

doing promotion or subsidies and welfare programs or having discount percentage, the sales will increase. Also, the negative ones are the number of stores, the number of COVID-19 new cases, the 2-lag promotion period and price factors. The number of COVID-19 new cases and the selling price factors are reasonable; when the number of COVID-19 new cases or the selling price increases, the sales will decrease. The number of stores factor that was found to have a negative effect on sales may be due to the store being located in an area far away from the target group or an area with competing stores. As a result, opening more stores does not increase sales. Moreover, the lag-2 promotion period factor has a negative effect on sales; it may be a result of the fact that customers bought a large amount of that product when promoting in the last two periods, causing them not to buy more product, leading to the sales decrease. Moreover, Figure 94 demonstrates the summary table showing the number of products and their contribution order to sales or order of importance.

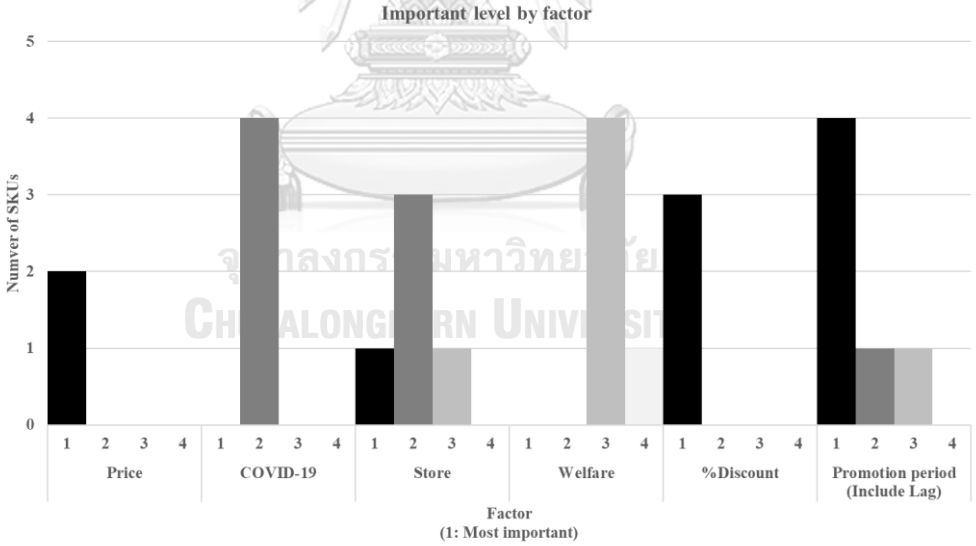


Figure 94 Summary factors and important order level of the 10 products
(Note: the lowest number of orders, the most important factor)

From Figure 94, the price and promotion factors had the most significant impact on the sales quantity of each product compared to other factors. It can be concluded that the most significant factors influencing sales quantity are price and promotion factors, especially the promotion factors

including discount percentage and promotion period, since they were the first highest contribution factors in four, three and two products, respectively.

According to the interpretation of MAPE values (Table 1), at a level of MAPE lower than 10% or highly accurate forecasting, it can be concluded that sales are highly sensitive to price, as seen in Figure 95, which illustrates the price factors and the sales quantity plot of an example of a product (SKU02). From Figure 95, it can be observed that when the price decreases, it results in an increase in the sales volume, where the lower the price, the greater the increase in quantity. Corresponding to SHAP value, as shown in Figure 83, high values of the selling price factor have a high negative contribution to the prediction or sales quantity, while low values have a high positive contribution.

Figure 95 Quantity and price of SKU02

As for the MAPE level of 10% to 20%, or good accuracy prediction level, it also has similar results to the highly accurate prediction level. The sales quantity of the products at this level seems to be sensitive to the percentage discount or price factor. The greater the increase in discount percentage (decrease in price), the more sales increase, as shown by the example of products (SKU03) in Figure 96. According to the SHAP value of the SKU03, as shown in Figure 84, high values of the discount percentage

factor have a high positive contribution to the prediction or sales quantity, while low values have a high negative contribution.

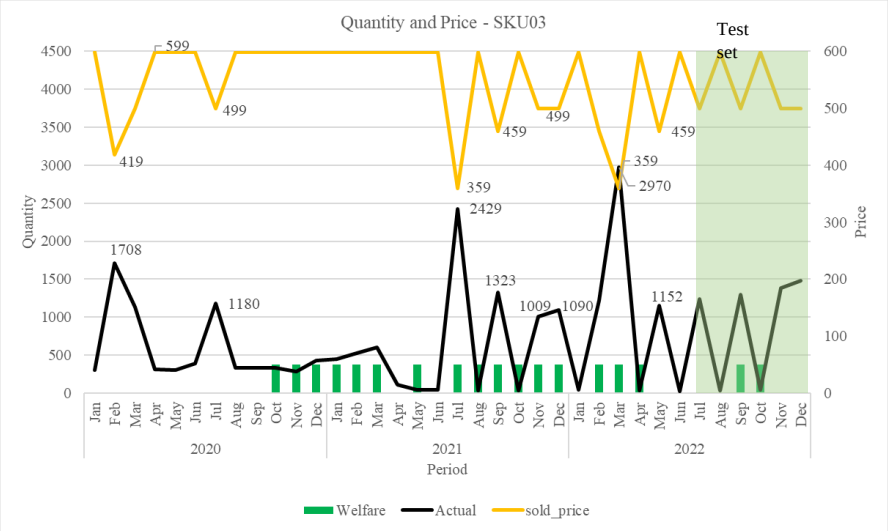


Figure 96 Quantity and price of SKU03

While the MAPE level of 20% to 50%, or reasonable prediction, has many patterns or conclusions about the affecting factors depending on each product. For the price factor, it affects sales quantity because reducing prices leads to increased sales, but it may not be as sensitive as at lower levels, where a reduction in price always results in an increase in sales. For instance, as presented as an example of a product (SKU04) in Figure 97, it can be observed that some periods at the shallow reducing price have higher sales than at the deeper reducing price, such as at the price of 34.5, which has more sales than at the price of 34, or at the price of 35, which has higher sales than at the price of 34.

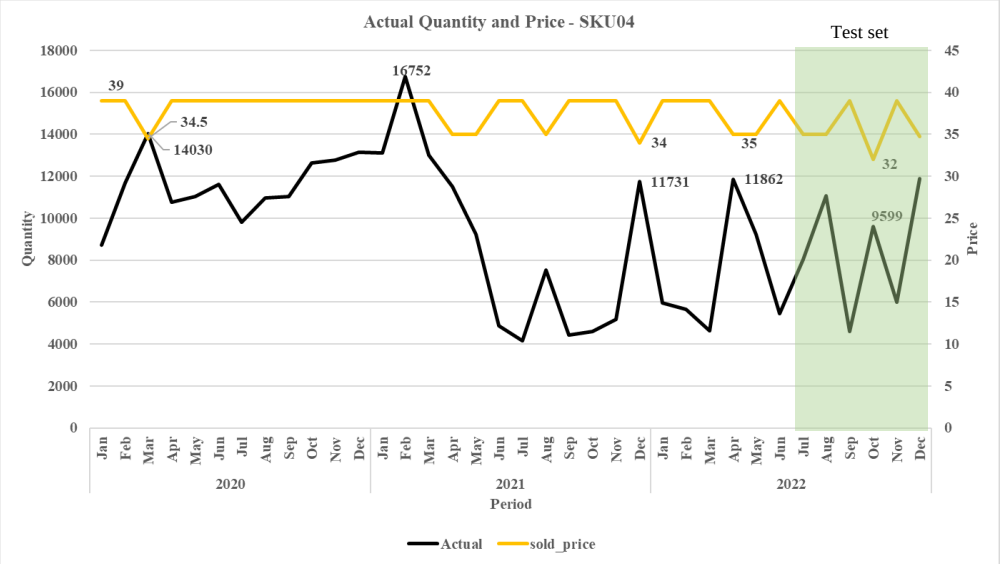


Figure 97 Quantity and price of SKU04

Figure 98 shows another product example (SKU10). There are other factors affecting sales, as the increase in sales from November 2020 to March 2021 might be a result of subsidies and welfare program effects. In addition, at the beginning of the first peak of COVID-19 new cases (around March 2021 to July 2022), sales were visibly dropping.

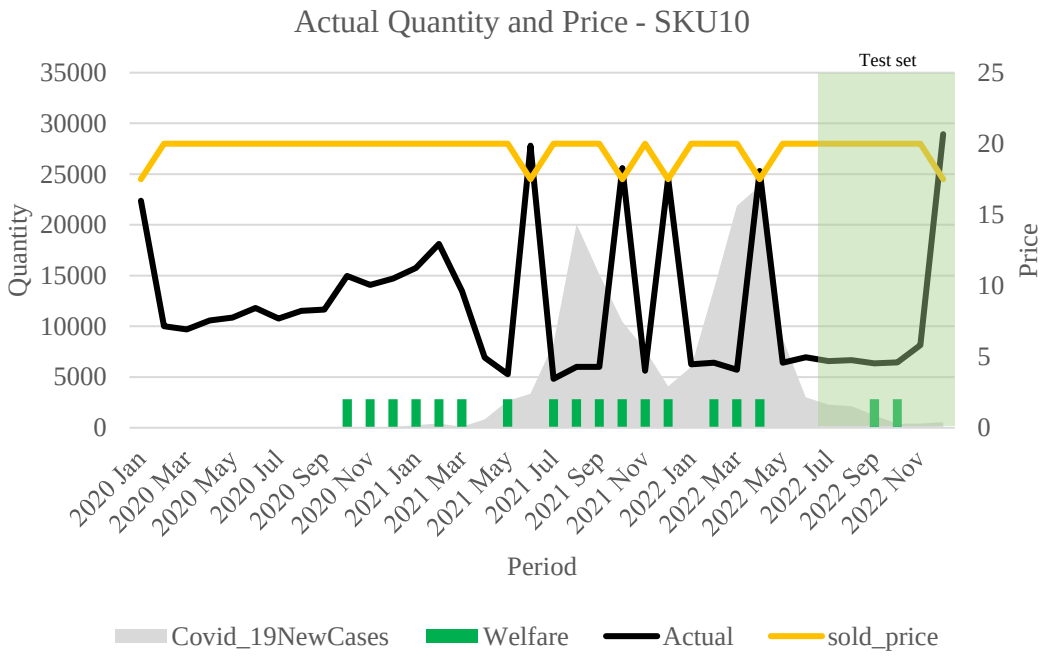


Figure 98 Quantity and price of SKU10

As for the level of MAPE greater than 50% or inaccurate accuracy prediction, the product is at the level shown in Figure 99. The product has a high MAPE on the testing data set, caused by a sudden increase in sales on the testing dataset (e.g., August 2022 and December 2022), which may be due to other factors that were not considered in the model such as other promotion campaigns or events. The result found that only the promotion period factor was significant. It seems that the price or percentage discount factors may have small effects or be insignificant since the sales increase while not reducing price at the first period caused by doing promotion (promotion period factor). Moreover, from October 2020 to February 2021, sales tended to increase, which may result in subsidies and welfare programs factor.

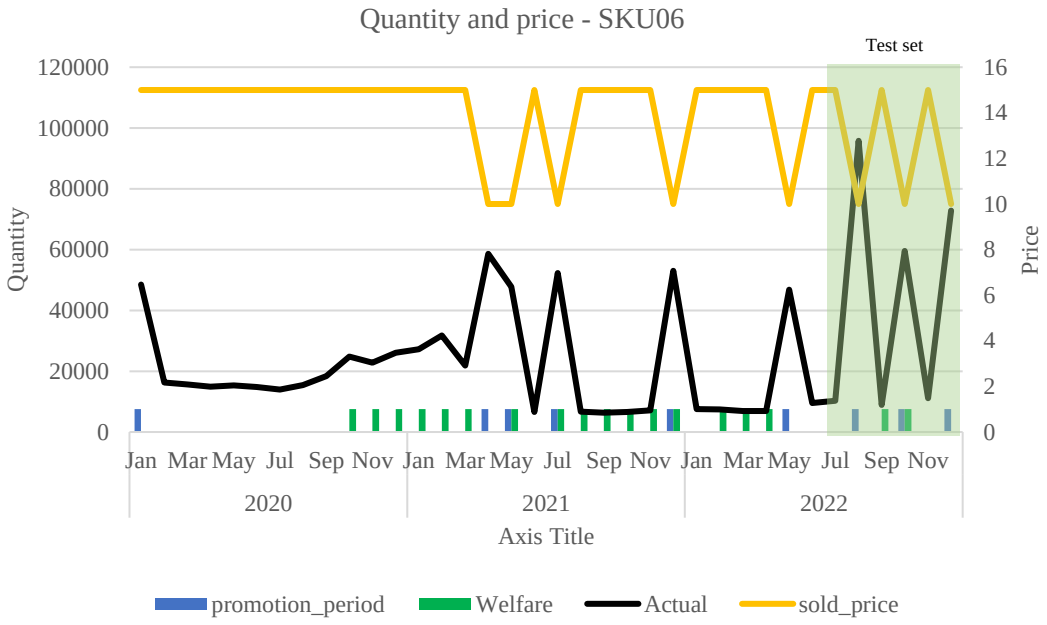


Figure 99 Quantity and price plot of SKU06

4.4.2 Performance by products

The best overall performance model was selected by the lowest WMAPE, however, the prediction of the 10 products may have both good and poor accuracy. For real-world application in retail business, the accurate prediction of each product may be used to make promotion plan. Therefore,

the most accurate model of individual products and their factors influencing sales quantity were analyzed.

