Leading Indicators Selection and Forecasting Packaging Consumption in Thailand

Pawitra Nopsuwan and Ittipong Khemapech*

Department of Computer Science, School of Science and Technology, University of the Thai Chamber of Commerce, Bangkok, 10400, Thailand

*Corresponding author. E-mail: ittipong_khe@utcc.ac.th

ABSTRACT

Packaging is one of the essential factors in the production and selling process. Knowing the leading indicators and predicting the demand of consumer goods that use a high volume of packaging can help packaging manufacturers plan their production to satisfy consumers' needs and plan production costs effectively. In this research, we present the economic factors that affect the demand for consumer goods in Thailand and predict the production for each product using machine learning approaches. We use model-based selection, f-regression, mutual information, and recursive feature elimination (RFE) for the feature selection. Moreover, for forecasting, we compare the performance of machine learning models that can describe algorithms inside included multiple linear regression, random forest regression, and gradient boosting. The results show that using random forest regression with model-based feature selection gives the best R square score for 29 out of 47 products or more than 60% of the total products.

Keywords: *Feature Selection, Regression Problem, Demand Forecast, Packaging Consumption*

INTRODUCTION

Thailand is one of the countries that produce products for domestic use and export to many countries worldwide. One of the essentials in the process of selling and exporting goods is the packaging. Many products cannot be sold and transported without packaging. Therefore, knowing the factors affecting the demand for each product and forecasting will help the packaging manufacturer plan the production to produce enough to meet customers' needs. Moreover, the production cost can be effectively controlled and generate the most benefit. The 2020 annual report of one of the major packaging manufacturers in Thailand shows that food, beverages, fast-moving consumer goods (FMCG), and electrical appliances consume around 70% of total packaging demand. Therefore, this report focuses on those products.

Predicting packaging demand is now possible with experts' knowledge and information from published documents from various relevant agencies. They have usually used known factors from historical data. However, factors affecting the demand for some products may change over time and using only the previously known factors may result in weak forecasting results. Finding other factors that

Article history:
Received 27 February 2022; Received in revised from 20 May 2022;
Accepted 7 June 2022; Available online 25 June 2022.

affect the demand for each product helps us identify factors that were not previously known, and it also helps to make forecasts more accurate.

This research aims to determine the economic factors affecting the products that use a high amount of packaging in Thailand, namely food, beverages, consumer goods, and electrical appliances. Furthermore, demand forecast six months in advance using machine learning. We use the python package from scikit-learn to select factors and forecasts. For forecasting, we use a machine learning model that can explain internal operations to know the relationship between product demand and economic factors. The first selected model in this study was the multiple linear regression model, which provides a straightforward linear relationship. Moreover, we selected ensemble models that included random forest regression and gradient boosting are explainable models and can also extract internal logic. We used regression models together with the various features selection methods, SelectFromModel, SelectKBest (f_regression and mutual_info_regression), and Recursive Feature Elimination (RFE). We finally select the machine learning model and feature selection method that the most appropriate for each product based on mean absolute error (MAE) and coefficient of determination (R square).

In the next section, we will talk about the relevant research in section 2, which discusses the research we have studied to develop and apply what we have learned in our study. And then, we will show the methods we used in this study in section 3, including data gathering, data processing, feature selection, prediction models, and performance measurement metrics that we use in our study. Finally, we will show the results of our experiment in section 4, our conclusion about our study, and our future work that we plan to improve in section 5.

LITERATURE REVIEW

From researching relevant research focusing on feature selection and demand forecasting, recursive feature elimination (RFE) is the most popular and effective method for feature selection. Research by Xing-Ling Wang, Xue-Lian Wu, and Bing-Yu Sun (2012) experimented with using RFE to select essential features and found that RFE resulted in more accurate prediction results. In their research, only feature selection from RFE was used in conjunction with a support vector machine (SVM). The feature selection by other methods and other types of machine learning models has not been compared. In a study by Sofia D'Souza, Prema K.V., and Balaji S (2020), RFE was combined with statistical methods including stepwise, best subsets, and lesso. It was found that RFE combined with the best subsets yielded the best results.

