Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

A Study on the Method of Predicting the Success of Bank Telemarketing

SeungYong Lee[1], Youngkeun Choi[2], Yangmi Lim[3]

[1]College of Liberal Arts, Namseoul University, 91 Daehak-ro, Seonghwan-eup, Seobuk-gu, Cheonan-si, Chungcheongnam-do, Republic of Korea, leesky@nsu.ac.kr
[2]Division of Business Administration, College of Business, Sangmyung University Seoul, Republic of Korea, penking1@smu.ac.kr
[3]IT media department, Duksung Woman's University, 33 Samyang-ro 144-gil, Dobong-gu, Seoul, Republic of Korea, yosimi@duksung.ac.kr

This study applied a machine learning technique to a dataset originating from direct marketing campaigns conducted by a Portuguese banking institution, sourced from the UC Irvine Machine Learning Repository. The findings reveal that, firstly, among all the variables, age, balance, loan, day, duration, campaign, pdays, and poutcome exert a significant influence on the success of bank telemarketing, whereas job, marital status, education, default, housing, contact method, month, and previous interactions exhibit no statistical significance. Secondly, for the complete model, the accuracy rate stands at 0.784, indicating an error rate of 0.216. Within the subgroup of individuals predicted not to succeed in bank telemarketing, the accuracy of this prediction is 75.63%, while among those predicted to achieve success in bank telemarketing, the accuracy reaches 82.61%.

Keywords: Machine learning, Decision tree, Financial industry, The success of bank telemarketing prediction.

1. Introduction

In the field of bank direct marketing, various data mining techniques have been employed for classifying marketing services, including the One-R Algorithm, naive Bayes classifier, classification, and association rule mining [1]. Exploratory data analysis will be utilized to uncover relationships between variables and the class variable, as well as the interplay between two variables in relation to the class variable. Data mining algorithms, specifically classification, will be employed to categorize bank customer data. The objective of classification is to forecast whether a client will subscribe to a term deposit, which is represented by the outcome variable. Classification entails the process of constructing a model (or function) that characterizes and distinguishes data classes or concepts, enabling the model to predict the class of objects with unknown class labels. The model is developed through the analysis of a training dataset, which comprises data objects with known class labels.

Machine learning, in recent times, encompasses transformations in systems engaged in tasks associated with artificial intelligence (AI) [2]. These tasks encompass activities such as recognition, analysis, planning, robot control, forecasting, and more. It delves into the study and development of algorithms capable of making predictions based on data. Machine learning involves the construction of programs equipped with tuning parameters that autonomously adapt to enhance their performance by learning from past data. This rapidly advancing technique mimics human cognitive processes, representing multi-level data and adeptly addressing the selectivity-invariance dilemma [3].

Machine learning finds application across various domains, with notable relevance in the financial sector [4]. In fact, telemarketing constitutes a central strategy for numerous banks in their customer interactions. In the realm of direct marketing, banks can employ machine learning to predict a customer's receptivity to marketing efforts, automating the entire feature validation process. However, a drawback of this model is its propensity to assign varying weights to individual factors. In reality, the success of bank telemarketing may, at times, hinge on a single predominant factor, a scenario not effectively accommodated by this system.

The primary objective of this study is to conduct an analysis and prediction of the effectiveness of bank telemarketing, enabling banks to formulate strategic responses based on the study's findings. The methodology outlined in this paper serves as a structured guide for readers to replicate the investigative process and apply similar procedures to address various other issues.

This paper is aimed at providing a streamlined, expeditious, and user-friendly approach for identifying eligible customers, offering distinct advantages to the bank. The bank telemarketing success prediction system automatically computes the significance of each feature contributing to the bank telemarketing success. When applied to new test data, these features are evaluated based on their respective significance.

A predefined time frame can be established for applicants to determine whether their bank telemarketing endeavors can be authorized. The bank telemarketing success prediction system facilitates prioritized processing for specific applications as needed.

2. Related Study

The listed researchers are renowned figures in this specific field of study. In contrast to techniques employing deep auto-encoders, the proposed method demonstrated superior performance, surpassing previous works in terms of results.