4.4.2.1 Model evaluation and selection by products

The MAPE of the 10 products on the testing dataset using all the models and the summarized most accurate prediction model by SKU are demonstrated in tables 39 and 40, respectively.

Table 39 MAPE of the models of each SKU

Model	Product name										
	SKU01	SKU02	SKU03	SKU04	SKU05	SKU06	SKU07	SKU08	SKU09	SKU10	WMAPE
Linear Regression	**24.16%**	74.26%	52.09%	43.87%	36.18%	47.70%	22.51%	63.77%	51.81%	48.71%	43.92%
Random forest Case1	31.88%	50.47%	88.60%	50.46%	24.92%	40.35%	17.73%	30.89%	23.46%	22.38%	39.08%
Random forest Case2	38.52%	35.00%	176.56%	46.05%	**11.00%**	49.77%	15.73%	21.68%	25.46%	22.29%	45.75%
Random forest Case3	39.80%	20.80%	65.82%	46.66%	23.53%	**27.87%**	12.09%	19.49%	22.49%	**13.02%**	31.80%
Random forest Case4	38.03%	49.79%	88.53%	26.73%	17.47%	38.30%	**11.12%**	**12.41%**	22.45%	14.55%	34.58%
Random forest Case5	33.99%	8.52%	15.36%	31.81%	29.09%	55.22%	19.93%	27.55%	20.85%	33.48%	**28.15%**
Random forest Case6	37.34%	9.70%	114.83%	35.11%	27.69%	50.53%	16.36%	38.94%	23.24%	36.56%	39.47%
Random forest Case7	45.94%	**8.48%**	40.57%	32.31%	29.09%	55.24%	18.35%	26.79%	22.50%	37.85%	33.54%
Random forest Case8	43.12%	8.59%	17.84%	30.83%	29.09%	59.32%	16.50%	24.85%	**18.94%**	35.79%	30.26%
XGBoost Case1	35.02%	93.83%	151.07%	38.26%	47.78%	53.64%	40.01%	44.57%	32.43%	41.75%	57.43%
XGBoost Case2	59.66%	67.75%	122.20%	38.25%	47.32%	57.02%	34.86%	31.16%	38.00%	34.90%	56.02%
XGBoost Case3	37.13%	82.71%	135.60%	39.61%	31.06%	41.55%	36.54%	36.34%	29.43%	38.69%	51.24%
XGBoost Case4	40.73%	58.99%	123.93%	39.26%	59.07%	53.30%	31.14%	44.27%	35.22%	48.85%	53.47%
XGBoost Case5	32.94%	13.15%	21.68%	33.91%	40.08%	65.10%	18.63%	42.67%	36.51%	43.32%	33.98%
XGBoost Case6	49.74%	19.41%	19.24%	31.19%	21.61%	60.00%	25.11%	41.73%	30.06%	32.94%	34.76%
XGBoost Case7	31.01%	26.73%	29.95%	30.45%	28.92%	58.85%	16.22%	34.54%	35.00%	38.71%	**32.60%**
XGBoost Case8	40.53%	21.95%	23.59%	35.96%	50.62%	70.67%	24.37%	34.68%	27.12%	39.35%	37.54%
ANN Case1	37.32%	35.62%	39.04%	36.09%	31.23%	50.52%	39.73%	35.84%	40.48%	42.07%	**38.32%**
ANN Case2	47.02%	33.94%	512.47%	61.37%	73.45%	78.44%	47.54%	65.37%	51.02%	57.79%	100.58%
ANN Case3	36.54%	70.97%	352.08%	**16.37%**	76.93%	58.07%	87.61%	64.10%	76.83%	28.87%	83.19%
ANN Case4	34.32%	71.33%	66.43%	67.47%	93.99%	92.37%	91.86%	87.24%	42.77%	40.62%	65.24%
ANN Case5	44.09%	15.20%	66.35%	43.41%	15.70%	62.11%	17.96%	79.78%	53.49%	59.45%	44.06%
ANN Case6	50.19%	38.32%	537.93%	67.85%	72.39%	77.01%	50.83%	66.68%	53.42%	57.73%	105.16%
ANN Case7	38.14%	72.42%	389.05%	23.77%	96.77%	87.92%	87.22%	81.94%	86.60%	31.75%	94.90%
ANN Case8	36.64%	73.30%	105.65%	81.44%	94.21%	93.29%	92.47%	88.32%	63.16%	50.76%	73.49%
Random forest Case5 parallel XGBoost Case7	32.13%	33.12%	38.01%	30.01%	34.70%	58.18%	16.86%	37.30%	33.46%	41.54%	35.16%
Random forest Case5 parallel ANN Case1	33.46%	17.67%	39.48%	23.69%	24.50%	53.46%	38.72%	26.38%	22.77%	36.67%	31.56%
XGBoost Case7 parallel ANN Case1	29.15%	28.43%	32.68%	25.87%	31.61%	58.39%	16.45%	35.58%	37.26%	40.97%	32.81%
Random forest Case5 series XGBoost Case7	32.79%	11.97%	**13.26%**	31.94%	29.57%	50.91%	12.42%	28.46%	25.23%	35.03%	**27.65%**
Random forest Case5 series ANN Case1	33.99%	8.52%	15.16%	30.83%	29.09%	55.23%	19.24%	27.81%	20.87%	33.99%	28.04%
XGBoost Case7 series ANN Case1	31.00%	26.78%	30.70%	29.58%	28.92%	58.83%	16.23%	34.79%	34.97%	38.82%	32.61%
XGBoost Case7 series Random forest Case5	30.95%	26.74%	32.92%	30.41%	30.18%	59.33%	17.02%	39.73%	34.26%	38.95%	33.45%
ANN Case1 series Random forest Case5	40.28%	38.15%	37.73%	32.07%	42.56%	42.50%	36.62%	40.54%	38.45%	45.04%	39.35%
ANN Case1 series XGBoost Case7	35.98%	42.21%	65.22%	34.98%	28.27%	28.65%	38.83%	49.86%	52.48%	36.99%	40.35%
Min	24.16%	8.48%	13.26%	16.37%	11.00%	27.87%	11.12%	12.41%	18.94%	13.02%	27.65%

Table 40 Summarized the best model and MAPE of each SKU

Product name	Subcategory	Weight	Best Model	MAPE		
				Train	CV	Test
SKU01	Basic skin caere	0.22	Linear Regression	29.62%	-	24.16%
SKU02	Anti-aging	0.11	Random forest using significant factor & factors of other products in the same group by subcategory	8.81%	18.74%	8.48%
SKU03	Anti-aging	0.10	Hybrid model of Random forest and XGBoost	8.36%	37.67%	13.26%
SKU04	Whitening	0.10	ANN using all factors & factors of other products in the same group by subcategory	16.54%	50.73%	16.37%
SKU05	Men	0.10	Random forest using all factors & factors of other products in the same group by category	13.14%	15.94%	11.00%
SKU06	Whitening	0.09	Random forest using all factors & factors of other products in the same group by subcategory	9.40%	17.45%	27.87%
SKU07	UV protection	0.07	Random forest using all factors & factors of other products in the same group by K-means	12.01%	22.13%	11.12%
SKU08	Anti-aging	0.07	Random forest using all factors & factors of other products in the same group by K-means	7.65%	13.48%	12.41%
SKU09	Whitening	0.07	Random forest using significant factors & factors of other products in the same group by K-means	18.37%	23.47%	18.94%
SKU10	Whitening	0.07	Random forest using all factors & factors of other products in the same group by subcategory	8.09%	16.04%	13.02%
			WMAPE	15.38%	19.31%	16.67%

According to Table 40, each SKU has a different best model to predict sales quantity and the WMAPE on testing set were 16.67%. The MAPE values on the training dataset and cross-validation set of 10 products were quite close, except for SKU02, SKU03, and SKU04, which had a large difference in

MAPE between the training dataset and cross-validation set and low performance in the cross-validation set but high performance in the test set. It may be due to the small testing dataset, or the test data may be easier to predict than the training data, meaning the model seems to do well in the particular group of data that is present in the test set or lacks generalization and may not perform as well in real-world scenarios. So, increasing the number of testing data, considering different evaluation metrics, or using different cross-validation techniques may help provide a more comprehensive evaluation of the model's performance.

The single machine learning model was selected to be the most accurate prediction model in most products, except SKU01 and SKU03, which selected linear regression and the series hybrid model, respectively. The most selected model was a random forest model, which resulted in the best performance in 7 products. For the factors used to train the model, using all factors may be better than using significant ones, as shown by most products using all factors. Moreover, considering factors of other products in the same group resulted in the best performance of 8 products, including 4, 3, and 1 product clustering products by subcategory, K-means method and category, respectively. Therefore, clustering similar products into the same group, especially by subcategory or K-means method, and considering them as one of the factors to predict sales may improve the performance.

4.4.2.2 Factor analysis by products

For SKU01, the most accurate model was linear regression with MAPE 24.16%, which may be due to the assumptions not being violated, including a linear relationship between predictors and the target variable; the predictors are not highly correlated since VIF is small; and the residuals are independent and identically normal distributed, as shown by the linear regression result in Figure A1 (see in Appendix). The significant factor from Table 21 included selling price, number of active stores and lag-2 promotion period.

SKU02 selected the random forest using significant factors from the stepwise method and considered the factors of other products in the same subcategory. The SHAP value is shown in Figure 100. It found that the most important or high contribution to the prediction was the selling price and number of stores, which had a highly negative impact on sales quantity. Moreover, the other products, SKU14 and SKU18, may have a small effect on SKU02's sales.

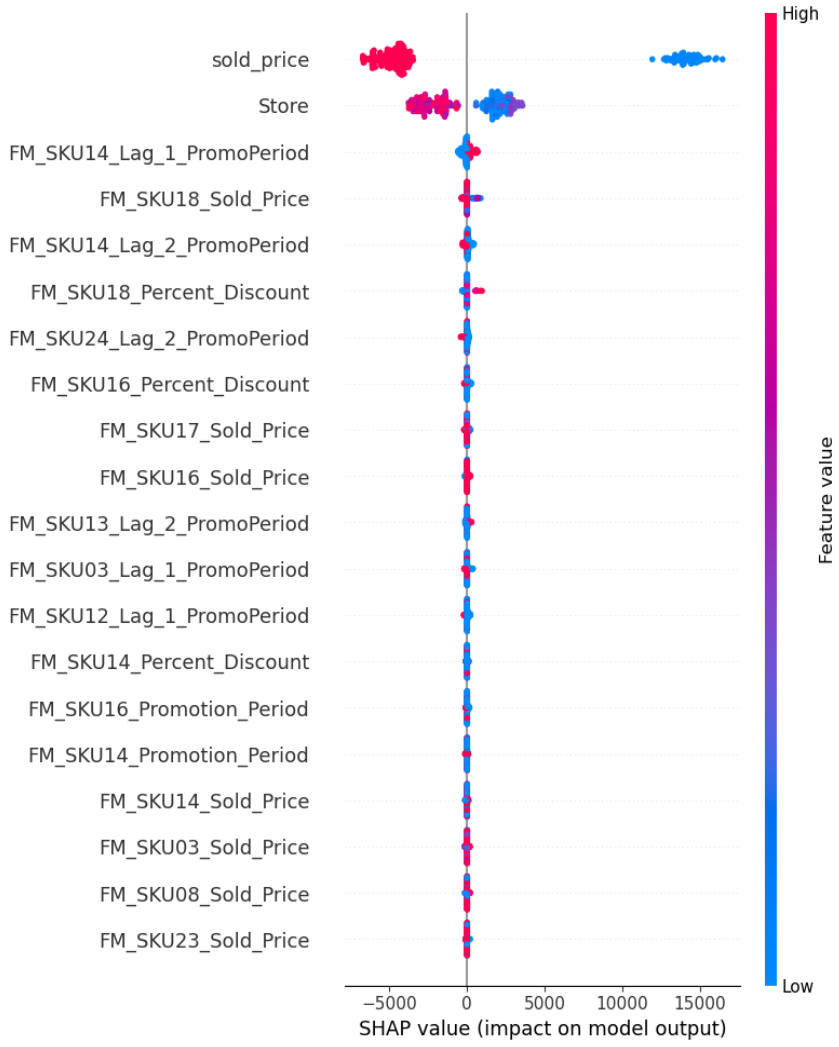


Figure 100 SHAP value of SKU02 using the best model

The most accurate model of SKU04 was the ANN model using all factors with considering factors of other products in the same subcategory. As

shown in Figure 101, the number of stores had a negative effect on sales, which correspond to linear regression and a random forest model using significant factors from stepwise method. Factors including subsidies and welfare programs and the number of COVID-19 new cases also had a greater impact on sales quantity. For the factors of other products in the same subcategory, the discount percentage and selling price of SKU09 and SKU73 influenced sales. When reducing the selling price of SKU09 or SKU73, sales may increase. While the discount percentage of SKU09 or SKU73 increases, sales will decrease.