In addition, the most famous scores for feature selection, the mutual information score Muhammad Aliyu Sulaiman and Jane Labadie (2015), focused solely on the mutual information score in feature selection, arguing that it was a method that could be used with a feature. However, their research has yet to compare them with other feature selections and comparing results between multiple machine learning models remains their future roadmap.

As for demand forecasting, many studies have avoided the use of a black box machine learning model. The most popular models for solving demand forecasting are multiple linear regression, and tree-based models namely decision tree, random forest regression, and gradient boosting. The Selcuk Cankurt research (2016) said they chose to use an ensemble of regression trees for demand forecasting because regression trees allow us to see how we think within the model. They compared each ensembling technique, which bagging showed the best results in their research. In addition, research by Jitendra Kumar Jaiswal and Rita Samikannu (2017) used only random forest regression, using random forest regression for both feature selection and prediction.

Many have chosen to compare the results of multiple machine learning models to see which model can best solve their problems, research by Sankalp Jain and Naresh Kumar Nagwani (2017), research by Narayana Darapaneni, Sreelakshminarayanan Muthuraj, Prabakar K and Madhavan Sridhar (2019), and research by YouLi Feng, ShanShan Wang (2017) used similar methods. They compared the effect of multiple linear regression and random forest regression. Their results, random forest regression, provided more accurate predictions than multiple linear regression when used to solve their problems.

Some studies have used machine learning models as black boxes in forecasting; for the most part, understanding how the model works is unnecessary. Their main aim is to find the model that produces the best results and compare the resources used by each model to select the model best suited to their problem. Neeraj Kumar, and M.M. Tripathi (2017) used only the neural network for forecasting found that the results were not very good. The forecast results for some months are accurate. Nevertheless, some months have a very high error, and they plan to improve by using other methods in the future. In the research by Engin Ilseven, and Murat Gol (2019), various machine learning models were compared, with their results concluding that the use of neural networks produced relatively better results than other models. At the same time, neural networks use more resources than other models as well. Therefore, choosing a suitable model for the job may require consideration of both accuracy and resource utilization.

However, we have not found research on finding leading indicators and forecasting the demand for packaging or product groups of interest. Nevertheless, it is helpful to study the relevant research and give us the idea to be applied in our study.

METHODOLOGY

After identifying the problem, having clear objectives, and reviewing related research, we designed an experimental approach to solving the problem, consisting of four main steps: data gathering, data processing, feature selection, prediction models, and performance measurement. Each step is detailed as follows:

Data Gathering

The data used in this research include the production data of each industry and economic factors. The two parts of the data have a monthly frequency. For monthly production data of each industry group, we collect information from Thailand's Office of Industrial Economics website. By downloading in excel format, the data is from 2000 to the present.

For economic factors, we take from a data provider website that provides economic data via API services. We manually select the inputs from Thailand and countries that import products from Thailand: China, Japan, Malaysia, Singapore, South Korea, Vietnam, Indonesia, France, Germany, UK, and Australia, including ASEAN and European Union countries by manual method with a total of 1279 factors.

Data Processing

Before using the data, we have data preparation steps to prepare the data in a suitable format for efficient processing. We start by removing the season from both production data and economic factors data. Because when there are no seasonal factors, we can more effectively determine the factors that affect the production volume. We use the Python package to remove the seasons from each column of data. The result of seasonal decomposition from the Python package included trend, seasonal, and residual components, as shown in Fig. 1.