Elsalamony outlined the knowledge discovery process as an interactive and iterative sequence of steps, including understanding the application domain, data selection, data preprocessing and cleaning, data integration, data reduction, transformation, algorithm selection for data mining, interpretation, description of results, and utilization of the discovered knowledge [5].

Data mining can be categorized into two primary classes: descriptive and predictive. In recent years, data mining has garnered significant attention in the business and banking sectors due to its adaptability in handling vast datasets and converting such data into comprehensible information and knowledge [1].

It's worth noting that many individuals may find it challenging to distinguish between the terms "knowledge discovery" and "data mining" across various domains. Knowledge discovery in databases refers to the process of identifying valid, novel, potentially useful, and ultimately comprehensible patterns or models within data.

On the contrary, data mining constitutes a pivotal stage within the knowledge discovery process, encompassing specialized data mining algorithms that, while adhering to acceptable computational efficiency constraints, unearth patterns or models within data [6].

Typically, selected customers are directly contacted through various means such as personal interactions, telephone, cellular, mail, email, or other communication channels to promote new products or services [7]. This marketing approach is referred to as direct marketing. Direct marketing serves as a primary strategy for numerous banks and insurance companies in their interactions with customers [8]. Some banks and financial services firms may rely exclusively on mass marketing strategies to promote new services or products to their customer base. In this approach, a single communication message is broadcast to all customers through various media outlets like television, radio, or advertising agencies, among others [6].

The literature suggests that direct marketing has evolved into a critical application in data mining. Data mining is extensively utilized in direct marketing to identify potential customers for new products by utilizing purchase data and predictive models to gauge a customer's likelihood to respond to promotional offers [6]. Especially within the banking sector [9], direct marketing is an invaluable and widely employed strategy for directly engaging with customers or potential customers, bypassing indirect channels.

However, it is crucial to identify the "right" customers to ensure the success of bank telemarketing. Unwanted product offerings to contacted customers can be seen as intrusive, and inundating inbound calls with excessive campaign content can be irritating. Therefore, greater emphasis should be placed on the task of selecting the most suitable clients or targeting specific customer segments—those more inclined to subscribe to a product [10].

Leveraging these data mining techniques, organizations and institutions can uncover previously unavailable information about their customers and products, enabling them to precisely define customer preferences and forecast future behaviors and requirements [11]. Recent technological advancements have significantly impacted the marketing domain, leading to a shift toward targeted marketing campaigns rather than mass marketing. The latter has become less effective due to heightened competition in a rapidly evolving market landscape [12].

3. Methodology

3.1 Dataset

The dataset utilized for experimentation in this paper pertained to direct marketing campaigns conducted by a Portuguese banking institution and is accessible via the UCI Machine Learning Repository [10]. These marketing campaigns predominantly relied on phone calls, with multiple contacts often being necessary to ascertain whether the client would subscribe to the bank term deposit ("yes") or not ("no"). The dataset encompasses the outcomes of 17 direct bank marketing campaigns carried out by a Portuguese bank between May 2008 and November 2010. Detailed descriptions of the 17 attributes are presented in Table 1.

Table 1: The variables in each category

Variables	Description
Age	numeric, age of client
job	categorical, type of job (admin, unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services)
marital	Categorical, marital status (married, divorced, single. Here ‖divorced‖ states the both divorced or widowed)
education	categorical (unknown, secondary, primary and tertiary)
default	binary, customer credit is in default (yes, no)
balance	numeric, average yearly balance (in euros)
housing	binary, status of housing loan (yes, no)
loan	binary, clients personal loan (yes, no)
contact	categorical, contact communication type (unknown, telephone, cellular)
day	numeric, the last contact day of the month range (1-31)
month	categorical, last contact month of the year
duration	numeric, last contact duration (in seconds)
campaign	numeric, number of contacts performed during this campaign
pdays	numeric, number of days that passed by after the client was last contacted from a previous campaign
previous	numeric, number of contacts which are made before this campaign
poutcome	categorical, result or outcome of the previous marketing campaign (unknown, other, failure, success)
y	binary, (desired target) client subscribed a term deposit or not

