Indonesian Journal of Business Analytics (IJBA)
Vol.1, No. 1, 2021: 1-16

A Telemarketing Guidance in Selling Banking Services: A Data Mining Approach

Kattareeya Prompreing[1*], Theera Prompreing[2]
[1*] Faculty of Business Administration and Liberal Arts
Rajamangala University of Technology Lanna, Thailand
Email: katt.rmutl.gmail.com, kattareeya14@rmutl.ac.th
[2]Mukdahan Community Colleage, Thailand Email:ptheera@mukcc.ac.th
*Corresponding author

ABSTRACT: In telemarketing activity, selecting the most potential customers are important because can reduce processing time and operational cost. Therefore, the ability to select the most likely buying customers are urgently needed. In this study, we propose a clear sequence in doing telemarketing activity based on the previous telemarketing data which applying data mining technique. We weight the importance of 16 customer characteristics through 45,211 observations from a Portuguese bank. Applying Random Forest algorithm along with Information Gain Ratio as a criterion and 10-fold Cross Validation, the model able to weight the importance of attributes and achieves 90.01 % accuracy in predicting telemarketing success. Furthermore, the rank of attribute importance was designed to be a guidance map in selecting potential targeted customers as a managerial implication.

Keywords*:* telemarketing, random forest, data mining, cross validation

Submitted: September 11, 2021; Revised: September 12, 2021; Accepted: September 14, 2021

INTRODUCTION

Providing a high quality of goods and services have been recognized as a prominent objective of marketing activity (Leonard, 1982; Rabin, 1983). However, the way traders or marketers introduce and promote the goods or services are affect the efficiency of company's operation, especially in promoting goods or services through telephone call or telemarketing. Ineffective telemarketing activity often leads into low motivation, high stress and high employee turnover (LaRoche, 1993). Based on operational perspective, ineffective telemarketing activity could be meant as high operational cost. On the other hand, customers also may feel uncomfortable to be targeted as the products or services not align with their interest and needs. Therefore, an accurate prediction of prospective customers can beneficial for both side of customers and company.

Nowadays, existing huge development and rapid increasing of information technology allows us to record huge dataset which can help decision makers addressing business problems. However, with large volume and complexity of existing data requires a robust and reliable technique. Furthermore, in 1950s several machine learning were invented and starting from this point, more and more people leveraging and applying this algorithm to reveal hidden pattern and structure of particular data (Kantardzic, 2003). Recently, data mining method is a rapid growing methodology in many businesses application which machine learning used as tool in revealing hidden information (Bose and Mahapatra, 2001).

Random forest is one of popular algorithm in businesses application (e.g. Depari, 2020; Booth et al., 2014; Larivière and Van den Poel, 2005; Kumar and Thenmozhi, 2006; Fantazzini and Figini, 2009; Liu et al., 2015; De Luca et al., 2010). Besides, Krauss et al (2017) compared the performance of random forest, gradient boosted tree and deep neural network in designing statistical arbitrage strategy and found that random forest outperform 2 other algorithms. Moreover, Depari (2021) utilizes random forest algorithm to characterize clusters for material for designing marketing strategy in marketing real estates. Therefore, random forest is a promising algorithm in business application and thus used as well in our study.

The objectives of this study are to predict the success of bank telemarketing activity and proposing a guidance map which can help the telemarketer in selecting prospective customers by applying weighting method of random forest algorithm. In order to achieve near optimum parameters, we applied a grid search technique and to maintain the accuracy and avoiding the overfitting problem, 10 fold cross validation technique also implemented.

THEORETICAL REVIEW

Tekouabou et al (2019) predicted the success of bank telemarketing in promoting long term deposits by using machine learning algorithms such as Naïve bayes, decision tree, artificial neural network, logistic regression, and support vector machine. The objective of that paper is to propose a new data modelling based on the result of algorithm compared. On the other hand, Lahmiri, (2017) also studied about how to improving the prediction model by comparing two-step system presented model and others single model. Eventually, he found that two-step system is outperform any single model. Surprisingly, Jiang (2018) also forecast the success of bank telemarketing with implementing naïve bayes, artificial neural network, support vector machine, logistic regression, and decision tree. This phenomenon shed the light that more work on algorithm comparisons but less on proposing a managerial implication in supporting manager works. Therefore, in this study we dig deeper on helping manager works by creating a guidance map in doing telemarketing activity based on weighting the importance of input variables by data mining approach.