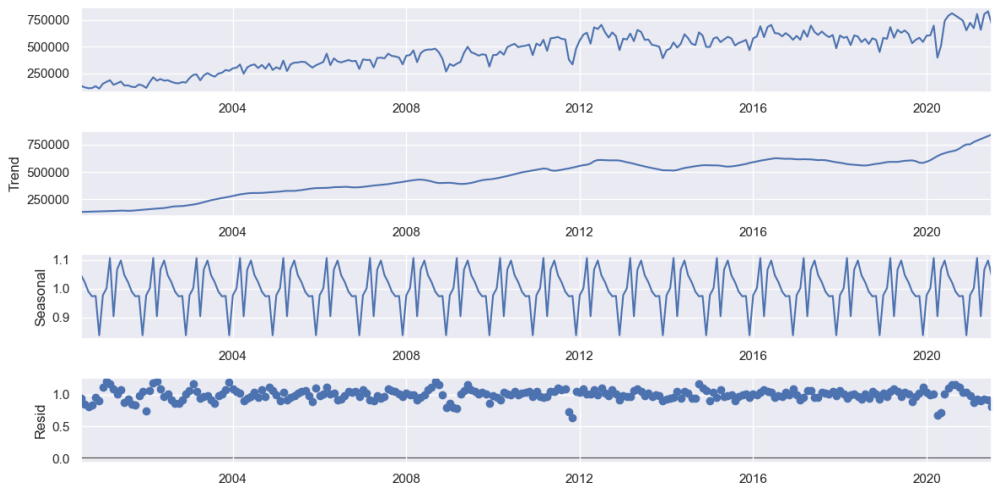

Figure 1 Example of seasonal decomposition.

After that, we merge production data and economic factors with the date as the key, then shift the economic data according to the number of months we want to find the leading indicator (1-6 months).

Moreover, some features contain large amounts of missing values or duplicate data. We need to remove those features because those features may lead to inaccurate predictions or not be helpful for our prediction. Then we set the criteria to remove the features with more than 10% missing or unique values. After that, we divide each dataset into two subsets for training and testing with a ratio of 80% and 20%.

Finally, since our data has different units, the scale of each feature also varies. Differences in the scale of each feature make different weights for each feature and lead to wrong feature selection and inaccurate predictions. So, we normalize all the data to equalize the scaling of all features. In this study, we choose a robust scaler, the outlier-tolerant method, to normalize the data because both production and economic data are prone to outlier data. We set the quantile range to (25, 75). The equation of robust scaler is shown in (1).

$$x_{scale} = \frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)} \quad (1)$$

Feature Selection

For features selection methods, we experimented with several methods to achieve subsets of different features using the feature selection package from scikit-learn: SelectFromModel, SelectKBest, and recursive feature elimination (RFE), each with its details. details as follows:

SelectFromModel: This method is a meta-transformer. The feature selection is based on the model used. The feature selection score is obtained from the used model, i.e., estimated coefficients (coef_) or feature importance after fitting the model. The working process of SelectFromModel is shown in Fig. 2.

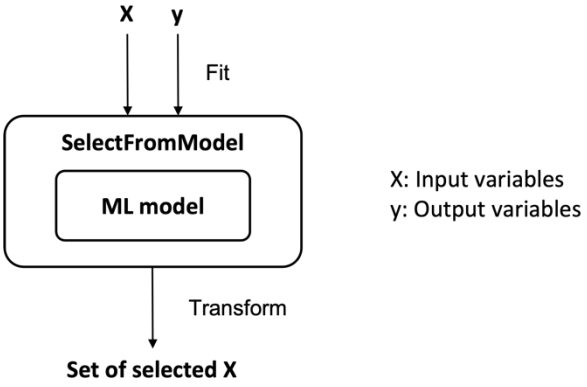

Figure 2 Workflow of SelectFromModel.

SelectKBest: This univariate feature selection works by selecting the best features based on univariate statistical tests, and it results in the feature with the top

k scores. The scores used in feature selections in this research are F-value and mutual information, and in addition, to find the optimal number of k. This research finds the optimal number of features by looping from 1 to the total number of features. The working process of SelectKBest is shown in Fig. 3.