3.2 Decision Tree

Among various analytical techniques, the decision tree (DT) stands out as a potent and widely adopted machine learning algorithm in contemporary data analytics, renowned for its predictive and classifying capabilities, especially in handling big data scenarios [13]. Decision trees are proficient in addressing both classification and regression tasks. One might wonder why we prefer employing a DT classifier over alternative classifiers. To address this query, two compelling reasons can be elucidated. Firstly, decision trees endeavor to emulate human cognitive processes, rendering data interpretation and the derivation of meaningful conclusions or insights a relatively straightforward task. Secondly, decision trees offer transparency in revealing the underlying data logic, in contrast to black-box algorithms such as Support Vector Machines (SVM) and Neural Networks (NN). This simplicity and clarity have made decision trees a favored choice among contemporary programmers [14].

Having discussed the merits of decision trees, let's delve deeper into understanding the decision tree classifier. A decision tree is a hierarchical structure composed of nodes, where each node represents a distinct feature or attribute. Branches connecting nodes signify decisions or rules, while the terminal leaves of the tree represent outcomes, which can be either categorical or continuous values [15]. The overarching concept is to construct a tree that encompasses the entire dataset and yields an outcome at each leaf.

Now that we have a better grasp of what a decision tree entails, let's proceed to explore the process of building a decision tree classifier. Decision trees can be constructed using two primary algorithms: CART (Classification and Regression Trees) and ID3 (Iterative Dichotomiser 3).

Starting with ID3, the initial step involves selecting an 'x' value from the column and a 'y' value, which is located in the last position of the column and possesses only "YES" or "NO" values. In the presented chart, our 'x' values consist of (outlook, temp, humidity, windy), and 'play,' which exclusively offers two options, namely 'YES' or "NO," serves as our 'y' value. The next stage is to establish the mapping between 'x' and 'y.' Since it's a binary classification problem, we will proceed to construct the tree utilizing the ID3 algorithm.

To build the tree, we must commence with a root node selection, and the initial choice for the root node [16] typically adheres to a general guideline: opt for the feature that exerts the most substantial influence on the 'y' value. Subsequently, we proceed to designate the next most influential feature as the subsequent node. In this context, we employ the concept of entropy, which quantifies the level of uncertainty within the dataset. It is imperative to compute the entropy for all categorical values pertinent to the binary classification problem [17].

To encapsulate the process succinctly, it entails computing the entropy for the entire dataset initially. Then, for each attribute or feature, we embark on the following steps: calculate entropy for all categorical values, determine the average value of information entropy for the current attribute, and ascertain the level of information gain achieved for the current attribute. Subsequently, the attribute with the highest information gain is selected, and this process iterates until the desired tree structure is attained. This, in essence, outlines the essence of the ID3 process.

As previously discussed, the decision tree classifier can also be constructed using another algorithm known as CART, an acronym for Classification and Regression Trees. In this algorithm, the Gini index serves as the cost function utilized to assess the efficacy of dataset splits. In our scenario, the target variable is indeed binary, with two possible values (yes and no), resulting in four possible combinations [18].

The key objective now is to determine the Gini score, which provides valuable insights into how the data can be effectively partitioned. An ideal scenario yields a Gini score of 0, signifying a perfect separation, whereas the worst-case scenario results in a 50/50 split. The question that naturally arises is how to compute the Gini index value.

If the target variable is a categorical variable with multiple levels, the computation of the Gini index remains similar. The method entails several steps, starting with the calculation of the Gini index for the dataset. Subsequently, for each feature, the Gini index for all categorical values must be calculated, followed by the determination of the average information entropy for the current attribute. Finally, the Gini gain is computed. Once these steps are completed, the attribute with the highest Gini gain is selected, and the process is repeated until the desired tree structure is achieved. This encapsulates the fundamental workings of the decision tree algorithm [19].

Decision tree (DT) classification methods involve the construction of tree models comprising a series of predictors. Within a training set, each of these predictors (attributes) undergoes repetitive splitting until pure subsets are obtained. This splitting process is influenced by the specific characteristics of entities, such as customers [20]. The fundamental structure of a DT includes both leaf nodes and decision nodes. Leaf nodes represent predictor variables and mark the points where binary splits occur. These leaf nodes are also referred to as internal nodes

[21]. On the other hand, decision nodes, alternatively known as terminal nodes, represent the output variable, typically a binary outcome variable, and are visually depicted as the endpoints of branches. It is the terminal node that serves as the foundation for churn prediction, as it reports the category with the majority of cases.