METHODOLOGY AND DATA

Table 1. Statistic Descriptive

Attributes	Min	Max	Average	Deviation
Age	18	95	40.936	10.619
Job	Unknown (88)	Blue-collar (79732)	-	-
Marital	Divorced (5207)	Married (27214)	-	-
Education	Unknown (1857)	Secondary (23202)	-	-
Default	Yes (815)	No (44396)		
Balance	-8019	102127	1362.272	3044.766
Housing	No (20081)	Yes (25130)	-	-
Loan	Yes (7244)	No (37967)	-	-
Contact	Telephone (2906)	Cellular (29285)	-	-
Day	1	31	15.806	8.322
Month	December (214)	May (13766)	-	-
Duration	0	4918	258.163	257.528
Campaign	1	163	2.764	3.098
Pdays	-1	871	40.198	100.129
Previous	0	275	0.58	2.303
Poutcome	success (1511)	Unknown (36969)	-	-
Y	yes (5289)	no (39922)	-	-

The dataset was collected by UCI Machine Learning repository consist of 17 attributes which is 16 as input variables and 1 as label variable. The original objective of this dataset is to predict the telemarketing success in selling bank long-term deposits. The attributes are completely described in table 1. Rapidminer 9.3.000 was used as data analytic tool along with data preprocessing, modelling, leveraging algorithms, deploying and visualization. Within preprocessing stage, we normalized all of input variables using Z-transformation technique which is formulated below. Where Z_k is transformed results, X_k is the dataset, $\bar{X}$ is the mean, and S is standard deviation.

$$Z_k = \frac{X_k - \bar{X}}{S} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(1)}$$

In dealing with prediction and weighting tasks, we performed Random Forest algorithm. Random forest was first introduced by Tin Kam Ho (Ho, 1995) which is the extension of decision tree algorithm. The appearance of Random forest was to address the overfitting problem which is often occurred in decision tree. Furthermore, random forest was developed by Breiman (2001). Random forest has also widely used in solving the problem of predicting and classification such as detecting financial statement manipulation (Patel et al., 2019), predicting student's academic success (Beaulac and Rosenthal, 2019), credit spread approximation (Mercadier and Lardy, 2019), financial early warning analysis (Xiong et al., 2019), assessing determinants of central bank independence (Cavicchioli et al., 2019), direct marketing campaign (Ładyżyński., 2019), measuring credit risk (Ładyżyński et al., 2019) and etc.

In order to have near optimum parameters, a grid search method was employed to find the most optimum parameters based on some predictor parameters. In this case, we optimized 2 random forest parameters such as number of tress and criterion. Number of trees determine the number of random tree in the forest to obtain. This can be achieved by implementing bootstrapping technique in selecting the subset of examples. The number of trees which optimized are 1, 11, 21, 31, 41, 51, 60, 70, 80, 90 and 100. Criterion is also selected to be used to split the data. For criterion type, we optimized information gain ratio, information gain, gini index and accuracy. Afterwards, to predict and evaluate the performance of the learning model, we then applied 10-fold cross validation technique. Cross validation has 2 major processes such as training and testing process. A training process is used to train the data and then apply to the model to test the testing data to see the performance of the model. To understand how cross validation works, the process is drawn on figure 2 below.

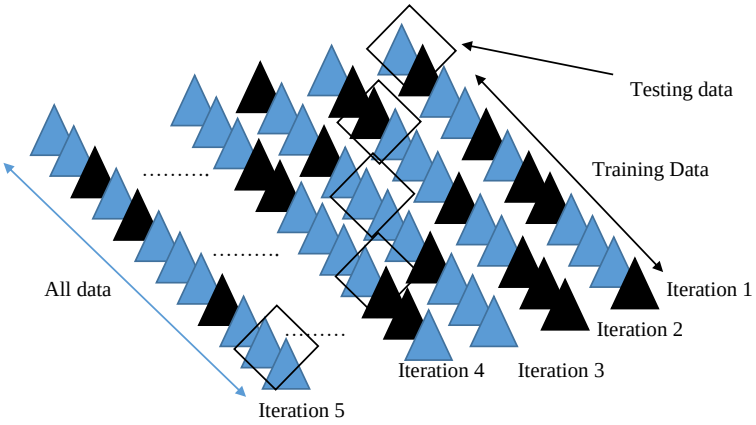

Figure 2. The concept of Cross Validation

Cross validation divides the dataset into k equal size of subsets. Furthermore, a single subset is separated from k subsets and use the single subset as a testing data or subset. The rest k-1 subsets are treated as training data. In order to generate the subsamples, we employed proportionate stratified random sampling technique. Proportionate stratified random sampling is a sampling method that used to generate subpopulation from a population based on equal proportion of each subset. The process of selecting samples are described on figure 3 below.