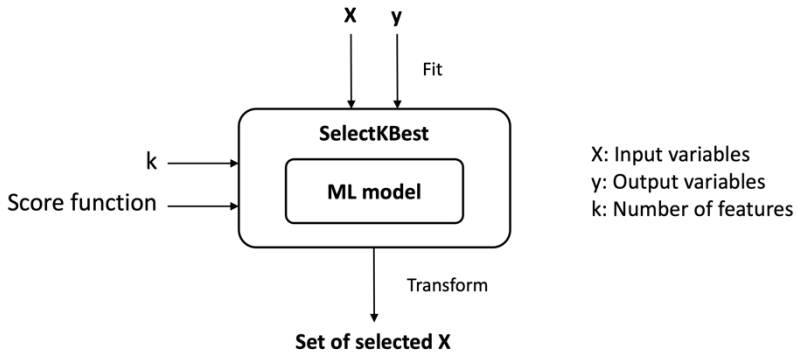

Figure 3 Workflow of SelectKBest.

Recursive feature elimination (RFE): This backward feature selection method works by training the model with all its features, gradually removing one feature with the lowest score, and repeating the trained model until the last feature. The results are the order of features by feature score. We show the algorithm of RFE in Figure 3. Additionally, this research used recursive feature elimination cross-validation (RFECV) to determine the optimal number of features. The working process of RFE is shown in Fig. 4.

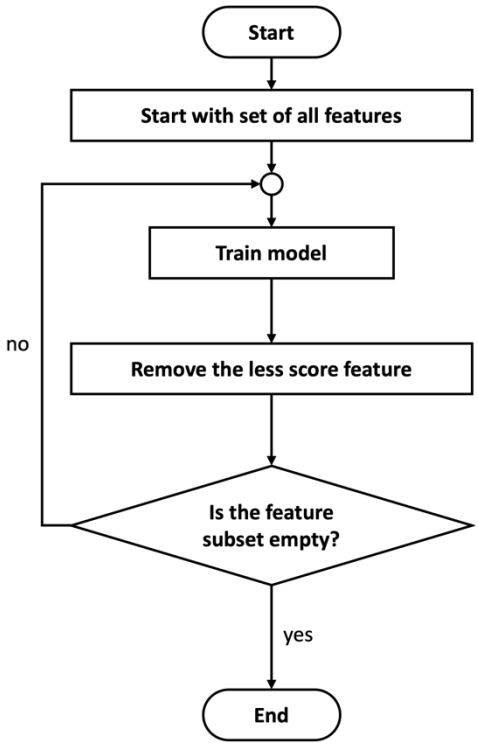

Figure 4 Workflow of RFE.

The methods mentioned above show that each method has a different workflow and feature selection criteria. We assume that the set of features selected by each method may differ. Each product of interest may be suitable for different feature selection methods, and the set of features selected by each method leads to different demand prediction accuracy.

Prediction Models

In this research, we compare the predictive performance of several models to find the most suitable model for forecasting the demand for each product. The models included multiple linear regression, random forest regression, and gradient boosting.

Multiple Linear Regression: Multiple linear regression, as in (2) is a statistical technique that uses explanatory variables to predict the outcome of response variables by finding the linear relationship between the explanatory and dependent variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \quad (2)$$

Random Forest Regression: Random forest regression is an algorithm that uses a bagging ensembling learning approach for regression problems. Within a random forest regression consists of several decision trees generated from datasets sampled from the training dataset. Results are predicted based on the mean of the answer trees, and this approach reduces overfitting and allows for more accurate results. We show the algorithm of random forest regression in Fig 5.