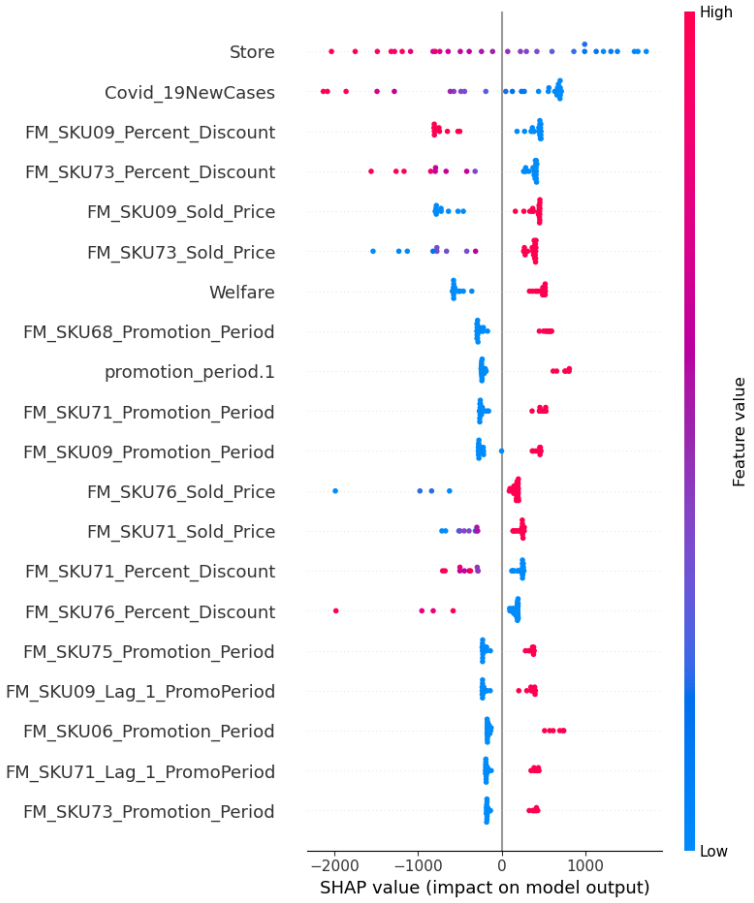


Figure 101 SHAP value of SKU04 using the best model

For FKU05, the best model was a random forest model using all the factors and considering the factors of all the products in the category. Figure 102 presents the SHAP value obtained from the model. It found that the first

three features that most affected sales were price and promotion factors, which are price factors, followed by discount percentage and promotion period factors, unlike the result from the random forest model using significant factors, where the promotion period has the highest contribution to prediction. The COVID-19 new cases and the number of active stores factors also had a negative effect on sales. Moreover, factors from other products had affected sales. For example, when SKU02 increases in price, it may affect the decrease in sales quantity of SKU05, or when SKU01 is doing promotions, it may increase SKU05's sales.

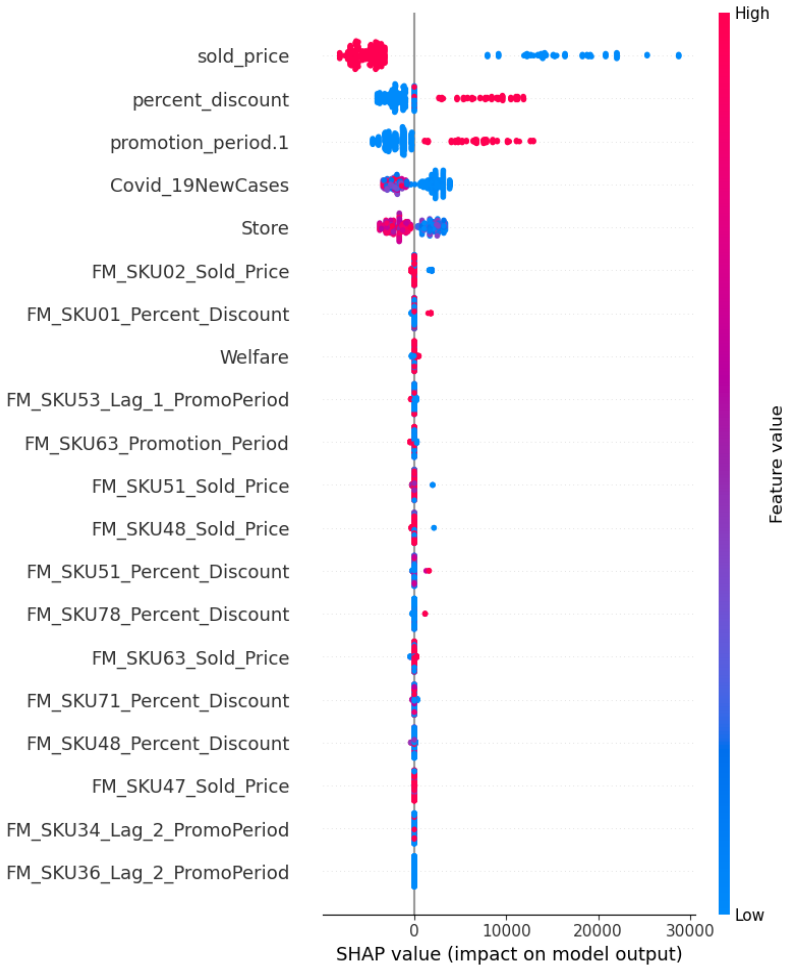


Figure 102 SHAP value of SKU05 using the best model

Figure 103 shows the SHAP value that was found using the random forest model with all factors and taking into account factors from other products in the same subcategory. The promotion period was the most significant factor positively affecting sales, followed by the discount percentage, the number of COVID-19 new cases, the number of stores, the price factors. The number of COVID-19 new cases, the number of stores, the price factors had negative effects, while the promotion period, discount percentage and subsidies and welfare programs had a positive effect on sales. The lag in the promotion period of SKU09 and SKU75 also had a noticeable impact on the model, indicating that when SKU09 or SKU75 are doing promotions, it may decrease SKU06's sales in the next period.

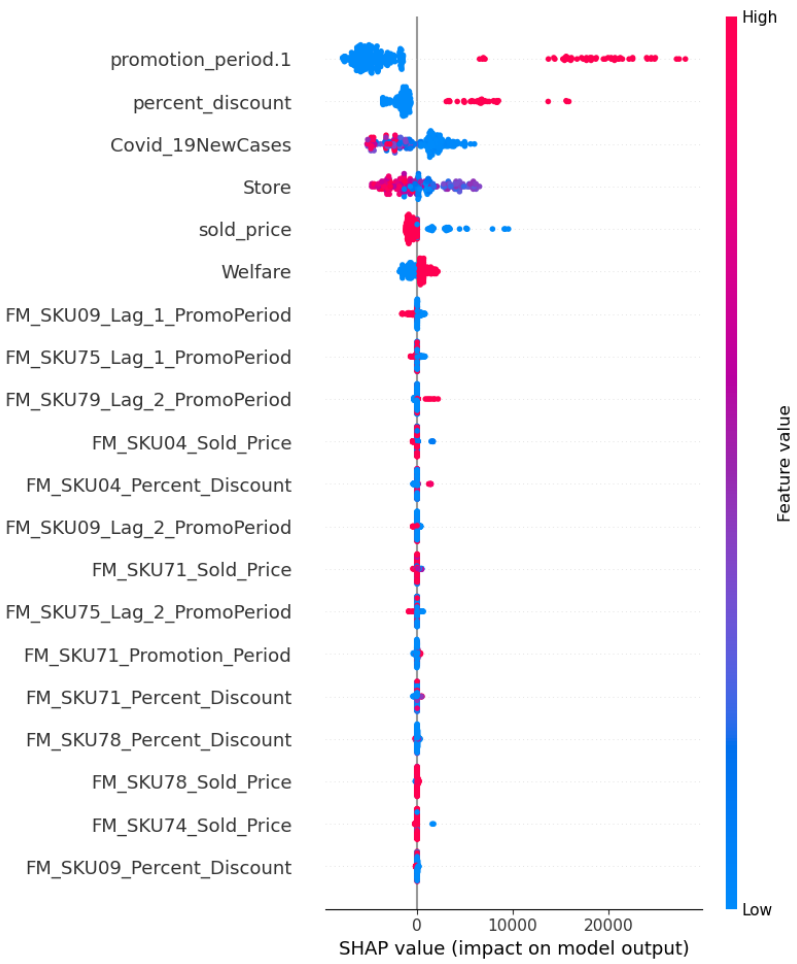


Figure 103 SHAP value of SKU06 using the best model

For SKU07, the SHAP value was obtained from the random forest model using all factors and considering factors of other products in the same K-means clustering group, as shown in Figure 104. The first three features that most affected sales were price and promotion factors, which are price factors that have a negative effect on sales, followed by discount percentage factors and promotion period factors that have a positive effect on sales. The number of stores and the number of COVID-19 new cases negatively affected sales, and the subsidies and welfare program factor had a positive effect on sales. For other products in the same K-means clustering group, product SKU69 seems to impact SKU07's sales. When SKU69 is doing promotions or reducing prices, it may lead to an increase in sales of SKU07.

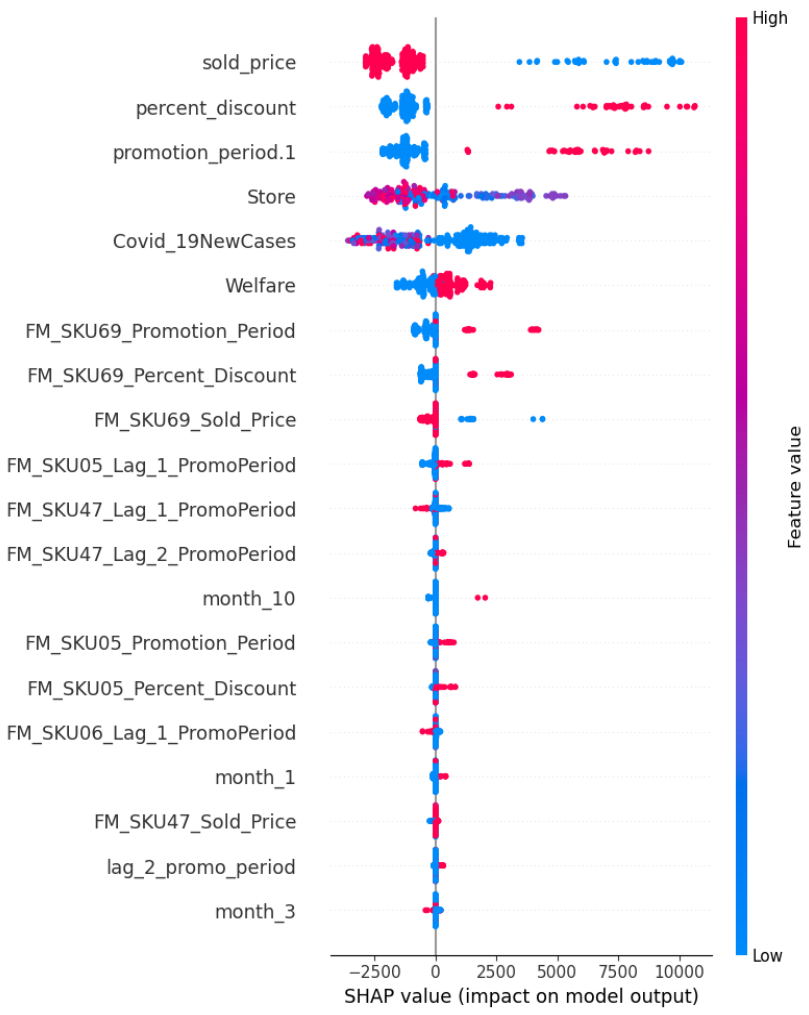


Figure 104 SHAP value of SKU07 using the best model

The random forest model, which used all factors and other products in the same K-means clustering group, was the best fit for SKU08. As shown in Figure 105, the first top five of the important features were similar to SKU07. Another product in the same cluster that may impact SKU08's sales is SKU01. When SKU01 does a promotion or increases the discount percentage, it may increase sales of SKU08.