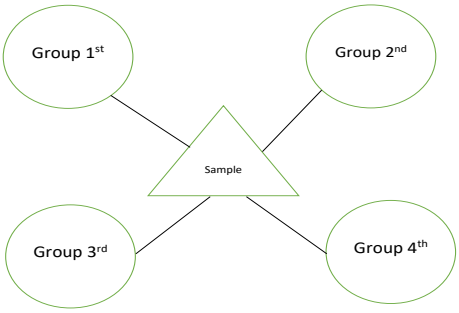

Figure 3. Proportionate Random Sampling

Afterwards, cross validation then iterated as many k subsets and then each of k subset employed once as a test data. In our study, we employed 10 fold cross validation which is meant there are 10 time iterations and the average of those iterations are treated as a model estimation. Cross validation is also known as a powerful method in solving overfitting problem (Cawley and Talbot, 2010). To understand more how our model implemented, the complete sequence using rapidminer is shown on figure 4 below.

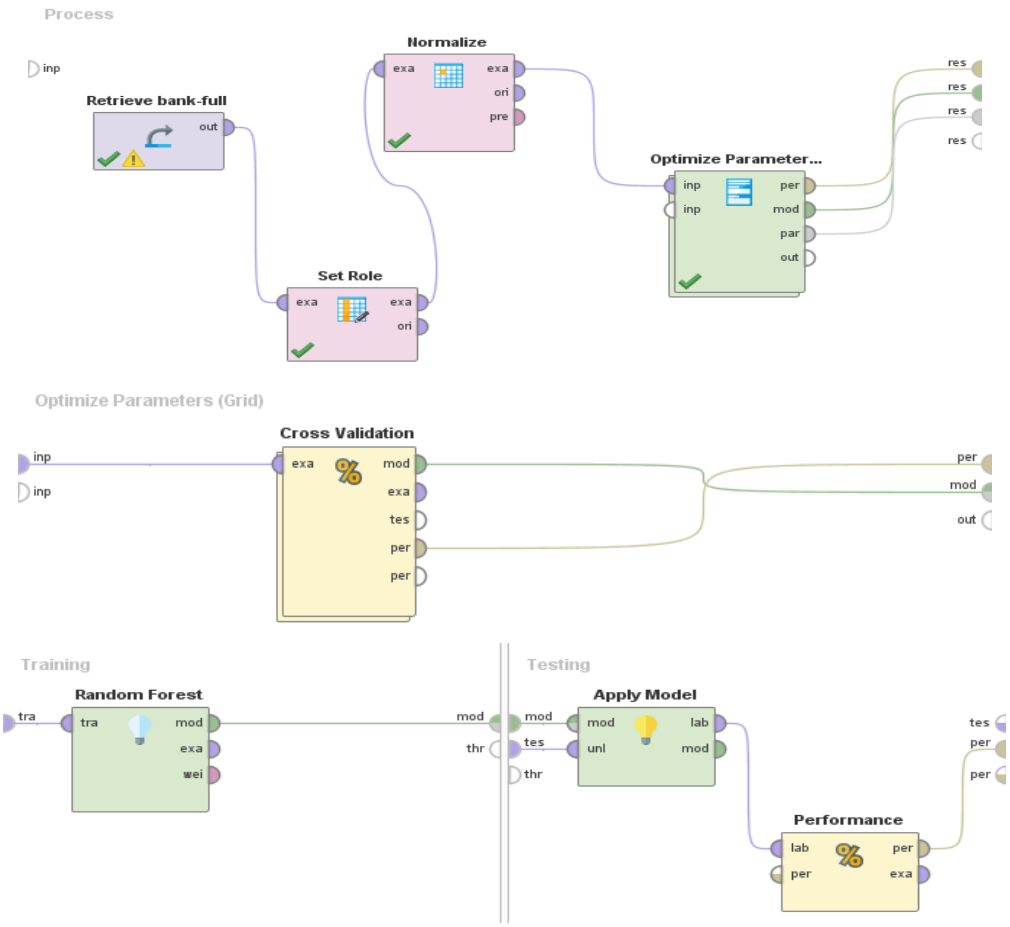

Figure 4. Data mining sequence using Rapidminer

RESULTS AND DISCUSSION

Since dataset contains only 7 numerical attributes (age, balance, day, duration, campaign, pdays, and previous), then z-transformation technique only applied to those 7 numerical attributes. The normalized data by z-transformation are meant to make the data comparable. The results are shown on table 2 below.