The existing literature identifies four major DT machine learning algorithms commonly employed:

1. Classification and Regression Trees (CART)

2. C4.5

3. Chi-squared automatic interaction detection (CHAID)

4. C5.0

DTs serve as the basis for other tree methods such as random forests and ensemble forests, which involve aggregating multiple decision trees to enhance predictive performance [20].

The process of binary splitting of attributes hinges on the selection of the appropriate attributes for splitting. The correct choice of attributes depends on the computation of entropy measures (C4.5) or the adoption of the Gini criterion (CART), depending on the specific DT algorithm employed[22]. DT analysis enjoys widespread popularity due to its simplicity, visual representation, and ease of interpretation. DTs offer a suitable framework for modeling both quantitative and qualitative decision-making questions without the need for creating dummy variables or transformations. Additionally, DTs can effectively capture non-linear relationships and are computationally efficient [23].

However, it's essential to acknowledge that DTs may not always deliver predictive accuracy comparable to other methods. Furthermore, minor alterations in the dataset can lead to non-robust predictions [20]. Nevertheless, this classification technique has found frequent use in modeling churn [24].

3.3 Data mining models

To thrive in an increasingly competitive marketplace, numerous companies are adopting data mining techniques for decision prediction analysis. To efficiently manage customers, it is imperative to construct a more precise and effective decision prediction model. Both statistical and data mining techniques have been leveraged to formulate decision prediction models. Data mining techniques prove valuable in uncovering noteworthy patterns or relationships within the data and making predictions or classifications by fitting models based on available data.

In situations where a learning dataset and a test dataset are segregated for machine learning, the test dataset must meet specific criteria. Firstly, the training dataset and the test dataset must adhere to the same format. Secondly, the test dataset should not include any data from the training dataset. Thirdly, the training dataset and the test dataset must exhibit data consistency. However, creating a test dataset that satisfies these criteria can be quite challenging. In the realm of data mining, several verification frameworks have been developed to address this challenge using a single dataset.

In this study, we employ the Split Validation operator provided by RapidMiner to address this issue. This operator partitions the input dataset into a training dataset and a test dataset, facilitating performance evaluation. Specifically, this study opts for relative segmentation among the segmentation method parameters of this operator, utilizing 70% of the input data as the learning dataset.

3.4 Performance evaluation

Performance assessment relies on training data to evaluate the effectiveness of the generated model. These assessments encompass technical performance measures and heuristic measures. In this study, the technical performance measures employed reveal performance outcomes through the generation of models using training data, transformation of test data into models, and the comparison of class labels from original verification cases with those predicted.

The evaluation of technical performance can be categorized into two primary domains: supervised and unsupervised learning. Within the realm of supervised learning in this study, classification and regression techniques are utilized. The data employed for both learning and testing purposes feature original class values, and performance evaluation entails a comprehensive analysis that involves the comparison of these original class values with the predicted outcomes.

The classification problem stands out as one of the most prevalent data analysis challenges. To gauge the efficacy of classification models, a multitude of metrics have been devised. In cases involving categorical classification, metrics such as accuracy, precision, recall, and f-measure find extensive utility. Within RapidMiner, Performance (Classification) is a tool that assesses performance metrics for common classification problems, while Performance (Binomial Classification) is tailored to provide performance indicators exclusive to binomial classification scenarios. Table 2 outlines the methodology for calculating these performance indicators.