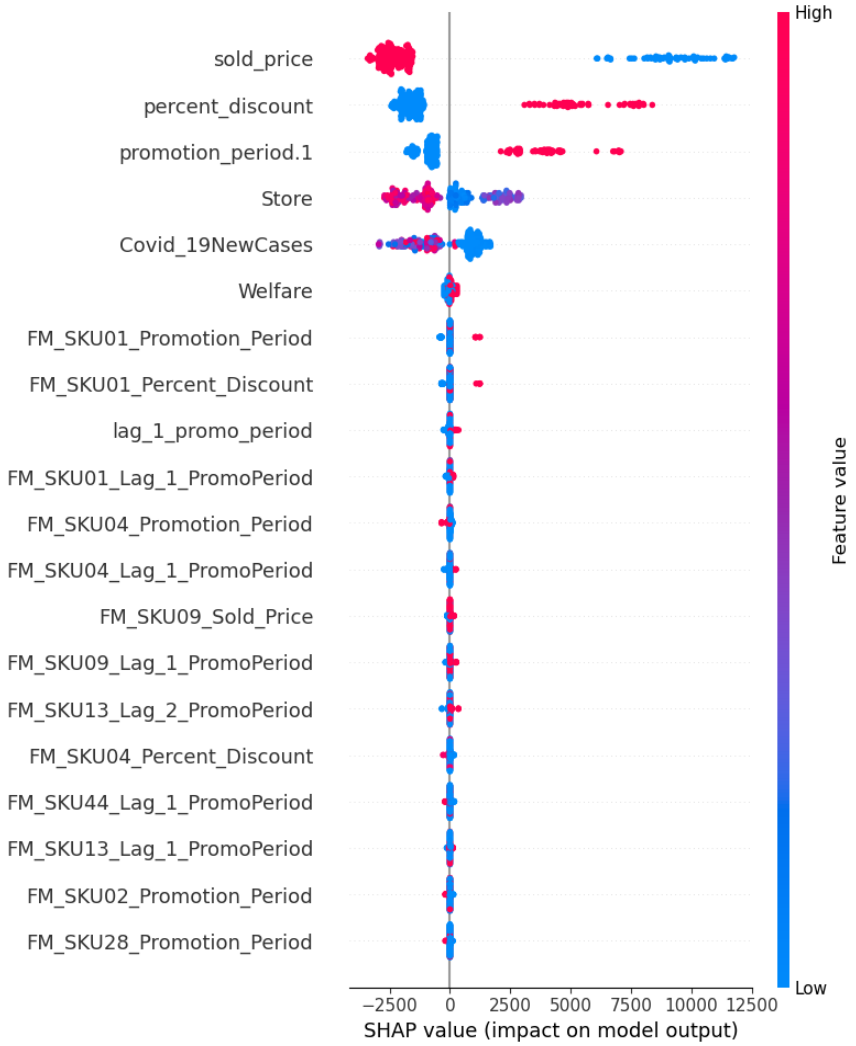


Figure 105 SHAP value of SKU08 using the best model

For SKU09, the random forest model using significant factors and considering factors of other products in the same K-means clustering group was outperformed. Figure 106 shows the SHAP value obtained by the model.

The first top three features had as similar an effect on sales as the random forest model using significant factors. Besides, promotion factors of SKU08 and SKU44 may affect sales. When SKU08 increases the discount percentage, it may increase in sales. After the period that SKU08 is promoting, the sales of SKU09 may decrease. While SKU44 is being promoted, the sales of SKU09 may increase in the next period.

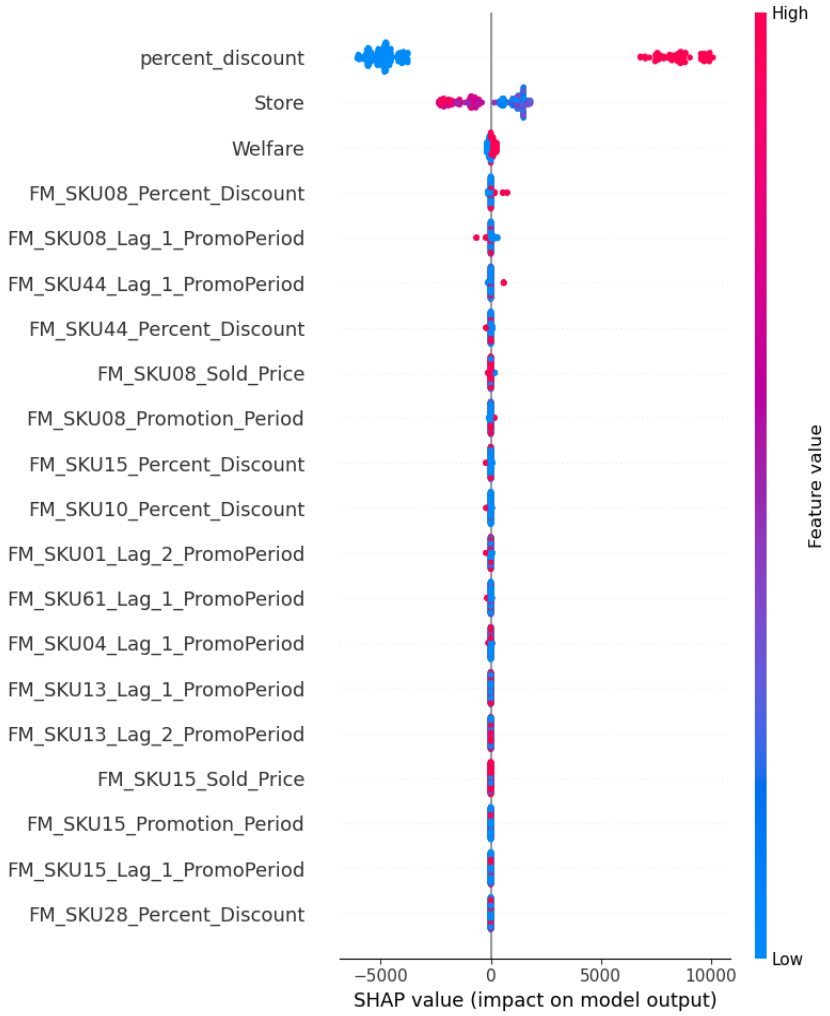


Figure 106 SHAP value of SKU09 using the best model

The SHAP value of SKU10, as shown in Figure 107, was obtained from the random forest model using all factors and considering factors of other products in the same subcategory. From the figure, the promotion period factor was the most important factor that positively influenced sales. The number of stores, the number of COVID-19 new cases, and the selling price

had a negative effect on sales, while the discount percentage and the subsidies and welfare programs were positively affecting sales. For the effect of other products, the price and promotion factors of SKU69 and SKU70 had an impact on sales quantity. As SKU69 does promotions, increases the discount percentage, or reduces the price, sales of SKU10 may increase. And when SKU70 is doing the promotion or reducing price, it may lead to an increase in the sales of SKU10.

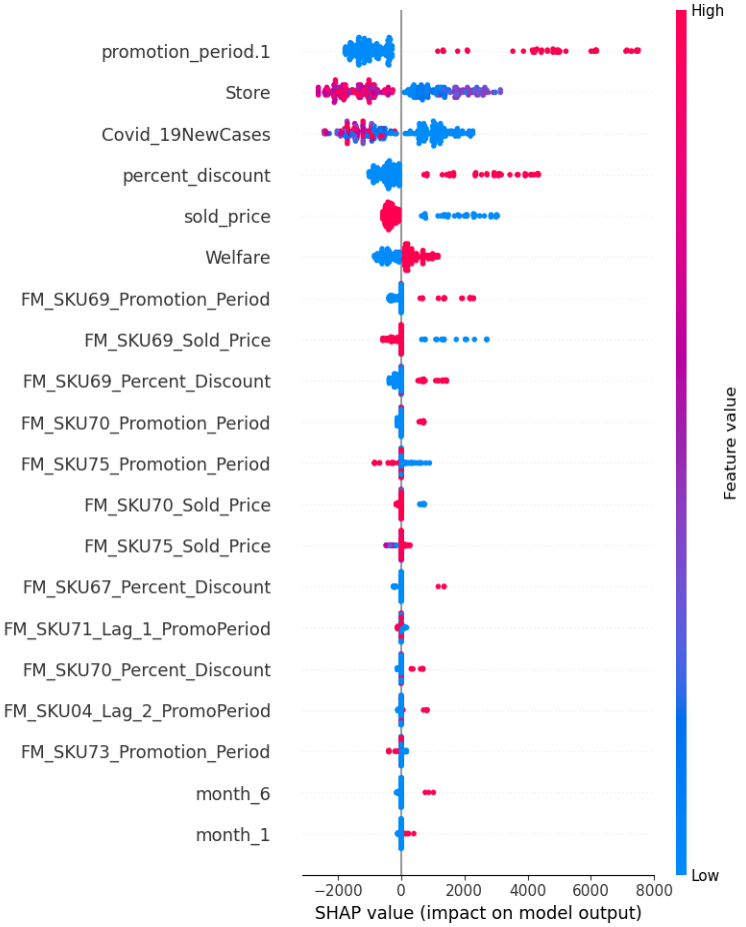


Figure 107 SHAP value of SKU10 using the best model

In conclusion, the factors that impact sales of individual SKU as shown in Table 41.

Table 41 Influencing factors considered by individual SKU

		Product name									
	Factors	SKU01	SKU02	SKU03	SKU04	SKU05	SKU06	SKU07	SKU08	SKU09	SKU10
	Price	(-)	(-) 1	Hybrid model		(-) 1	(-) 5	(-) 1	(-) 1		(-) 5
	Month_1										
	Month_2										
	Month_3										
	Month_4										
	Month_5										
	Month_6										
	Month_7										
	Month_8										
	Month_9										
	Month_10										
	Month_11										
	COVID-19				(-) 2	(-) 4	(-) 3	(-) 5	(-) 5		(-) 3
	Store	(+)	(-) 2		(-) 1	(-) 5	(-) 4	(-) 4	(-) 4	(-) 2	(-) 2
	Welfare				(+) 7		(+) 6	(+) 6	(+) 6	(+) 3	(+) 6
	%Discount					(+) 2	(+) 2	(+) 2	(+) 2	(+) 1	(+) 4
	Promotion period					(+) 3	(+) 1	(+) 3	(+) 3		(+) 1
	Promotion period Lag1										
	Promotion period Lag2	(-)									
SKU01	%Discount					(+) 7			(+) 8		
	Promotion period								(+) 7		
SKU02	Price					(-) 6					
SKU08	%Discount									(+) 4	
	Promotion period Lag1									(-) 5	
SKU09	Price				(+) 5						
	%Discount				(-) 3						
	Promotion period Lag1						(-) 7				
SKU14	Promotion period Lag1		(+) 3								
SKU18	Price										
SKU44	Promotion period Lag1									(+) 6	
SKU69	Price							(-) 9			(-) 8
	%Discount							(+) 8			(+) 9
	Promotion period							(+) 7			(+) 7
SKU70	Promotion period										(+) 10
SKU73	Price				(+) 6						
	%Discount				(-) 4						
SKU75	Promotion period Lag1						(-) 8				

Other Products

Note: (i) Highlight color: Red - negative impact to the output, Green - positive impact to the output, yellow – small impact or insignificant to output; (ii) The number shows the order of the important factors to the prediction (1 is the most important)

According to Table 41, from total products (exclude SKU03), the factors influencing sales included selling price, discount percentage, promotion period, the number of stores, the number of COVID-19 new cases, as well as factors from other products which are selling price, discount percentage, promotion period and lag-1 promotion period. The factor of monthly periods may not influence sales since the findings presented that it is insignificant to sales quantity. The selling price, the number of stores, the number of COVID-19 new cases and the lag-2 promotion period factors had a negative effect on sales quantity, which means that the increase in selling price, the number of stores and the number of COVID-19 new cases leading to decrease in sales. Doing promotions may decrease sales in the next two

periods. Moreover, factors with a positive effect on sales, including the discount percentage, the promotion period, the subsidies and welfare program factors, the increase in discount percentage, or when doing promotion or subsidies and welfare program leading to an increase in sales. Considering factors of other similar products in the same group, there were both positive and negative effects on sales depending on each product. However, they had a smaller contribution impacting sales compared to the product factors.