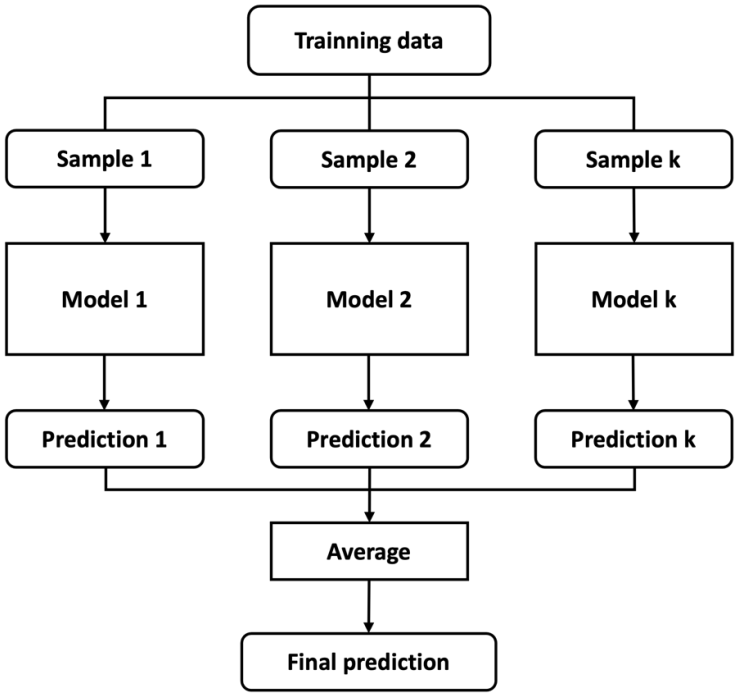

Figure 5 Random Forest Regression.

Gradient Boosting: Gradient boosting is an algorithm that uses a boosting ensembling learning approach for regression problems. It consists of several models working together in sequence. Each model takes the strengths and weaknesses of the previous model and improves it for more accurate results. That is to say, the model in the final sequence is the model that collects all the advantages and disadvantages of the previous model. We show the algorithm of gradient boosting in Fig 6.

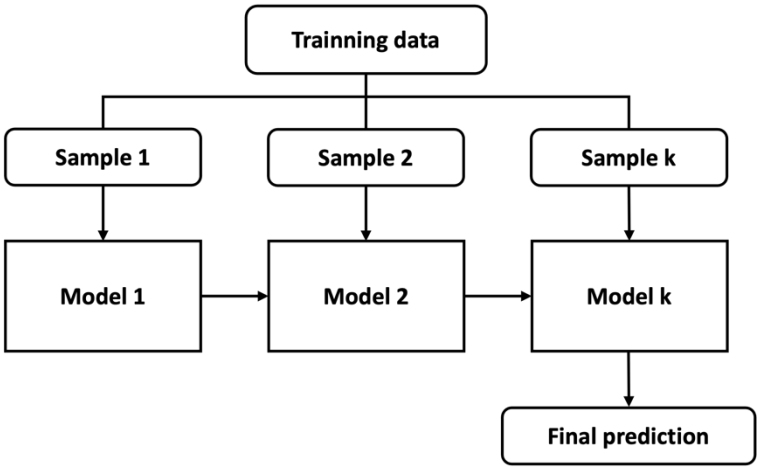

Figure 6 Gradient Boosting.

Performance Measurement

In this research, we use model performance measurement methods to compare and choose which model can best predict the needs of each product. The mean absolute error (MAE) and coefficient of determination (R square) are the measurement units considered.

Mean Absolute Error (MAE): Mean absolute error as in (3) is a measure that shows absolute error (absolute of residual) between actual and prediction values in the test dataset.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \quad (3)$$

Coefficient of Determination (R square): R square as in (4) is a measure that represents the proportion of the variance in the dependent variable, which the regression model explains. R square is in the range 0 to 1, where 1 means the model can best describe the relationship between explanatory and dependent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \quad (4)$$

RESULTS

As an experiment, we used the feature selection and forecasting methods described in the methodology section and measured the results using mean absolute error and R square. We have found that each product has different factors that affect demand, even within the same industry. The results show that the most suitable model for forecasting the demand of each product and the number of factors that

affect it differs. We found that ensembling models included random forest regression and gradient boosting are outstanding in forecasting the products of our interest, as 46 out of 47 products are the best results compared to multiple linear regression, and random forest regression is best for more than 60% of all products we are interested in. Tables 1 to 4 show the best method for using the R square score as a metric and the number of leading indicators affecting each product by industry.