Table 2: Key performance indicators of binomial classification

		Actual class (as determined by Gold Standard)	
		True	False
Predicted class	Positive	True Positive	False Positive(Type I error)
	Negative	False Negative(Type II error)	True Negative

Precision = TP/(TP+FP), Recall = TP/(TP+FN), True negative rate = TN/(TN+FP), Accuracy = (TP+TN)/(TP+TN+FP+FN), F-measure = 2·((precision·recall)/(precision + recall))

4. Results

4.1 Decision tree

Figure 1 displays the classification tree representing the full model, which has undergone pruning through cross-validation to mitigate overfitting, following the methodology proposed by Kuhn and Johnson [25]. The primary variables in the comprehensive model analysis encompass a total of eight, as delineated below, with each variable adhering to a predefined criterion. Within this context, it is worth noting that among all the variables, age, balance, loan, day, duration, campaign, pdays, and poutcome significantly influence the success of bank telemarketing. Conversely, variables such as job, marital status, education, default status, housing, contact method, month, and previous interactions exhibit no discernible significance in this regard.

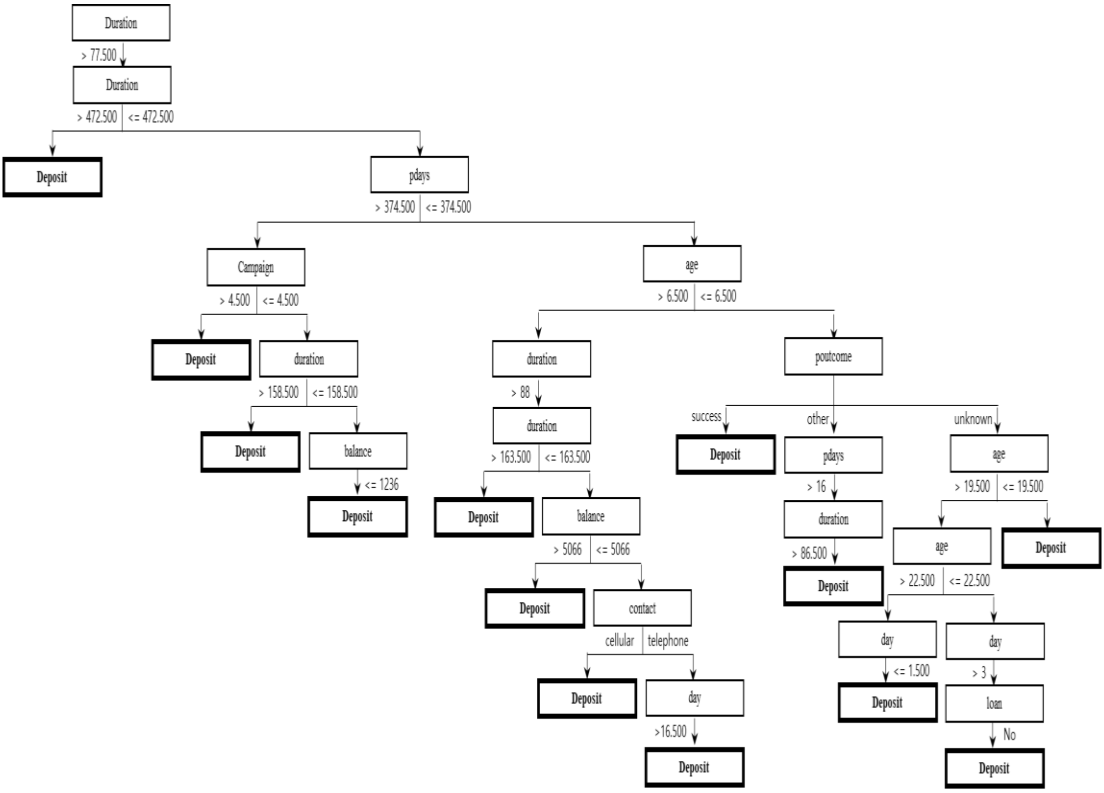

Figure 1: Classification Tree for the Full Model. Machine learning model showing a step-by-step decision-making process for telemarketing.

Tables 3 present a comprehensive breakdown of the various measures derived from the confusion matrix. Specifically, for the full model, the accuracy rate stands at 0.784, signifying an associated error rate of 0.216. Furthermore, among the subset of patients who were predicted not to achieve success in bank telemarketing, the accuracy in correctly identifying those who did not succeed was 75.63%. Conversely, the accuracy in correctly identifying those who did succeed in bank telemarketing among the patients predicted to do so stood at 82.61%.