Chapter 5 Conclusion

This research aims to identify accurate prediction models to predict monthly sales of beauty products of a case-study retail company and identify effects of factors that influence sales quantity. The dataset used to construct prediction model is from January 2020 to December 2022 (36 months), where 30 months are for model training and the rest 6 months are for model testing. Prediction models including linear regression, random forest, extreme gradient boosting (XGBoost), artificial neural networks (ANN), and hybrid models were applied and evaluated their overall performance accuracy by using weighted mean absolute percentage error (WMAPE), where weight is determined by product revenue. Meanwhile, the factors were used for training three machine learning techniques: random forest, XGBoost, and ANN, which consisted of 8 cases by using either all factors or selected significant factors from the stepwise method and without or with consideration of exogenous factors of other products in the same group clustered by three criteria: category, subcategory, or K-means method.

5.1 Result summary

The MAPE results on testing data found that the most accurate prediction model of each product was different, which had the WMAPE on the test of 16.67%. Nonetheless, the most accurate prediction model of all products may be chosen for use in retail. According to the testing results, the machine learning models have a higher overall accuracy in prediction than the traditional model of linear regression, since their WMAPE is lower. The best model of each machine learning technique was the random forest model using significant factors and not considering factors of other products; the XGBoost model using significant factors and considering factors of other products in the same subcategory; and the ANN model using all the factors and not considering factors of other products with WMAPE of 28.15%, 32.60%, and 38.32%, respectively. Meanwhile, the overall prediction performance of the series hybrid model of random forest and XGBoost outperformed with a WMAPE of

27.65%. However, it was 0.5% better than the random forest model, and it took about 5 times longer to compute. In conclusion, for overall performance, the random forest model using significant factors and not considering other products was selected to be the most suitable model to predict monthly sales of beauty products in this research.

From the SHAP values obtained from the random forest model of beauty products, the factors influencing sales quantity in most products were promotion period, subsidies and welfare programs, the number of stores, the number of COVID-19 new cases, discount percentage and selling price, while the monthly period factor does not significantly influence sales quantity. The price and promotion factors had the greatest impact on sales in most products. From the results, the promotion period was the most significant factor affecting most products' sales, followed by discount percentage and selling price factors. Products of high accuracy prediction (MAPE lower than 20%) are very sensitive to price or discount percentage, observing the lower price or the larger discount percentage, the greater sales quantity will could be. For other products, reducing the price or increasing the discount percentage may or may not help the sales increase; there could be other factors. During the promotion period, subsidies and welfare programs could increase in sales. The increase in the number of COVID-19 new cases also had a negative effect on sales; when COVID-19 cases are high, sales will drop.

In retail applications, the best performance model and factors impacting the sales quantity of individual products may be considered for accurate prediction. The WMAPE on the test set using each product's most accurate prediction models was 16.67%. For each product, the best models to accurately predict sales were different, including linear regression, the hybrid model of random forest and XGBoost, the random forest model and the ANN model. The result showed that the random forest model provided the lowest MAPE on most products. Using all factors and considering factors of other products in the same group results in low MAPE. Moreover, clustering similar products into the same group, especially by subcategory or K-means method, and taking other product factors into account to predict sales may improve performance. Factors including selling price, discount percentage, promotion period, the number of stores, and the number of new COVID-19 cases had similar effects on

the overall performance, and the month factor was also insignificant to sales quantity. Additionally, factors of other products in the same group, including selling price, discount percentage, promotion period and lag-1 promotion period, were found to have both negative and positive effects on product sales depending on the product. However, the factors of other products had a minor impact on the product sales compared to those of the product.

5.2 Limitation and Recommendation

As this paper used the monthly sales of only 36 months in total, there is a limited amount of available data. However, data in the retail industry is obtained weekly or daily and more observations, the result may be enhanced in future studies. For the products that do not have as good prediction performance on the test dataset as on the train dataset, it may be a result of the change in consumer behavior caused by other factors that may be considered in the future work. If the data is more detailed, such as promotion types, product placement, advertising campaigns, or company events, it may improve the prediction performance and provide more insights into which promotions affecting sales. Moreover, using the sales of each store instead of the total sales of all stores may provide insight into the store's effects on sales, such as store location or store area, for further analysis of which stores should do promotions to boost sales.

For future studies, the models may be applied to a larger number of products in the case-study retail company since this research selected only 10 products to predict sales. Instead of using elbow method to find optimal number of K in K-means method, the silhouette method can be conducted. Further investigations may apply more advance models, such as applying the Long Short-Term Memory (LSTM) or hybrid with this model to predict future demand. This approach has the potential to enhance the accuracy of predictions.

Appendix

Stepwise Selection of Terms

α to enter = 0.05, α to remove = 0.05

Regression Equation

QTY = 198417 - 6450 lag_2_promo_period - 4187 sold_price + 27.98 Store

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	198417	29797	6.66	0.000	
lag_2_promo_period	-6450	2363	-2.73	0.011	1.53
sold_price	-4187	561	-7.46	0.000	1.54
Store	27.98	8.24	3.40	0.002	1.65

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4189.65	85.51%	83.84%	73.86%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	2693055859	897685286	51.14	0.000
lag_2_promo_period	1	130758626	130758626	7.45	0.011
sold_price	1	977413543	977413543	55.68	0.000
Store	1	202437107	202437107	11.53	0.002
Error	26	456381268	17553126		
Total	29	3149437127			

Fits and Diagnostics for Unusual Observations

Obs	QTY	Fit	Resid	Std Resid	
1	12328	22199	-9871	-3.06	R X
3	5517	-579	6096	1.90	X
29	43957	32385	11572	3.06	R

R Large residual
X Unusual X

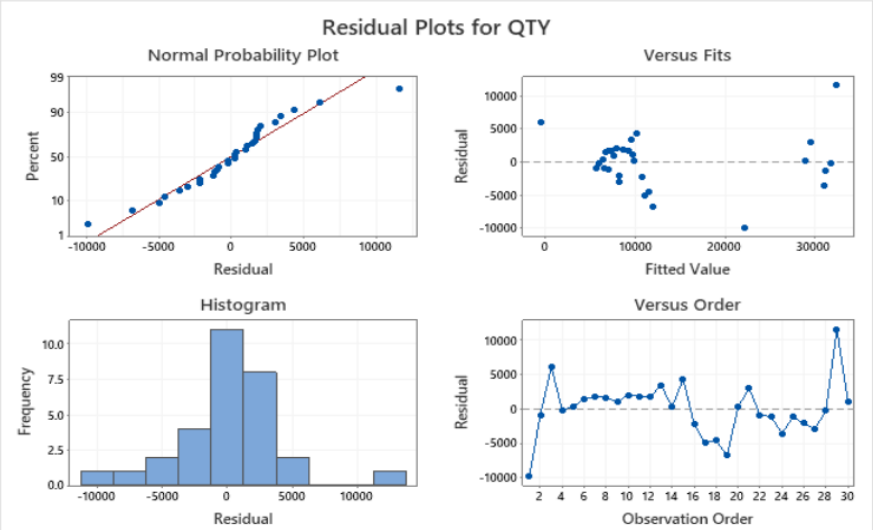


Figure A1 The result of SKU01 by Linear regression model

Stepwise Selection of Terms

α to enter = 0.05, α to remove = 0.05

Regression Equation

QTY = 103357 - 2126.0 sold_price + 3042 month_2 - 23.00 Store

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	103357	4641	22.27	0.000	
sold_price	-2126.0	97.3	-21.86	0.000	1.09
month_2	3042	1330	2.29	0.031	1.01
Store	-23.00	3.47	-6.64	0.000	1.08

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2176.75	95.00%	94.43%	93.25%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	2342020859	780673620	164.76	0.000
sold_price	1	2264305620	2264305620	477.88	0.000
month_2	1	24775905	24775905	5.23	0.031
Store	1	208677182	208677182	44.04	0.000
Error	26	123193966	4738229		
Total	29	2465214825			

Fits and Diagnostics for Unusual Observations

Obs	QTY	Fit	Resid	Std Resid	
2	34878	34600	278	0.19	X
15	11226	6526	4700	2.21	R
26	3285	5956	-2671	-1.65	X

R Large residual
X Unusual X

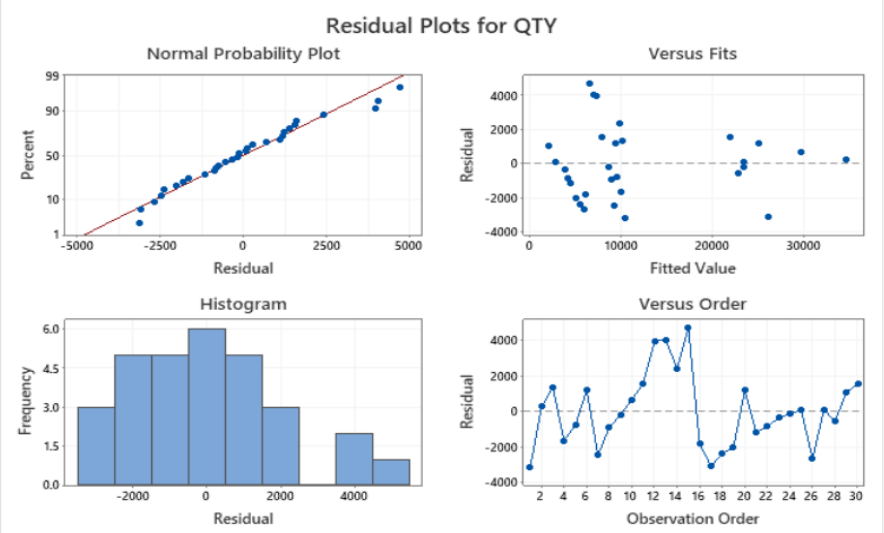


Figure A2 The result of SKU02 by Linear regression model

Stepwise Selection of Terms

α to enter = 0.05, α to remove = 0.05

Regression Equation

QTY = 678 + 65.07 percent_discount - 345 promotion_period + 326.7 month_3 - 0.852 Store
+ 121.1 Welfare

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	678	143	4.73	0.000	
percent_discount	65.07	5.34	12.18	0.000	6.71
promotion_period	-345	141	-2.45	0.022	6.44
month_3	326.7	91.9	3.56	0.002	1.11
Store	-0.852	0.248	-3.43	0.002	1.28
Welfare	121.1	58.2	2.08	0.048	1.23

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
143.329	96.80%	96.13%	92.30%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	14919494	2983899	145.25	0.000
percent_discount	1	3049370	3049370	148.44	0.000
promotion_period	1	123575	123575	6.02	0.022
month_3	1	259693	259693	12.64	0.002
Store	1	241575	241575	11.76	0.002
Welfare	1	88957	88957	4.33	0.048
Error	24	493037	20543		
Total	29	15412531			

Fits and Diagnostics for Unusual Observations

Obs	QTY	Fit	Resid	Std Resid	
3	1136.0	1361.4	-225.4	-2.38	R
27	2970.0	2735.1	234.9	2.61	R X