Table 2. The result of normalization by Z-transformation technique

No	Age	Balance	Day	Duration	Campaign	pdays	Previous
1	1.606947	0.256416	-1.29846	0.011016	-0.56934	-0.41145	-0.25194
2	0.288526	-0.43789	-1.29846	-0.41612	-0.56934	-0.41145	-0.25194
3	-0.74738	-0.44676	-1.29846	-0.70735	-0.56934	-0.41145	-0.25194
4	0.571045	0.047205	-1.29846	-0.64522	-0.56934	-0.41145	-0.25194
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
45209	2.925368	1.429577	0.143417	3.37376	0.721803	1.436173	1.050462
45210	1.512774	-0.22802	0.143417	0.970136	0.399016	-0.41145	-0.25194
45211	-0.37068	0.528359	0.143417	0.399324	-0.24656	1.476121	4.523527

Applying grid search technique, eventually we generate the near optimum parameters. For the number of trees, there are 100 trees created. In term of criterion, information gain ratio outperform information gain, accuracy, and gini index. Information gain ratio able to achieve 90.06% accuracy combined with 100 trees generated. The results are described on figure 5 and table 3 below. Implementing this combination, we then weighted the importance of 16 importance attributes in order to design a guidance map in doing telemarketing activity.

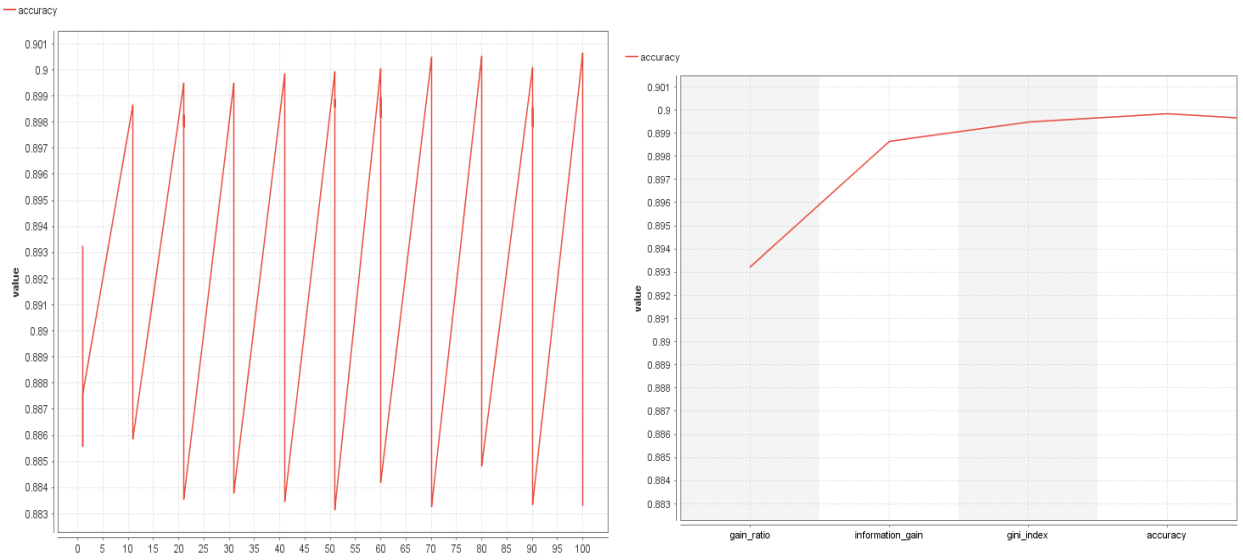

Figure 5. The results of grid search technique for number of trees and criterion

Tabel 3. The results of grid search technique

Number of trees	Criterrion	Accuracy
100	gain_ratio	0.900622
100	information_gain	0.898918
100	gini_index	0.898631
100	accuracy	0.883325
90	gain_ratio	0.900091
.	.	.
.	.	.
1	information_gain	0.887306
1	gini_index	0.885603
1	accuracy	0.887572

The overall prediction accuracy achieves 90.06%. In order to predict the customers who don't buy the long term deposits, the model able to achieve 91.58% accuracy. However, in predicting the small number of success customers, the model only achieve 65.24% accuracy (Table 4). This phenomenon would be a promising area for future research. This further research opportunity can be studied by involving several machine learning algorithms and applying dimensionality reduction technique to reduce the computational cost.

Table 4. Confusion Matrix Result

	true no	true yes	class precision
pred. no	39014	3585	91.58%
pred. yes	908	1704	65.24%
class recall	97.73%	32.22%	

MANAGERIAL IMPLICATION

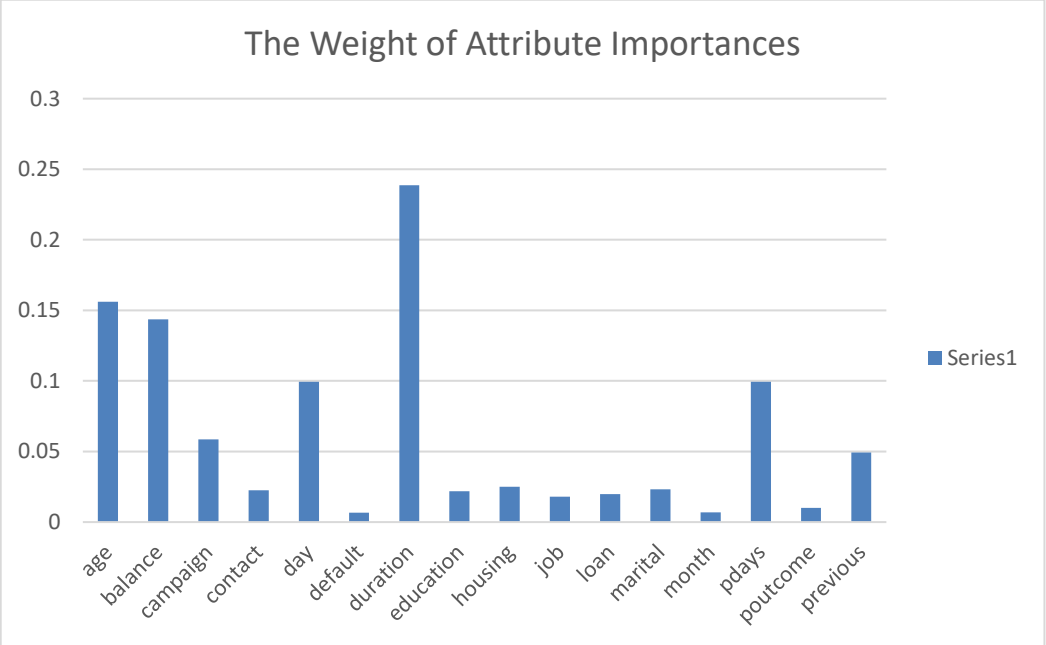

Figure 6. The Weight of Attribute Importance

The results show that the duration of call (previous call) is the most importance factor in predicting prospective customers. However, the duration is only known after the calling process finished. Therefore, the duration is not included as our strategy here but this phenomenon can be used as a reference which emphasizing the longer marketer can persuade prospective customers, the higher opportunity to be closing. Ignoring the duration attribute, then we have age as the most importance attribute, followed by balance, pdays, day and so on. The most importance attribute denotes the most important task that need to do first. Therefore, in this case, age was categorized by 3 groups such as 18-30 years old (young range age), 31-60 years old (Medium range age) and 61-85 (high range age). This categorization is based on recommendations of some expert in data mining field. The results of the age importance are shown on figure 7 below. The age range from 61-85 are found the most importance range age which is

considered as the influential factor in predicting the telemarketing success. Therefore, the age of target customers should be on within this range.

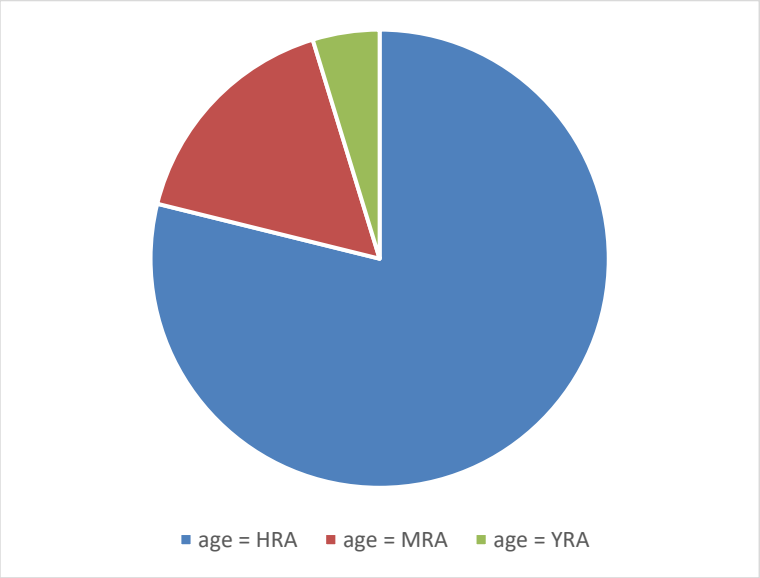

Figure 7. The Importance of Age categorization

Eventually, based on this results, we can draw a guidance map before a telemarketer doing a telemarketing activity. This guidance map is intended to select the most potential customers based on the rank of importance attributes above and then complimented with the rank of sub-attributes presented below (table 5). For several sub-attributes such as housing (yes or no), default (yes or no), and loan (yes or no) are not ranked because of equal importance. It means, for housing attribute, the customer who possess or not possess house, not contribute to predict the success of telemarketing activity, so do for loan and default attributes. In order to select the most potential customers we then applied the rank of attribute and sub-attribute importance below to select the most potential customers, for example, first step we choose customer who is high range age (60-95 years old). The second step is to select the customers who has high balance in their bank account and so on. Table 5 shows the step by following the attributes and sub-attributes rank.

Table 5. The guidance map based on the importance of attributes and sub-attributes

Rank of importance	Attributes	Sub-attributes	Rank applied
1	duration		
2	age	age = HRA	Rank applied
		age = MRA	
		age = YRA	
3	balance		
4	pdays		
5	day		
6	campaign		
7	previous		
8	housing	No	No difference
		yes	
9	marital	Single	Rank applied
		Married	
		Single	
10	contact	unknown	Rank applied
		cellular	
		telephone	
11	education	Tertiary	Rank applied
		Primary	
		Secondary	
		unknown	
12	loan	No	No difference
		Yes	
13	job	job = retired	Rank applied
		job = student	
		job = blue-collar	
		job = unemployed	
		job = entrepreneur	
		job = services	
		job = management	
		job = housemaid	
		job = technician	
		job = admin.	
		job = self-employed	
		job = unknown	
14	poutcome	poutcome = success	Rank applied
		poutcome = unknown	
		poutcome = other	
		poutcome = failure	

15	month	month = mar month = sep	Rank applied
		month = oct	
		month = dec	
		month = may	
		month = apr	
		month = feb	
		month = jul	
		month = jun	
		month = nov	
		month = jan	
		month = aug	
16	default	No	No difference
		yes	

CONCLUSION AND FUTURE RESEARCH

This study employed random forest algorithm to weight the attribute importance and to predict bank telemarketing success based on previous characteristic and transaction records of a Portuguese bank. In order to avoid overfitting problem, 10 fold cross validation are performed. Besides, to have a near optimum parameters, we employed grid search technique to optimize the number of trees and select the fittest criterion to our dataset. Finally, information gain ratio and 100 trees are found as the most optimum parameters which can achieve 90.06% accuracy. We also visualized the parameters so easier to understand in helping decision maker to addressing business problems.

This study contributes at least through 2 aspects. First, a simple, accurate and reliable method are presented which achieves 90.06% accuracy. Second, propose a guidance map in selecting prospective customers in promoting bank long term deposits. The methodology presented in this study also can be applied in fraud detection, maintain the prospective customers and so on. Future potential research could be the implementation of other algorithms along with parameters optimization which somehow has a high computational cost. Therefore, a robust algorithm with near optimum parameters along with cheap computational cost are promising future research.

ACKNOWLEDGMENT

This article and the research behind it would not have been possible without the exceptional support of our affiliated institutions. We would like to express our gratitude to the Rajamangala University of Technology Lanna (RMUTL) in Thailand and Mukdahan Community College, Thailand. Moreover, we sincerely appreciate the reviewer and journal editor of **Indonesian Journal of Business Analytics (IJBA)** for providing a space for international researcher to disseminate research results.

REFERENCES

Beaulac, C. and Rosenthal, J.S., 2019. Predicting University Students' Academic Success and Major Using Random Forests. *Research in Higher Education*, pp.1-17.

Booth, A., Gerding, E. and Mcgroarty, F., 2014. Automated trading with performance weighted random forests and seasonality. *Expert Systems with Applications*, *41*(8), pp.3651-3661.

Bose, I. and Mahapatra, R.K., 2001. Business data mining—a machine learning perspective. *Information & management*, *39*(3), pp.211-225.

Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.

Cavicchioli, M., Papana, A., Papana Dagiasis, A. and Pistoresi, B., 2019. A Random Forests Approach to Assess Determinants of Central Bank Independence. *Journal of Modern Applied Statistical Methods*, *17*(2), p.12.

Cawley, G.C. and Talbot, N.L., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*(Jul), pp.2079-2107.

De Luca, G., Rivieccio, G. and Zuccolotto, P., 2010. Combining random forest and copula functions: a heuristic approach for selecting assets from a financial crisis perspective. *Intelligent Systems in Accounting, Finance & Management*, *17*(2), pp.91-109.

Depari, G. S. (2020). Iklan Berbayar di Social Media: Sebuah Sistem Pendukung Keputusan. Journal of Accounting and Management Innovation, 4(2), 58-71.

Depari, G. S. (2021). Real Estate Segmentation: A Model of Real estate Decision Support System. Sang Pencerah: Jurnal Ilmiah Universitas Muhammadiyah Buton, 7(2), 233-250.

Fantazzini, D. and Figini, S., 2009. Random survival forests models for SME credit risk measurement. *Methodology and computing in applied probability*, *11*(1), pp.29-45.

Ho, T.K., 1995, August. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.

Jiang, Y., 2018. Using Logistic Regression Model to Predict the Success of Bank Telemarketing. *International Journal on Data Science and Technology*, *4*(1), p.35.

Kantardzic, M., 2003. Data Mining: Concepts, Models, Methods, and Algorithms. *Technometrics*, *45*(3), p.277.

Krauss, C., Do, X.A. and Huck, N., 2017. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, *259*(2), pp.689-702.

Kumar, M. and Thenmozhi, M., 2006, January. Forecasting stock index movement: A comparison of support vector machines and random forest. In *Indian institute of capital markets 9th capital markets conference paper*.

Ładyżyński, P., Żbikowski, K. and Gawrysiak, P., 2019. Direct marketing campaigns in retail banking with the use of deep learning and random forests. *Expert Systems with Applications*.

Lahmiri, S., 2017. A two-step system for direct bank telemarketing outcome classification. *Intelligent Systems in Accounting, Finance and Management*, *24*(1), pp.49-55.

Larivière, B. and Van den Poel, D., 2005. Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, *29*(2), pp.472-484.

LaRoche, N.J., Nokia Bell Labs, 1993. *Arrangement for motivating telemarketing agents*. U.S. Patent 5,239,460.

Leonard, F.S., 1982. The incline of quality. *Harv. Bus. Rev.*, pp.163-171.

Liu, C., Chan, Y., Alam Kazmi, S.H. and Fu, H., 2015. Financial fraud detection model: based on random forest. *International journal of economics and finance*, *7*(7).

Mercadier, M. and Lardy, J.P., 2019. Credit spread approximation and improvement using random forest regression. *European Journal of Operational Research*.

Patel, H., Parikh, S., Patel, A. and Parikh, A., 2019. An Application of Ensemble Random Forest Classifier for Detecting Financial Statement Manipulation of Indian Listed Companies. In *Recent Developments in Machine Learning and Data Analytics* (pp. 349-360). Springer, Singapore.

Rabin, J.H., 1983. Accent is on quality in consumer services this decade. *Marketing News*, *17*(4), p.12.

Tang, L., Cai, F. and Ouyang, Y., 2019. Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in china. *Technological Forecasting and Social Change*, *144*, pp.563-572.

Tekouabou, S.C.K., Cherif, W. and Silkan, H., 2019, March. A data modeling approach for classification problems: application to bank

telemarketing prediction. In *Proceedings of the 2nd International Conference on Networking, Information Systems & Security* (p. 56). ACM.

XIONG, S.Y., Chen, L.U., CHANG, L. and XIE, A.R., 2019. Impact Analysis of Financial Early Warning Indicators Based on Random Forest. *DEStech Transactions on Computer Science and Engineering*, (iteee).