Table 3: Performance evaluation

	True Y	True N	Class precision
Pred. Y	1102	232	82.61%
Pred. N	491	1524	75.63%
Class recall	69.18%	86.79%	

5. Conclusions

The primary objective of this paper is twofold: to assess model accuracy and develop a novel predictive model for forecasting the success of bank telemarketing. This study encompasses two key goals. Firstly, it aims to enhance our comprehension of the role of variables in predictive modeling for bank telemarketing success. Secondly, it endeavors to assess the predictive capabilities of decision tree models.

In light of the aforementioned findings, several implications emerge. With regard to the first goal, the study's results suggest that evaluating the role of variables is a complex endeavor, with their impacts varying depending on the classification methods utilized. Decision tree methods emphasize the significance of explanatory power in the analysis. Therefore, it is evident that there is no unanimous consensus on which explanatory variables are of utmost importance in predicting the success of bank telemarketing across all methods employed.

Nevertheless, this study does provide additional insights into the customer's profile. Banks and financial companies should strive to predict the success of bank telemarketing using the classification methods employed. For instance, among all variables, age, balance, loan, day, duration, campaign, pdays, and poutcome significantly influence the success of bank telemarketing, while job, marital status, education, default, housing, contact method, month, and previous interactions have no significant impact. Furthermore, for the full model, the accuracy rate stands at 0.784, implying an error rate of 0.216. Among the individuals predicted not to achieve success in bank telemarketing, the accuracy of this prediction was 75.63%, while the accuracy for those predicted to succeed was 82.61%.

This study makes both research and practical contributions. Firstly, it expands the existing literature by empirically investigating the collective impact of various variables on the modeling of bank telemarketing success. While numerous studies have explored the prediction of bank telemarketing success, it remains a challenge to create a universal predictive tool due to its complexity and multifaceted nature. Researchers often use a limited set of factors, overlooking the effects of other variables. Customer demographics frequently change and require ongoing monitoring, posing challenges for banks and financial institutions and raising concerns about data privacy. This study contributes to the literature on bank telemarketing success prediction by presenting a comprehensive model that summarizes the determinants of success, considering various customer factors.

Secondly, the methodology employed in this paper can serve as a roadmap for readers to follow when addressing similar problems. It offers a systematic approach to identify the root causes of various issues. The paper strives to develop the most effective predictive model for bank telemarketing success using a restricted set of features, including customer-related factors. Machine learning techniques, such as decision trees and neural networks, in conjunction with feature importance analyses, are employed to achieve optimal accuracy. This methodology facilitates the identification of patterns related to bank telemarketing success prediction.

In practical terms, this application facilitates the management of customers' personal records by bank and finance companies, streamlining decision-making processes when user data is readily available. The paper outlines a prototype of the model that organizations can utilize to make informed decisions regarding the approval or rejection of customer requests for successful bank telemarketing. Importantly, this study is designed exclusively for the managerial authority within the bank company, ensuring the privacy and integrity of the prediction process. The results pertaining to the success of bank telemarketing IDs can be transmitted to various departments within the banks, enabling them to take appropriate actions on applications. This streamlined communication helps other departments carry out their respective formalities efficiently.

In the proposed system, a database is essential to store customer records. As the customer base expands, the volume of generated data increases, posing potential storage challenges. To address this concern, a future release will incorporate cloud-based storage, ensuring secure data protection and enabling remote access with appropriate authorization. Additionally, in upcoming iterations, the smart device will be synchronized with our application. This synchronization will enable real-time monitoring of customers' financial statuses, allowing banks and finance companies to receive alerts in case of financial needs.

In the future, the machine learning model will leverage an extensive training dataset, potentially encompassing over a million distinct data points stored within an electronic financial record system. While this transition would necessitate significant advancements in computational power and software sophistication, an AI-driven system could empower financial practitioners to expedite the decision-making process, facilitating tailored decisions for individual customers.