R Large residual
X Unusual X

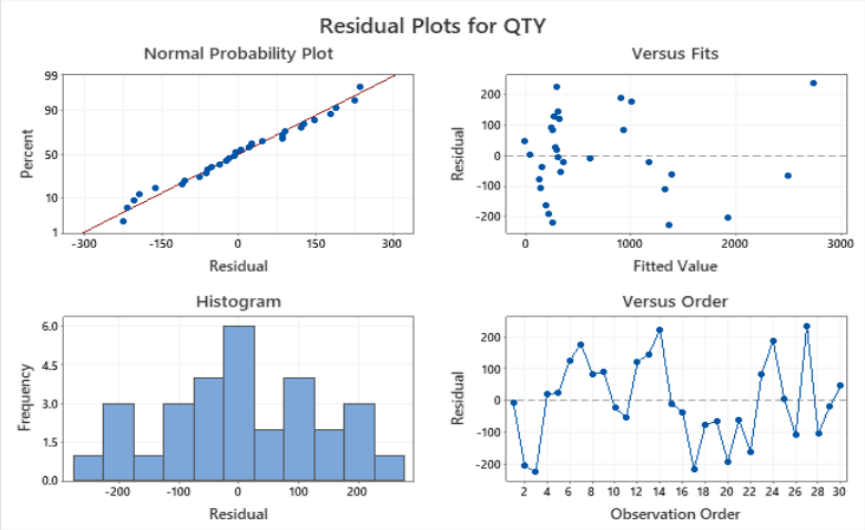


Figure A3 The result of SKU03 by Linear regression model

Stepwise Selection of Terms

α to enter = 0.05, α to remove = 0.05

Regression Equation

QTY = 18282 - 3466 lag_2_promo_period - 13.04 Store

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	18282	2580	7.09	0.000	
lag_2_promo_period	-3466	1257	-2.76	0.010	1.08
Store	-13.04	4.21	-3.10	0.005	1.08

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2652.37	46.52%	42.56%	35.34%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	165231741	82615871	11.74	0.000
lag_2_promo_period	1	53483790	53483790	7.60	0.010
Store	1	67490991	67490991	9.59	0.005
Error	27	189947380	7035088		
Total	29	355179121			

Fits and Diagnostics for Unusual Observations

Obs	QTY	Fit	Resid	Std Resid	
14	16752	10585	6167	2.38	R

R Large residual

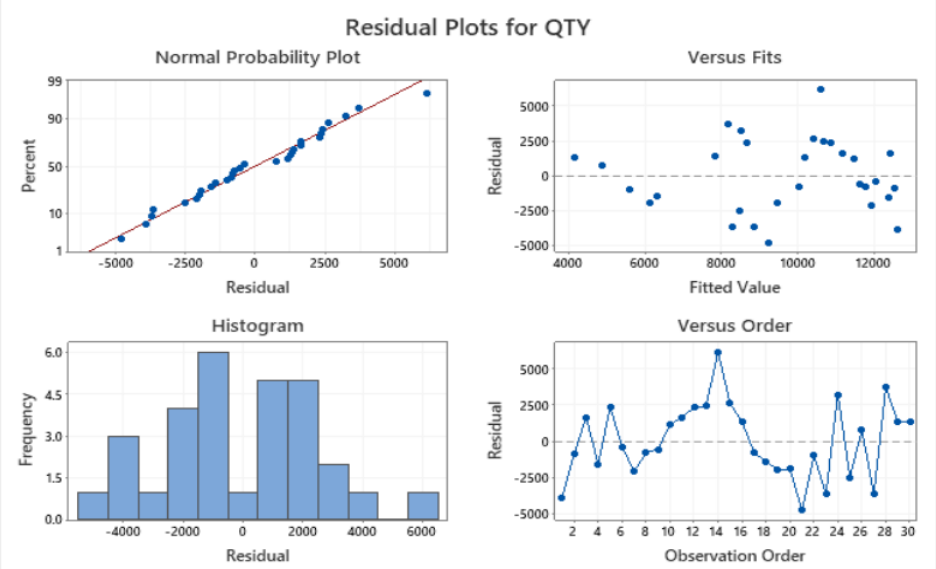


Figure A4 The result of SKU04 by Linear regression model

Stepwise Selection of Terms

α to enter = 0.05, α to remove = 0.05

Regression Equation

QTY = 19939 + 41330 promotion_period - 0.01616 Covid_19NewCases

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	19939	1775	11.23	0.000	
promotion_period	41330	3289	12.57	0.000	1.04
Covid_19NewCases	-0.01616	0.00655	-2.47	0.020	1.04

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
7487.38	85.40%	84.32%	80.54%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	8852645882	4426322941	78.96	0.000
promotion_period	1	8851586986	8851586986	157.89	0.000
Covid_19NewCases	1	341234932	341234932	6.09	0.020
Error	27	1513644030	56060890		
Total	29	10366289912			

Fits and Diagnostics for Unusual Observations

Obs	QTY	Fit	Resid	Std Resid	
1	44937	61269	-16332	-2.41	R
10	81237	61266	19971	2.95	R
28	54133	49716	4417	0.72	X

R Large residual
X Unusual X

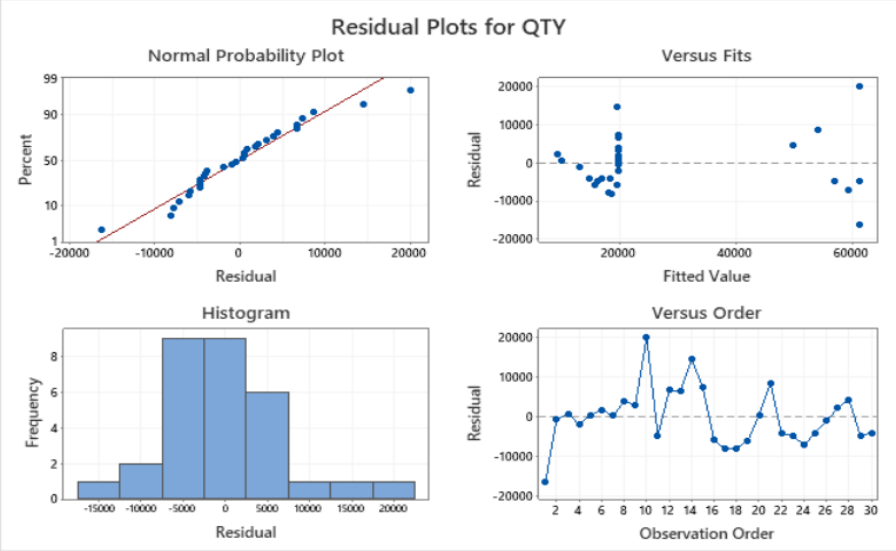


Figure A5 The result of SKU05 by Linear regression model

Stepwise Selection of Terms

α to enter = 0.05, α to remove = 0.05

Regression Equation

QTY = 15305 + 35846 promotion_period - 0.02949 Covid_19NewCases + 7300 Welfare

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	15305	1316	11.63	0.000	
promotion_period	35846	2082	17.22	0.000	1.00
Covid_19NewCases	-0.02949	0.00440	-6.70	0.000	1.27
Welfare	7300	1871	3.90	0.001	1.26

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4551.80	93.18%	92.40%	90.53%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	7362755555	2454251852	118.46	0.000
promotion_period	1	6141880229	6141880229	296.44	0.000
Covid_19NewCases	1	929850702	929850702	44.88	0.000
Welfare	1	315388308	315388308	15.22	0.001
Error	26	538690141	20718852		
Total	29	7901445695			

Fits and Diagnostics for Unusual Observations

Obs	QTY	Fit	Resid	Std Resid
14	31737	22245	9492	2.22 R
16	58578	50423	8155	2.00 R
17	47785	56116	-8331	-2.07 R

R Large residual

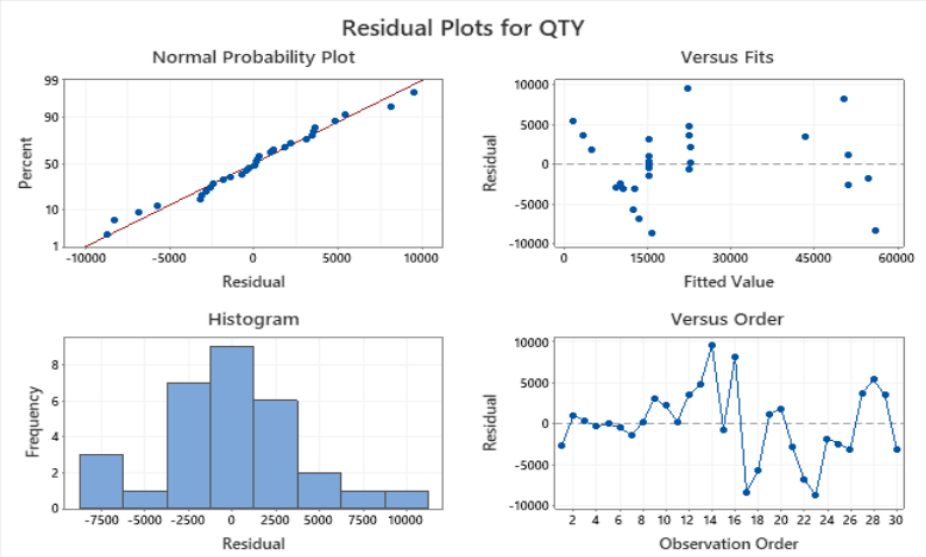


Figure A6 The result of SKU06 by Linear regression model

Stepwise Selection of Terms

α to enter = 0.05, α to remove = 0.05

Regression Equation

QTY = 13021 + 1940 percent_discount

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	13021	1097	11.87	0.000	
percent_discount	1940	161	12.03	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
5484.45	83.79%	83.22%	81.81%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	4355012651	4355012651	144.78	0.000
percent_discount	1	4355012651	4355012651	144.78	0.000
Error	28	842218115	30079218		
Total	29	5197230765			

Fits and Diagnostics for Unusual Observations

Obs	QTY	Fit	Resid	Std Resid	
1	38380	45350	-6970	-1.42	X
11	46361	45350	1011	0.21	X
14	26329	13021	13308	2.48	R
22	49027	45350	3677	0.75	X
24	46175	45350	825	0.17	X
28	46808	45350	1458	0.30	X

R Large residual
X Unusual X

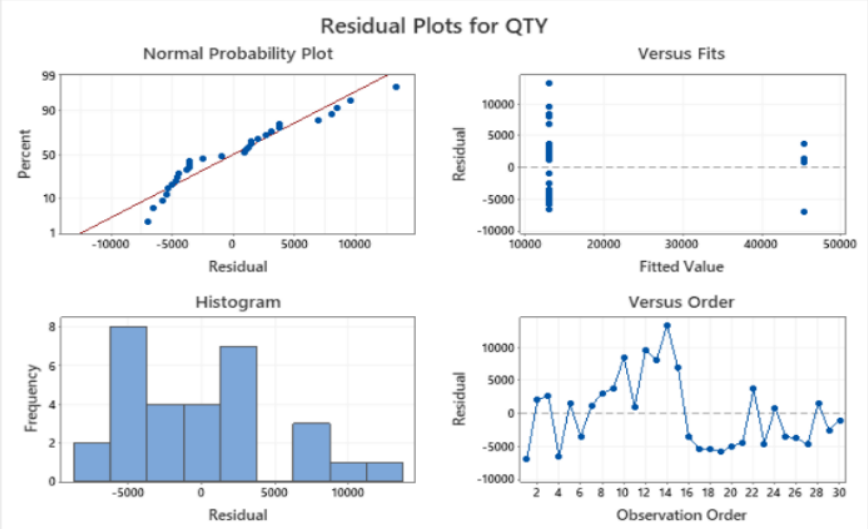


Figure A7 The result of SKU07 by Linear regression model

Stepwise Selection of Terms

α to enter = 0.05, α to remove = 0.05

Regression Equation

QTY = 8540 + 25640 promotion_period + 4522 month_9 - 0.01534 Covid_19NewCases + 2732 Welfare

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	8540	781	10.93	0.000	
promotion_period	25640	1397	18.35	0.000	1.40
month_9	4522	2036	2.22	0.036	1.03
Covid_19NewCases	-0.01534	0.00310	-4.95	0.000	1.73
Welfare	2732	1157	2.36	0.026	1.33

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2738.87	94.00%	93.03%	91.77%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	2935752414	733938104	97.84	0.000
promotion_period	1	2526398814	2526398814	336.79	0.000
month_9	1	37022760	37022760	4.94	0.036
Covid_19NewCases	1	183809944	183809944	24.50	0.000
Welfare	1	41818094	41818094	5.57	0.026
Error	25	187535530	7501421		
Total	29	3123287945			

Fits and Diagnostics for Unusual Observations

Obs	QTY	Fit	Resid	Std Resid	
9	11894	13061	-1167	-0.66	X
17	3835	10057	-6222	-2.38	R
21	35647	34480	1167	0.66	X
24	4127	9390	-5263	-2.00	R

R Large residual
X Unusual X

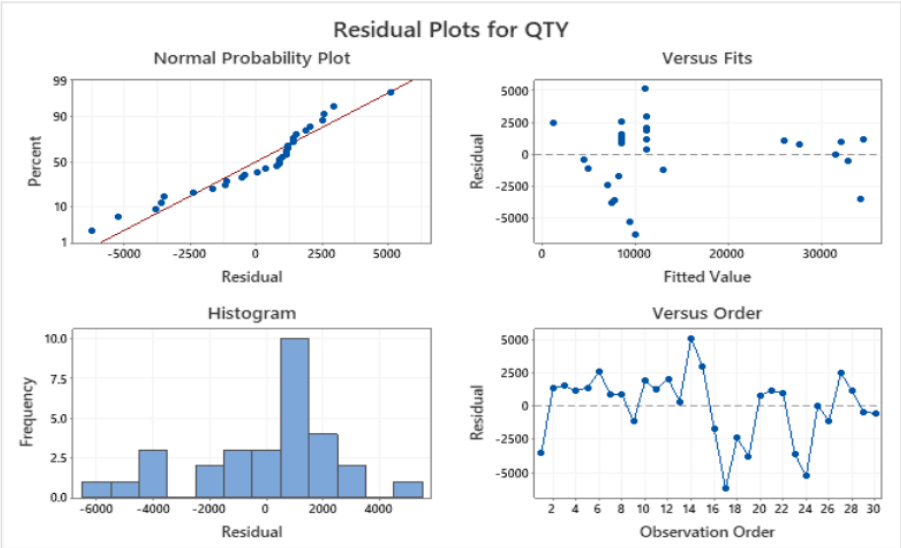


Figure A8 The result of SKU08 by Linear regression model

Stepwise Selection of Terms

α to enter = 0.05, α to remove = 0.05

Regression Equation

QTY = 9894 + 16591 promotion_period - 0.01217 Covid_19NewCases + 2831 Welfare

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	9894	811	12.20	0.000	
promotion_period	16591	1462	11.35	0.000	1.05
Covid_19NewCases	-0.01217	0.00288	-4.23	0.000	1.32
Welfare	2831	1199	2.36	0.026	1.26

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2915.47	83.73%	81.86%	76.52%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	1137573360	379191120	44.61	0.000
promotion_period	1	1094554853	1094554853	128.77	0.000
Covid_19NewCases	1	152127910	152127910	17.90	0.000
Welfare	1	47354286	47354286	5.57	0.026
Error	26	220999775	8499991		
Total	29	1358573135			

Fits and Diagnostics for Unusual Observations

Obs	QTY	Fit	Resid	Std Resid	
14	18105	12577	5528	2.01	R
17	5268	11761	-6493	-2.34	R
28	25327	20616	4711	2.03	R

R Large residual

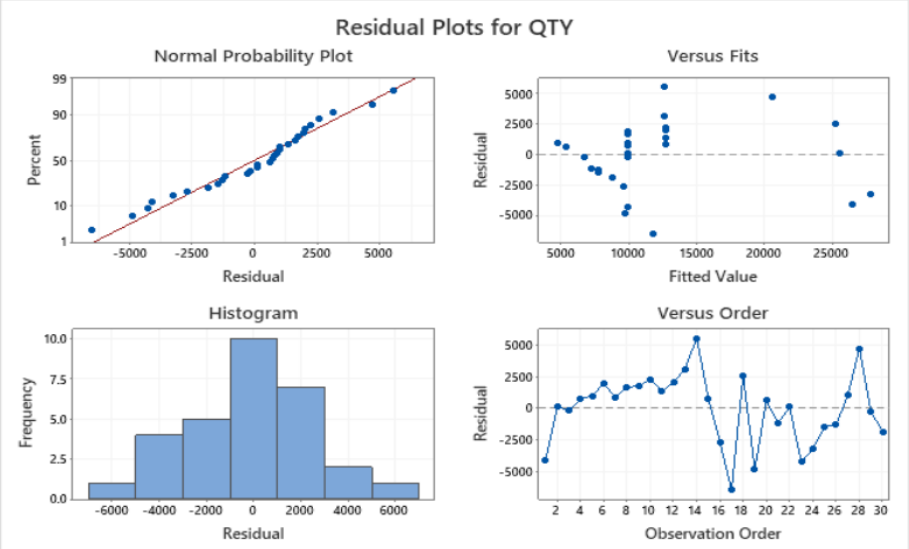


Figure A9 The result of SKU09 by Linear regression model

Stepwise Selection of Terms

α to enter = 0.05, α to remove = 0.05

Regression Equation

QTY = 9894 + 16591 promotion_period - 0.01217 Covid_19NewCases + 2831 Welfare

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	9894	811	12.20	0.000	
promotion_period	16591	1462	11.35	0.000	1.05
Covid_19NewCases	-0.01217	0.00288	-4.23	0.000	1.32
Welfare	2831	1199	2.36	0.026	1.26

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2915.47	83.73%	81.86%	76.52%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	1137573360	379191120	44.61	0.000
promotion_period	1	1094554853	1094554853	128.77	0.000
Covid_19NewCases	1	152127910	152127910	17.90	0.000
Welfare	1	47354286	47354286	5.57	0.026
Error	26	220999775	8499991		
Total	29	1358573135			

Fits and Diagnostics for Unusual Observations

Obs	QTY	Fit	Resid	Std Resid	
14	18105	12577	5528	2.01	R
17	5268	11761	-6493	-2.34	R
28	25327	20616	4711	2.03	R

R Large residual

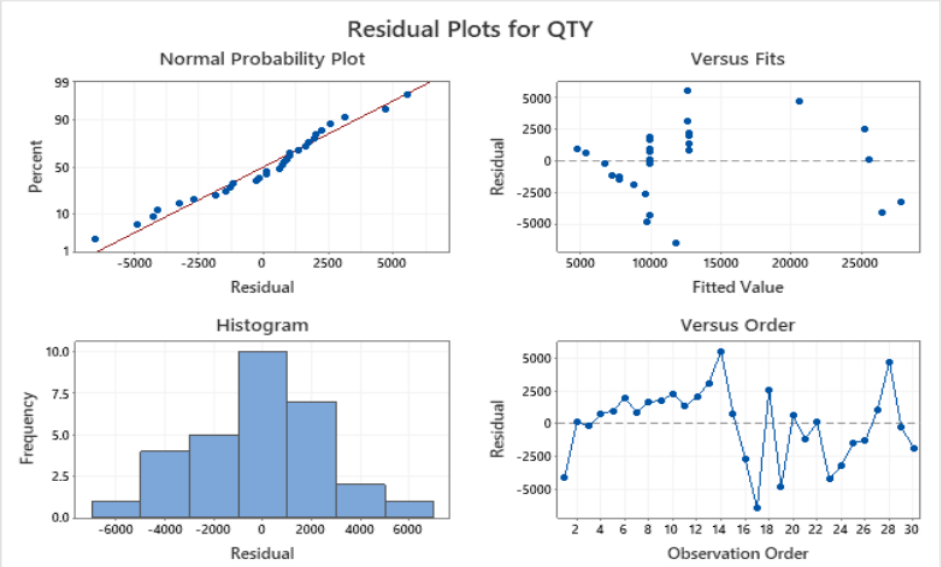


Figure A10 The result of SKU10 by Linear regression model

REFERENCES

Abolghasemi, M., Hurley, J., Eshragh, A., & Fahimnia, B. (2020). Demand forecasting in the presence of systematic events: Cases in capturing sales promotions. *International Journal of Production Economics*, *230*, 107892. https://doi.org/https://doi.org/10.1016/j.ijpe.2020.107892

Aburto, L., & Weber, R. (2007). Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing*, *7*(1), 136-144. https://doi.org/https://doi.org/10.1016/j.asoc.2005.06.001

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717-727. https://doi.org/https://doi.org/10.1016/S0731-7085(99)00272-1

Ahmed, H. U., Mohammed, A. A., & Mohammed, A. (2022). Soft computing models to predict the compressive strength of GGBS/FA-geopolymer concrete. *PloS one*, *17*(5), e0265846.

Ali, Ö. G., Sayın, S., van Woensel, T., & Fransoo, J. (2009). SKU demand forecasting in the presence of promotions. *Expert Systems with Applications*, *36*(10), 12340-12348. https://doi.org/https://doi.org/10.1016/j.eswa.2009.04.052

Almeida, F. M. D., Martins, A. M., Nunes, M. A., & Bezerra, L. C. T. (2022, 15-17 June 2022). Retail sales forecasting for a Brazilian supermarket chain: an empirical assessment. 2022 IEEE 24th Conference on Business Informatics (CBI),

Amiri, S. S., Mueller, M., & Hoque, S. (2023). Investigating the application of a commercial and residential energy consumption prediction model for urban Planning scenarios with Machine Learning and Shapley Additive explanation methods. *Energy and Buildings*, *287*, 112965. https://doi.org/https://doi.org/10.1016/j.enbuild.2023.112965

Archer, B. (1987). Demand forecasting and estimation. *Demand forecasting and estimation.*, 77-85.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*.

Arunraj, N. S., & Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal of Production Economics*, *170*, 321-335. https://doi.org/https://doi.org/10.1016/j.ijpe.2015.09.039

Auppakorn, C., & Phumchusri, N. (2022). *Daily Sales Forecasting for Variable-Priced Items in Retail Business* 2022 4th International Conference on Management Science and Industrial Engineering,

Aye, G. C., Balcilar, M., Gupta, R., & Majumdar, A. (2015). Forecasting aggregate retail sales: The case of South Africa. *International Journal of Production Economics*, *160*, 66-79. https://doi.org/https://doi.org/10.1016/j.ijpe.2014.09.033

Ayers, J. B., & Odegaard, M. A. (2017). *Retail supply chain management*. CRC Press. https://books.google.co.th/books?hl=th&lr=&id=TXJQDwAAQBAJ&oi=fnd&pg=PP1&dq=retail+business+role+in+supply+chain+management+pdf&ots=ZITC1iq6KV&sig=UiRM9PxEz9uobncw1EO6iBSBL8o&redir_esc=y#v=onepage&q&f=false

Azadi, M., Yousefi, S., Saen, R. F., Shabanpour, H., & Jabeen, F. (2023). Forecasting sustainability of healthcare supply chains using deep learning and network data envelopment analysis. *Journal of Business Research*, *154*, 113357.

Badorf, F., & Hoberg, K. (2020). The impact of daily weather on retail sales: An empirical study in brick-and-mortar stores. *Journal of Retailing and Consumer Services*, *52*, 101921. https://doi.org/https://doi.org/10.1016/j.jretconser.2019.101921

Berrar, D. (2019). Cross-Validation. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *1–3*, 542-545.

Biau, G., & Scornet, E. (2015). A Random Forest Guided Tour. *TEST*, *25*. https://doi.org/10.1007/s11749-016-0481-7

Boukadida, H., Hassen, N., Gafsi, Z., & Besbes, K. (2011). A HIGHLY TIME-EFFICIENT DIGITAL MULTIPLIER BASED ON THE A2 BINARY REPRESENTATION. *International Journal of Engineering Science and Technology*, *3*.

Boyapati, S. N., & Mummidi, R. (2020). Predicting sales using Machine Learning Techniques. In.

Brownlee, J. (2018). What is the Difference Between a Batch and an Epoch in a Neural Network. *Machine Learning Mastery*, *20*.

Carter, M. (2019). Competition and Profit Margins in the Retail Tฟrade Sector | Bulletin – June Quarter 2019. RBA Bulletin,

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS international journal of geo-information*, *7*(5), 168.

Chiewpanich, T., & Mokkhamakkul, T. (2019). Forecasting Model for Promotional Sales. *Journal of Applied Research on Science and Technology (JARST)*, *18*(2), 31-40.

Chung, D., Lee, K., & Baek, Y. (2023). New Product Demand Forecasting Using Hybrid Machine Learning: A Combined Model of K-Means, Ann, and Qrnn. *SSRN Electronic Journal*.

Development, D. o. C. P. a. U. (2021). สถิติข้อมูลการค้าปลีกในเขตกรุงเทพมหานครปี พ.ศ. 2563.

E, E., Yu, M., Tian, X., & Tao, Y. (2022). Dynamic Model Selection Based on Demand Pattern Classification in Retail Sales Forecasting. *Mathematics*, *10*, 3179. https://doi.org/10.3390/math10173179

EIC, E. I. C. (2022). *Retail Trade 2022*. Retrieved 10 January 2023 from https://www.scbeic.com/th/detail/file/product/8348/gaygdow7yx/Industry-Insight_Retail-trade-2022_20220622.pdf

Ensafi, Y., Amin, S. H., Zhang, G., & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. *International Journal of Information Management Data Insights*, *2*(1), 100058. https://doi.org/https://doi.org/10.1016/j.jjimei.2022.100058

Fadillah, A. D., Wantuah, A. F., Surya, M. H. B., Azka, M. Z. A., & Nurprawito, D. (2022). Improving Inventory Management by Better Forecasting Method for Healthcare Industry Company. *European Conference on Industrial Engineering and Operations Management*

Falatouri, T., Darbanian, F., Brandtner, P., & Udokwu, C. (2022). Predictive Analytics for Demand Forecasting – A Comparison of SARIMA and LSTM in Retail SCM. *Procedia Computer Science*, *200*, 993-1003. https://doi.org/https://doi.org/10.1016/j.procs.2022.01.298

Fox, E., & Sethuraman, R. (2006). Retail Competition. In (pp. 193-208). https://doi.org/10.1007/3-540-28433-8_13

García, V., Sánchez, J., & Marqués, A. (2019). Synergetic Application of Multi-Criteria Decision-Making Models to Credit Granting Decision Problems. *Applied Sciences*, *9*, 1-15. https://doi.org/10.3390/app9235052

Ghafari, E., Bandarabadi, M., Costa, H., & Júlio, E. (2014). *Full Paper-BMC 2012-EG-04-1 revEJ*.

Ghalehkhondabi, I., Ardjmand, E., Weckman, G. R., & Young, W. A. (2017). An overview of energy demand forecasting methods published in 2005–2015. *Energy Systems*, *8*, 411-447.

Güven, İ., & Şimşir, F. (2020). Demand forecasting with color parameter in retail apparel industry using artificial neural networks (ANN) and support vector machines (SVM) methods. *Computers & Industrial Engineering*, *147*, 106678. https://doi.org/https://doi.org/10.1016/j.cie.2020.106678

Hajirahimi, Z., & Khashei, M. (2019). Hybrid structures in time series modeling and forecasting: A review. *Engineering Applications of Artificial Intelligence*, *86*, 83-106. https://doi.org/https://doi.org/10.1016/j.engappai.2019.08.018

Hamzaçebi, C. (2008). Improving artificial neural networks' performance in seasonal time series forecasting. *Information Sciences*, *178*(23), 4550-4559.

Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting*, *36*(4), 1420-1438. https://doi.org/https://doi.org/10.1016/j.ijforecast.2020.02.005

Jain, A., Menon, M. N., & Chandra, S. (2015). Sales forecasting for retail chains. *San Diego, California: UC San Diego Jacobs School of Engineering*.

Jones, S. S., Evans, R. S., Allen, T. L., Thomas, A., Haug, P. J., Welch, S. J., & Snow, G. L. (2009). A multivariate time series approach to modeling and forecasting demand in the emergency department. *Journal of Biomedical Informatics*, *42*(1), 123-139. https://doi.org/https://doi.org/10.1016/j.jbi.2008.05.003

Kadam, R., & Lingras, P. (2023). Understanding Sales Patterns using Unsupervised Machine Learning. *Symposium on AI, Data and Digitalization (SAIDD 2023), 42)*.

Khan, M. Z., Yousuf, R. I., Shoaib, M. H., Ahmed, F. R., Saleem, M. T., Siddiqui, F., & Rizvi, S. A. (2023). A hybrid framework of artificial intelligence-based neural network model (ANN) and central composite design (CCD) in quality by design formulation development of orodispersible moxifloxacin tablets: Physicochemical evaluation, compaction analysis, and its in-silico PBPK modeling. *Journal of Drug Delivery Science and Technology*, *82*, 104323. https://doi.org/https://doi.org/10.1016/j.jddst.2023.104323

Khan, S. M. (2020). Forecasting of sales data using support vector regression.

Kiran, J. S., Rao, P. S. V. S., Rao, P. V. R. D. P., Babu, B. S., & Divya, N. (2022, 25-27 Jan. 2022). Analysis on the Prediction of Sales using Various Machine Learning Testing Algorithms. 2022 International Conference on Computer Communication and Informatics (ICCCI),

KResearch, K. R. C. (2022). ค้าปลีกปี' *66* ... ขยายตัวต่อเนื่อง ท่ามกลางปัจจัยท้าทายรอบด้าน (กระแสทรรศน์ ฉบับที่ *3377)*. Retrieved 10 January 2023 from https://www.krungsri.com/th/research/industry/industry-outlook/wholesale-retail/modern-trade/io

Laaroussi, H., Guerouate, F., & Sbihi, M. (2023). A novel hybrid deep learning approach for tourism demand forecasting. *International Journal of Electrical and Computer Engineering*, *13*(2), 1989.

Lahouar, A., & Ben Hadj Slama, J. (2015). Day-ahead load forecast using random forest and expert input selection. *Energy Conversion and Management*, *103*, 1040-1051. https://doi.org/https://doi.org/10.1016/j.enconman.2015.07.041

Lek, S., & Park, Y. S. (2008). Artificial Neural Networks. In S. E. Jørgensen & B. D. Fath (Eds.), *Encyclopedia of Ecology* (pp. 237-245). Academic Press. https://doi.org/https://doi.org/10.1016/B978-008045405-4.00173-7

Loureiro, A. L. D., Miguéis, V. L., & da Silva, L. F. M. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*, *114*, 81-93. https://doi.org/https://doi.org/10.1016/j.dss.2018.08.010

Lu, C.-J., Lee, T.-S., & Lian, C.-M. (2012). Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks. *Decision Support Systems*, *54*(1), 584-596. https://doi.org/https://doi.org/10.1016/j.dss.2012.08.006

Ma, S., & Fildes, R. (2021). Retail sales forecasting with meta-learning. *European Journal of Operational Research*, *288*(1), 111-128. https://doi.org/https://doi.org/10.1016/j.ejor.2020.05.038

Mahesh, B. (2019). *Machine Learning Algorithms -A Review*. https://doi.org/10.21275/ART20203995

Mitra, A., Jain, A., Kishore, A., & Kumar, P. (2022). A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company: A Novel Hybrid Machine Learning Approach. *Operations Research Forum*, *3*(4), 58. https://doi.org/10.1007/s43069-022-00166-4

Montaño, J., Palmer, A., Sesé, A., & Cajal, B. (2013). Using the R-MAPE index as a resistant measure of forecast accuracy. *Psicothema*, *25*, 500-506. https://doi.org/10.7334/psicothema2013.23

Moon, K. S., Lee, H. W., Kim, H. J., Kim, H., Kang, J., & Paik, W. C. (2022). Forecasting Obsolescence of Components by Using a Clustering-Based Hybrid Machine-Learning Algorithm. *Sensors (Basel)*, *22*(9). https://doi.org/10.3390/s22093244

Mustika, W. F., Murfi, H., & Widyaningsih, Y. (2019 2019). Analysis accuracy of xgboost model for multiclass classification-a case study of applicant level risk prediction for life insurance. 2019 5th International Conference on Science in Information Technology (ICSITech),

Nenni, M. E., Giustiniano, L., & Pirolo, L. (2013). Demand forecasting in the fashion industry: a review. *International Journal of Engineering Business Management*, *5*(Godište 2013), 5-36.

Nguyen, D. T., Wang, C. N., Dang, T. T., Ming-Hsien, H., & Do, N. H. (2023). Enhancement of Sales Forecast Using Hybrid Sarima and Extreme Machine Learning: A Case for a Jewelry Retailer in Viet Nam. *Available at SSRN 4330707*.

Nunnari, G., & Nunnari, V. (2017, 24-27 July 2017). Forecasting Monthly Sales Retail Time Series: A Case Study. 2017 IEEE 19th Conference on Business Informatics (CBI),

Parhi, R., & Nowak, R. D. (2020). The Role of Neural Network Activation Functions. *IEEE Signal Processing Letters*, *27*, 1779-1783. https://doi.org/10.1109/LSP.2020.3027517

Prabhakar, V., Sayiner, D., Chakraborty, U., Nguyen, T., & Lanham, M. (2018). Demand Forecasting for a large grocery chain in Ecuador. *Data. Published*.

Pratama, K., & Kang, D.-K. (2021). Trainable activation function with differentiable negative side and adaptable rectified point. *Applied Intelligence*, *51*(3), 1784-1801. https://doi.org/10.1007/s10489-020-01885-z

Priyadarshi, R., Panigrahi, A., Routroy, S., & Garg, G. K. (2019). Demand forecasting at retail stage for selected vegetables: a performance analysis. *Journal of Modelling in Management*, *14*(4), 1042-1063. https://doi.org/10.1108/JM2-11-2018-0192

Punia, S., Nikolopoulos, K., Singh, S. P., Madaan, J., & Litsiou, K. (2020). Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. *International Journal of Production Research*, *58*, 1-16. https://doi.org/10.1080/00207543.2020.1735666

Punia, S., & Shankar, S. (2022). Predictive analytics for demand forecasting: A deep learning-based decision support system. *Knowledge-Based Systems*, *258*, 109956. https://doi.org/https://doi.org/10.1016/j.knosys.2022.109956

Saha, P., Gudheniya, N., Mitra, R., Das, D., Narayana, S., & Tiwari, M. K. (2022). Demand Forecasting of a Multinational Retail Company using Deep Learning Frameworks. *IFAC-PapersOnLine*, *55*(10), 395-399. https://doi.org/https://doi.org/10.1016/j.ifacol.2022.09.425

Samang, P. (2020). *Forecasting of Thailand major petroleum product consumption using machine learning techniques* https://digital.car.chula.ac.th/chulaetd/3827

Singh, P. K., Gupta, Y., Jha, N., & Rajan, A. (2019). Fashion retail: Forecasting demand for new items. *arXiv preprint arXiv:1907.01960*.

Swami, D., Shah, A. D., & Ray, S. K. B. (2020). Predicting Future Sales of Retail Products using Machine Learning. *arXiv preprint arXiv:2008.07779*.

Tekin, A., & Sarı, C. (2022). Store-based Demand Forecasting of a Company via Ensemble Learning. In (pp. 14-23). https://doi.org/10.1007/978-3-031-09176-6_2

Thoplan, R. (2014). Qualitative v/s Quantitative Forecasting of Yearly Tourist Arrival in Mauritius. *International Journal of Statistics and Applications*, *4*, 198-203. https://doi.org/10.5923/j.statistics.20140404.04

Tian, X., Cao, S., & Song, Y. (2021). The impact of weather on consumer behavior and retail performance: Evidence from a convenience store chain in China. *Journal of Retailing and Consumer Services*, *62*, 102583. https://doi.org/https://doi.org/10.1016/j.jretconser.2021.102583

Wang, J. (2020, 4-6 Dec. 2020). A hybrid machine learning model for sales prediction. 2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI),

Wang, W., Chakraborty, G., & Chakraborty, B. (2020). Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm. *Applied Sciences*, *11*, 202. https://doi.org/10.3390/app11010202

Wen, X., Choi, T.-M., & Chung, S.-H. (2019). Fashion retail supply chain management: A review of operational models. *International Journal of Production Economics*, *207*, 34-55. https://doi.org/https://doi.org/10.1016/j.ijpe.2018.10.012

Wijaya, D., Prayogo, D., Santoso, D. I., Gunawan, T., & Widjaja, J. A. (2020). Optimizing Prediction Accuracy of Concrete Mixture Behavior Using Hybrid K-means Clustering and Ensemble Machine Learning. *Journal of Physics: Conference Series*, *1625*, 012022.

Witten, D., & James, G. (2013). *An introduction to statistical learning with applications in R*. springer publication.

Yang, C.-L., & Nguyen, T. (2022). Sequential Clustering and Classification Approach to Analyze Sales Performance of Retail Stores Based on Point-of-Sale Data. *International Journal of Information Technology & Decision Making*, *21*, 1-26. https://doi.org/10.1142/S0219622022500079

Yoon, D., Park, S., Song, Y., Chae, J., & Chung, D. (2023). Methodology for Improving the Performance of Demand Forecasting Through Machine Learning. *International Journal of Logistics Research and Applications*, *25(4):1-24*. https://doi.org/10.1080/13675567.2020.1803246

VITA

NAME	Nichakan Phupaichitkun
DATE OF BIRTH	10 October 1999
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Chulalongkorn University
HOME ADDRESS	Bangkok, Thailand