Table 1 The best method and results of food products.

Product	Best Model	Feature selection method	Number of selected features	R square
Dried fruits and vegetables	RandomForestRegression	SelectFromModel	216	0.98
Cake	GradientBoosting	SelectFromModel	234	0.98
Sausage	GradientBoosting	SelectFromModel	168	0.96
Soybean oil	GradientBoosting	SelectFromModel	153	0.94
Waffle	RandomForestRegression	SelectFromModel	236	0.94
Molasses	GradientBoosting	SelectFromModel	449	0.94
Ice cream	RandomForestRegression	SelectFromModel	361	0.94
Instant noodles	RandomForestRegression	SelectFromModel	398	0.93
Raw sugar	GradientBoosting	SelectFromModel	438	0.93
Cookie	GradientBoosting	SelectFromModel	266	0.92
Ham	RandomForestRegression	SelectFromModel	274	0.90
Seasoning	RandomForestRegression	SelectFromModel	845	0.84
Biscuits / Crackers	RandomForestRegression	SelectFromModel	576	0.81
Bacon	RandomForestRegression	SelectFromModel	717	0.79
Bran oil	RandomForestRegression	SelectFromModel	544	0.76
Coconut oil	RandomForestRegression	SelectFromModel	445	0.69
White sugar	RandomForestRegression	SelectFromModel	490	0.66
Palm oil	LinearRegression	RFE	133	0.63
Other crispy pastries	RandomForestRegression	SelectFromModel	319	0.55
Soy sauce / Bean sauce	RandomForestRegression	SelectFromModel	634	0.48

Table 2 The best method and results of beverages products.

Product	Best Model	Feature selection method	Number of selected features	R square
Drinking water	RandomForestRegression	SelectFromModel	252	0.99
Milk	GradientBoosting	SelectFromModel	243	0.98
Curd	RandomForestRegression	SelectFromModel	250	0.97
Soft drink	RandomForestRegression	SelectFromModel	357	0.95
Juice	GradientBoosting	SelectFromModel	191	0.93
Soda	RandomForestRegression	SelectFromModel	410	0.92
Beer	RandomForestRegression	SelectFromModel	420	0.81
fruit-flavored juice	GradientBoosting	SelectFromModel	321	0.67
Tea	RandomForestRegression	SelectFromModel	336	0.62
Yoghurt	GradientBoosting	SelectFromModel	505	0.55
Mixed spirits	RandomForestRegression	SelectFromModel	1244	0.45
White spirits	GradientBoosting	SelectFromModel	480	0.39

Table 3 The best method and results of fast-moving consumer goods products.

Product	Best Model	Feature selection method	Number of selected features	R square
Shampoo	RandomForestRegression	SelectFromModel	349	0.92
Powder	RandomForestRegression	SelectFromModel	333	0.91
Dishwasher	GradientBoosting	SelectFromModel	233	0.87
Sanitary pad	RandomForestRegression	SelectFromModel	528	0.81
Cleanser	RandomForestRegression	SelectFromModel	821	0.68
Softener	RandomForestRegression	SelectFromModel	401	0.67
Soap and skincare	GradientBoosting	SelectFromModel	405	0.67
Washing powder	RandomForestRegression	SelectFromModel	872	0.45

Table 4 The best method and results of electric appliance products.

Product	Best Model	Feature selection method	Number of selected features	R square
Washing machine	RandomForestRegression	SelectFromModel	422	0.96
Modular air conditioning Condensing Units	GradientBoosting	SelectFromModel	232	0.96
Modular air conditioning Girlfriend coil	GradientBoosting	SelectFromModel	224	0.96
Home fan	GradientBoosting	SelectFromModel	185	0.96
Fridge	RandomForestRegression	SelectFromModel	361	0.93
Compressor	RandomForestRegression	SelectFromModel	441	0.88
Rice cooker	GradientBoosting	SelectFromModel	282	0.69

We considered both the R square and MAE values to compare the leading indicator and the most suitable model for a single product. From the example of the instant noodles demand shown in Table 5. We found that random forest regression gives the best score with the highest R square value and lowest MAE.

Table 5 The results of instant noodle from multiple methods.

Model	Feature selection method	Number of selected features	R square	MAE
LinearRegression	SelectFromModel	778	0.75	1647.28
LinearRegression	f_regression	1000	0.82	1356.33
LinearRegression	mutual_information	919	0.83	1270.23
LinearRegression	RFE	98	0.91	1020.94
RandomForestRegression	SelectFromModel	398	0.93	864.52
GradientBoosting	SelectFromModel	219	0.92	925.12

The forecast results for most of the products are acceptable. However, some products' forecasts are not very reliable, which may be due to all the factors that we consider not affecting the demand of that product. We may need to find other factors or experiment with other forecasting methods in the future to better forecast the results of the product.

CONCLUSIONS

This study compared machine learning methods for identifying leading indicators and forecasting demand for 47 products from four industries that use a high packaging volume. Random forest regression gives the best results for more than 60% of the products we are interested in. We found that the leading indicators affecting demand and the forecasting method differ for each product. The demand forecasts for most of the products in this study were acceptable results. However, some products had weak forecasts, which could be because we had not found any factors affecting demand or forecasting methods suitable for that product. In future studies, we will focus on products with poor forecasting results. We will look for other factors and experiment with other machine learning models such as deep neural networks to make more accurate forecasting on such products.

ACKNOWLEDGMENTS

I would like to thank all people who supported and made this study successful.

First, I would like to thank my advisors Asst.Prof.Dr.Ittipong Khemapech, Dr.Manachai Toahchoodee, Asst.Prof.Dr.Watsawee Sansrimahachai, and Dr.Chatchai Wangwiwattana for giving me helpful advice throughout this study. His advice helped me get through the problems I encountered during my studies and led to the success of this study.

Second, I would like to thank my mentor and the people who gave me insight into the packaging business. And thank my friend who besides me during my studies and this project.

Finally, I would like to thank my parents and brother for fully supporting me. They have always encouraged and supported everything.

REFERENCES

Cankurt, S. (2016). Tourism demand forecasting using ensembles of regression trees. *2016 IEEE 8th International Conference on Intelligent Systems (IS).* Sofia.

D'Souza, S., V., P. K., & S, B. (2020). Feature Selection and Modeling using Statistical and Machine learning Methods. *2020 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER).* Udupi.

Darapaneni, N., Muthuraj, S., Prabakar, K., & Sridhar, M. (2019). Demand and Revenue Forecasting through Machine Learning. *2019 International Conference on Communication and Signal Processing (ICCSP).* Chennai.

Feng, Y., & Wang, S. (2017). A forecast for bicycle rental demand based on random

forests and multiple linear regression. *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS).* Wuhan.

Ilseven, E., & Gol, M. (2019). A Comparative Study on Feature Selection Based Improvement of Medium-Term Demand Forecast Accuracy. *2019 IEEE Milan PowerTech.* Milan.

Jain, S., & Nagwani, N. K. (2021). A Study on Impact of Feature Selection on Product Valuation. *2021 6th International Conference for Convergence in Technology (I2CT).* Maharashtra.

Jaiswal, J. K., & Samikannu, R. (2017). Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression. *2017 World Congress on Computing and Communication Technologies (WCCCT).* Tiruchirappalli.

Kumar, N., & Tripathi, M. (2017). Evaluation of effectiveness of ANN for feature selection based electricity price forecasting. *2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT).* Dehradun.

Sulaiman, M. A., & Labadin, J. (2015). Feature selection based on mutual information. *2015 9th International Conference on IT in Asia (CITA).* Sarawak.

Wang, X.-L., Wu, X.-L., & Sun, B.-Y. (2012). Factor selection and regression for forecasting relief food demand. *2012 8th International Conference on Natural Computation.* Chongqing.