Acknowledgement: Funding for this paper was provided by Namseoul University

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data availability statements

The generated during and/or analysed during the current study are available in the UCI Machine Learning Repository The marketing campaigns repository, https://archive.ics.uci.edu/ml/datasets/bank+marketing

References

1. Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. MA: Morgan Kaufmann Publishers. https://doi.org/10.1016/B978-0-12-381479-1.00001-0
2. Simon, J. (2019). Artificial intelligence: scope, players, markets and geography. Digital Policy, Regulation and Governance, 21(3), 208-237. https://doi.org/10.1108/DPRG-08-2018-0039
3. Cun, Y. Le, Bengio, Y., & Hinton, G. E. (2015). Deep learning. Nature 521, 436-444. https://doi.org/10.1038/nature14539.
4. Kose, U. (2019). Using Artificial Intelligence Techniques for Economic Time Series Prediction. Contemporary Issues in Behavioral Finance, Emerald Publishing Limited, 13-28.
5. Elsalamony, A. H. (2014). Bank Direct Marketing Analysis of Data Mining Techniques. International Journal of Computer Applications 85.7, (2014): 12–22. https://doi.org/10.5120/14852-3218.
6. Ian H. Witten, Eibe Frank, M. A. A. H. (2013). Data Mining Practical Machine Learning Tools and Techniques with Java implementations. Acm Sigmod Record 31(1), 76-77. Retrieved from http://ir.obihiro.ac.jp/dspace/handle/10322/3933
7. Jeongmin Y, Younghwan P., (2020), Classification of Generation Z Utility for Function of Mobile Financial Services : Based on KANO Model, Journal of Next-generation

Convergence Information Services Technology, Vol.9, No.3, pp. 197-210, DOI : 10.29056/jncist.2020.09.02

8. Ayetiran, E. F., & Adeyemo, A. B. (2012). A data mining-based response model for target selection in direct marketing. IJ Information Technology and Computer Science, 1(1), 9–18. https://doi.org/10.1016/j.cie.2016.07.006.
9. Nachev, A. (2014). Application of Data Mining Techniques for Direct Marketing. Computational Models for Business and Engineering Domains, 86–95.
10. Moro, S., Cortez, P. & Rita, P. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, 62, 22-31. https://doi.org/10.1016/j.dss.2014.03.001.
11. Moin, I. K., & Ahmed, B. Q. (2012). Use of Data Mining in Banking. International Journal of Engineering Research and Applications, 2(2), 1–5.
12. Su, C. T., Chen, Y. H., & Sha, D. Y. (2006). Linking innovative product development with customer knowledge: a data-mining approach. Technovation, 26(7), 784–795. https://doi.org/10.1016/j.technovation.2005.05.005.
13. Gonzalez-Cava, J. M., Reboso, J. A., Casteleiro-Roca, J. L., Calvo-Rolle, J. L., & Pérez, J. A.M. (2018). A Novel Fuzzy Algorithm to Introduce New Variables in the Drug Supply Decision-Making Process in Medicine. Complexity 2018, 1–15. https://doi.org/10.1155/2018/9012720.
14. Arnold, C., Heart disease. New York: Franklin Watts, 1990
15. Canfield, J., Hansen, M. V., and Rackner, V. (2005). Heart disease. Health Communications Books.
16. Dittmer, L. (2012). Heart disease. MN Creative Education.
17. Goetz, T. (2010). The decision tree: taking control of your health in the new era of personalized medicine., New York: Rodale Books
18. Gold, J. C. & Cutler, D. J. (2000). Heart disease. Enslow Publishers.
19. Healey, J. (2005). Heart disease. Australia, Spinney Press.
20. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: springer.
21. Mendez, G., Buskirk, T. D., Lohr, S., & Haag, S. (2008). Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. Journal of Engineering Education, 97(1), 57-70. https://doi.org/10.1002/j.2168-9830.2008.tb00954.x.
22. Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211-229. https://doi.org/10.1016/j.ejor.2011.09.031.
23. Höppner, S., Stripling, E., Baesens, B., & Verdonck, T. (2017). Profit Driven Decision Trees for Churn Prediction. arXiv preprint arXiv:1712.08101. Retrieved from: https://arxiv.org/abs/1712.08101.
24. De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research 269.2, 760-772. https://doi.org/10.1016/j.ejor.2018.02.009.
25. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer.