Open
Full Length Article

ORIGINAL ARTICLE

Product-Pair Recommendation for Customers using Machine Learning Technique

Maryam Maqbool[1], Syed Saood Zia[2], Muhammad Naseem[3], Syed Abbas Ali[4]

1Department of Computer Information System, NED University of Engineering & Technology Karachi, Pakistan
[2]Department of Software Engineering, Sir Syed University of Engineering & Technology Karachi, Pakistan
[3]Department of Software Engineering, Sir Syed University of Engineering & Technology Karachi, Pakistan
[4]Department of Computer Information System, NED University of Engineering & Technology Karachi, Pakistan

ABSTRACT

Bank marketing campaigns mostly rely on human expert's opinions on choosing potential customers. This method is time consuming and lacks accuracy. Marketing offers being send out to customers are mostly targeting the customers with irrelevant products that are not suitable according to their financials and transactional history. This mismatch means, the bank is wasting their precious marketing resources and missing opportunities to please loyal customers and hence lost the profit. In this paper, a Pakistani banking dataset has been analyzed using two Machine Learning algorithms i.e. Decision Tree (Cart 4.5) and Random Forest to perform data analytics and recommend the best customer-product pair. The main intent of this research paper is to identify different banking products (Two Different Term Deposits Categories) suitable for various contrasting customers. The experimental results obtained will help in predicting and contacting the relevant customers with the best Term Deposit Category suitable according to their financials and also identifying the key features contributing towards the subscription of different Term Deposit categories. The performance evaluation of these algorithms are computed and determined by three of the statistical measures: classification accuracy, specificity and sensitivity.

Keywords: Bank Marketing, Machine Learning, Data Analytics, Customer Product-Pair

Author`s Contribution
[1,2,3,4]*Manuscript writing, Data Collection Data analysis, interpretation, Conception, synthesis, planning of research, and discussion*

Address of Correspondence
Syed Saood Zia
Saood_zia@hotmail.com

Article info.
Received: Aug 12, 2019
Accepted: May 22, 2020
Published: June 30,2020

Cite this article: *Maqbool M, Zia SS, Naseem M, Ali SA S.Product-Pair Recommendation for Customer using Machine Learning Technique. J. Inf. commun. technol. robot. appl.2020; 11(1):38-46.*

Funding Source: *Nil*
Conflict of Interest: *Nil*

INTRODUCTION

Banks have very structured and detailed information and transaction records of every single customer that could be termed as big data[1][2]. This banking big data can be used for many purposes by extracting useful information from it. The under-utilization of keeping complete portfolio of the customers provide no other big benefit other than anti money laundering tracking or capturing fraudulent activities. However, this big data using machine learning models can be used for marketing different banking products for each type of customer and maximize the direct marketing campaigns of the bank and bring out more revenue, and profit for the bank [3][4]. Marketing not only helps in bringing more customers to the bank or selling out more products to customer, but it also improves the ranking and position of bank in this competitive environment which is quite important [5].

As the technology advances, the era of machine learning and big data analytics urges us to replace the old method of marketing strategies being used in banks [6] be replaced by new and innovative methods that provide maximum accuracy and offers fast processing, resulting in improved and timely results [7]. Similarly, banks can use supervised machine learning algorithms that can help the marketing teams to identify which banking product would be suitable for which customer and they would contact the customers accordingly [8]. Since they will not be blindly targeting the customers with the products not of customers' interest.

For example, a person who is maintaining just an average balance of 25 thousand PKR will not be interested in having a luxurious car loan instead; he would be more inclined towards saving products. A person's credit and debit monthly history, total number of accounts, average monthly balance maintained, category (type) of the account already existing, source of income and many other similar parameters could be used to predict the best customer-product pair. This customer-product pair will definitely help the marketing teams to streamline their customer targeting process and will also increase the customer satisfaction as the customers will not be bothered by wrong advices and wrong products.

This research is crafted into following sections: The successive section presents the related work and literature review. Section 3 shows the research methodology. Section 4 provides the experimental results and discussions. Whereas, Section 5 shows the conclusion and future directions.

LITERATURE REVIEW

The term 'direct marketing' introduced by Lester Wunderman means to eliminate the distributors and put the producer and customer into direct contact [9]. Numerous diverse studies have been conducted by many researchers and data miners in the aspect of bank direct marketing. Analysis of data mining techniques to help direct marketing of bank by use of multilayer perception neural network (MLPNN), Nominal regression or logistic regression (LR), tree augmented Naïve Bayes (TAN) known as Bayesian networks and Ross Quinlan new decision tree model (C5.0) in order to know which customers will subscribe to Term Deposit and which will not [10]. A Portuguese bank data for direct marketing has been used to perform three iterations of CRISP-DM methodology for tuning the Data mining model results. The best model found to be was SVM. Sensitivity analysis was used to measure the input variable importance in the SVM that can be used by Bank managers to enhance their marketing campaigns[11]. Two main objectives are assessed in of a classical bank direct marketing campaign dataset. The first main objective was to foresee the customer response to bank's direct marketing campaign [12] through using four classifiers: Decision Tree (C4.5), Random Forest (RF), Multilayer Perceptron Neural Network (MLPNN) and Logistic Regression [13]. Out of which Random Forest proved to be the best Classifier with an accuracy of 87%. Identification of key features of customer to subscribe to a term deposit or not, using clustering analysis was the second objective. In AWS Machine Learning is used to predict whether a customer will acquire a product being offered by the bank or not [14][15]. Predictive analysis has been used using Logistic Regression and AUC as a metric for evaluation of model in the banking sector [16]. A Portuguese banking dataset has been analyzed in [11] to predict the success of telemarketing calls for selling bank long-term deposits to customers. Data mining analysis has used four different types of techniques named Neural Networks, Support Vector Machine, Decision Trees and Logistic Regression [17]. Two metrics that were used for data mining model analysis is AUC and ALIFT (area of the LIFT cumulative curve) [18]. The NN gives the best results. Deep Neural Networks, the most famous and attractive machine learning technique has been used in [19] to predict whether the given customer is suitable for contacting through telemarketing or not. The accuracy achieved by DCNN was 76.70%. The dataset from a U.S.-based catalog direct marketing company was analyzed in [20] and results suggest that Bayesian Network are better in terms of accuracy of prediction and in modeling consumer response. An e-commerce that helps in customer satisfaction and also promotes the sales is proposed in [21] that uses an improved Apriori-based personal recommendation algorithm.

RESEARCH METHODOLOGY

- **Corpus Collection and Specification**

Pakistan's private bank customer corpus has been used to assess and evaluate the performances of

Decision Tree (C 4.5) and Random Forest classifiers. The name of bank cannot be disclosed due to privacy and security concerns. The dataset contains (4000) samples with (6) attributes and no missing values. The facet of data set is composed of two types of attributes: numeric and nominal (categorical) attributes. Numerical: which are Age, Average Balance, Monthly Debit turnover, Monthly Credit turnover and nominal attributes are Category (nature of account), Sector (type of customer either individual, sole proprietor etc.). The class that have to be predicted also consists of nominal attribute named Product. Detailed attribute description is given in Table 1.

Table 1. Bank's Customer Corpus Description

#	Attributes	Kind	Attribute Illustration
1.	Age	Numeric	Age of different customers
2.	Category	Nominal	• 1001 (XYZ RUPEE CURRENT ACCOUNT) • 1026 (XYZ ASAAN CURRENT ACCOUNT) • 6001 (XYZ RUPEE SAVINGS ACCOUNT) • 6002 (XYZ BACHAT ACCOUNT) • 6004 (XYZ KAROBARI MUNAFA ACCOUNT) • 6008 (XYZ BUSINESS PLUS ACCOUNT) • 6012 (XYZ Non Res. Pak ACCOUNT) • 6017 (XYZ KIDS CLUB ACCOUNT) • 6018 (XYZ TEENS CLUB ACCOUNT) • 6025 (XYZ ASAAN SAVINGS ACCOUNT)
3.	Sector	Nominal	• 1001 (Individual) • 105 (Corporate) • 1002 (Staff) • 1003 (Photo Account) • 1006 (Minor) • 1009 (Non – Resident) • 1100 (Sole Proprietor) • 1110 (Partnership) • 1120 (Private Ltd. Company) • 1126 (Government Organization) • 1140 (Trust/Club/NGO)
4.	Monthly Credit Turnover	Numeric	It is the summation of all the credit transactions within a month of a particular account
5.	Monthly Debit Turnover	Numeric	It is the summation of all the debit transactions within a month of a particular account
6.	Average Monthly Balance	Numeric	Average monthly balance sums the closing balance of account at the end of each day and divides it by the number of calendar days in the month.
7.	Product	Nominal(categorical)	Output Variable- which type of TDR has been chosen by the customer. • TDR1 = COII ON MATURITY PAYMENT (Deposit Amount equal or greater than 50 Thousand PKR) • TDR2 = COII ON MONTHLY PAYMENT (Deposit Amount equal or greater than 2 Lac PKR)

The column titled Attributes is illustrating the different classes for each attribute. The age attribute defines the age of customer who has issued a particular Term Deposit. The second attribute, named Monthly Credit Turnover outlines the sum of monthly credit transactions done on an account. Similarly, third attribute name Monthly Debit Turnover sums up the total debit transactions that have been done on an account in the time period of one month. Average balance is the sum of all the End of Day (EOD) balances and dividing it by the number of days in that particular month. The Product attribute defines the category of two types of different Term Deposit Categories.

RESULTS AND ANALYSIS

Classification algorithms work by anticipating the best group to which a new observation appertains to by learning from labeled observations that were already known [22]. Classification algorithms are being used extensively in machine learning tasks: which provide a very good modeling and a best fit for categorizing and assembling the data that have never been seen before and unite them into their various respective groups [23]. The whole process is summarized in the following flowchart.

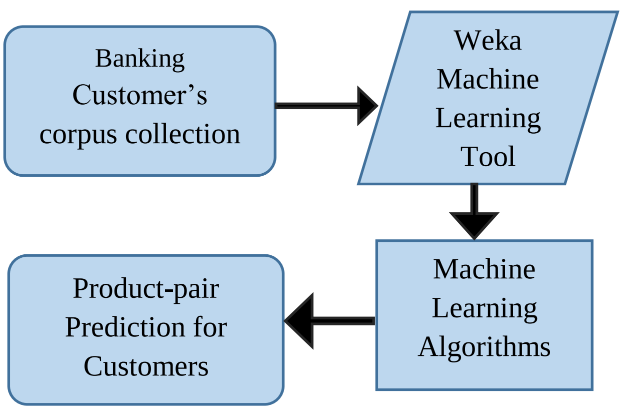

Figure 3. Process Flow Diagram

In this paper, J48 an implementation of Cart 4.5 Decision Tree and Random Forest in WEKA are used as learning classifiers.

J48: Professor Ross Quinlan in 1993 unfold a decision tree algorithm known as Cart 4.5; it epitomizes the result of research that follows back to the ID3 algorithm (which was also proposed by Ross Quinlan in 1986). A greedy approach is implemented using C4.5 algorithms that builds decision tree in top-down recursive divide and conquer vogue. In Top down approach, training set is recursively reduced into smaller subsets as the tree is being built up with a set of labeled training set and their concomitant class labels [24] [25].

RANDOM FOREST: Random Forest Tree (RFT) is an ensemble learner developed by Leo Breiman and formulated by the data mining tool Weka. An unpruned decision tree set is build up by the learner which perform the predictions using the majority voting. Ensemble learners employ randomness in order to produce a robust learner [26] [27].

RF proficiently works on vast informational collections (training data sets) and gives even exactness than different algorithms. RF may create an exceedingly precise classifier for more information sets [28]. RF is much effortlessness and gives a quick learning approach [29].

To predict the best suitable customer-product pair WEKA software is used. Weka is a comprehensive package of machine learning algorithms that was developed at the University of Waikato in New Zealand. Weka uses Java language for the implementation of learning algorithms [30]. Different classification algorithms were applied on corpus from which Decision Tree and Random Forest were found to be the best classifiers. 70% of the dataset is used for training purpose and the rest is used for testing purpose. Figure 1 is the visualization of all attributes in WEKA. And Figure 2 shows the decision tree produced by Cart 4.5 in Weka.

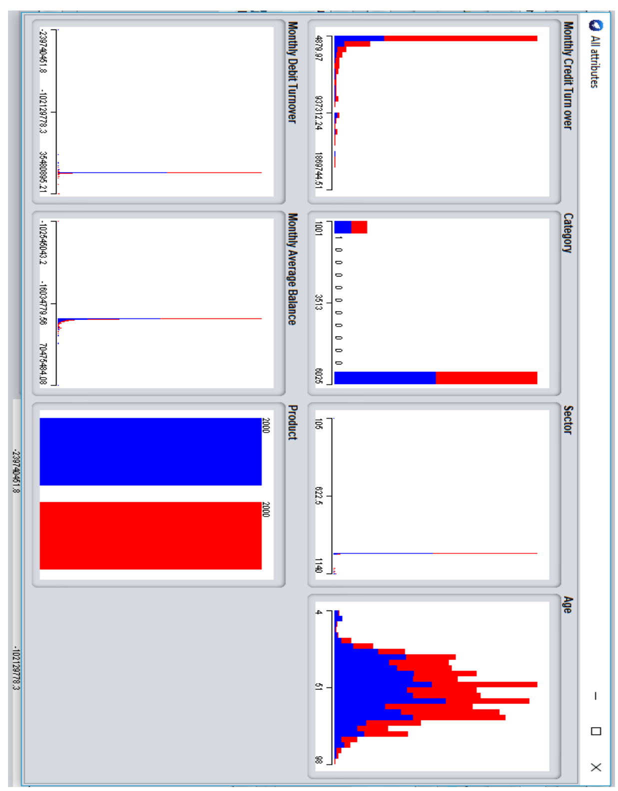

Figure 1. Attribute Visualization

Figure 2. Decision Tree in Weka

DISCUSSION

The performance measures for the classification and prediction are used to evaluate and assess the final result, which are as follows:

- **Confusion Matrix**

The confusion matrix is utilized to measure two-class problem performance for the given data set in Table 1. Instances are correctly classified by the right diagonal elements TP(true positive) and TN(true negative) whereas FP(false positive) and FN(false negative) are the instances that are classified incorrectly [29]. Table 2 and Table 3 is the illustration of Confusion Matrix.

Total number of instances = Accurate classified instance + In Accurate classified instance

Accurate classified instance = TP + TN

Inaccurate classified instance = FP + FN [29]

Table 2. Confusion Matrix

	PREDICTED		
ACTUAL	POSITIVE	TP	FP
	NEGATIVE	FN	TN

Table 3. Confusion Matrix of Decision Tree and Random Forest

DECISION TREE	547	55
	59	539
RANDOM FOREST	567	35
	45	553

- **Accuracy, Specificity and Sensitivity**

The accuracy of classification is implied as the ratio of the correct number of cases classified and is equivalent to the aggregate of TP and TN divided by the aggregate number of cases N [9]. Specificity alludes to the rate of negative correctly classified and is equal to the TN ratio to the TN and FP sum. Sensitivity alludes to the rate of positives correctly classified and is equal to TP divided by TP and FN aggregates. Sensitivity can be known as a true positive measure [9].

Percentages of statistical measures of Decision Tree and Random Forest for test set is shown in Table 4.

Table 4. Percentages of Statistical Measures

	Accuracy	Specificity	Sensitivity
DECISION TREE	90.5%	90.74%	90.26%
RANDOM FOREST	93.33%	94.04%	92.6%

The above results show that Random Forest provides the best classification in terms of accuracy, specificity and sensitivity. RF gives the most accurate prediction of Term Deposit Category (TDR1 and TDR2) suitable for each customer present in the dataset.

In order to determine the importance of features and to know which of the attributed has contributed the most for classification, two methods are used through Weka named CorrelationAttributeEval(CA) and InfoGainAttributeEval(IG).

CorrelationAttributeEval: The evaluation method of CA is done by assessing the attributes/features with respect to the target class attribute [31]. In statistics, correlation is commonly referred to as Pearson's correlation method. Pearson correlation indicates the extent to which two attributes are related linearly to each other through a number that ranges from -1 to 1[32]. CorrelationAttributeEval is one of the technique offered by Weka that uses Pearson's correlation techniques and correlates each of the attribute classes with the target class. CorrelationAttributeEval is used along with the Ranker Search method [33] on banking corpus and the first five attributes ranking are Age as foremost attribute that contribute to the classification followed by Monthly Debit Turnover, Sector, Monthly Credit Turnover and Monthly Average balance.

InfoGainAttributeEval (IG): The importance of each attribute in this method (Information Gain Attribute Evaluation) is calculated through information gain measure with respect to target class. The formula used to calculate InfoGain is as follows [34]:

InfoGain (Class,Attribute) = H(Class) - H(Class | Attribute) [15]

This method provides the ranking with Monthly Debit Turnover, Monthly Credit Turnover, Age, Monthly Average Balance and sector with being the first, second, third, fourth and fifth ranks respectively.

Table 5. Ranking of Features

	1st	2nd	3rd	4th	5th	6th
CorrelationAttributeEval	Age	Monthly Debit Turnover	Sector	Monthly Credit Turnover	Monthly Average Balance	Category
InfoGainAttributeEval	Monthly Debit Turnover	Monthly Credit Turnover	Age	Monthly Average Balance	Sector	Category

Table 5 shows that both of these methods gives the last ranking to customer account's nature i.e. Category attribute evaluating that Category attribute has no significant effect on the TDR Category prediction for the customer. The results also show that Monthly Debit Turnover and Age are the key features and has the most important impact in the TDR Category prediction for customers as these two attribute are common in the top three positions of both the methods used.

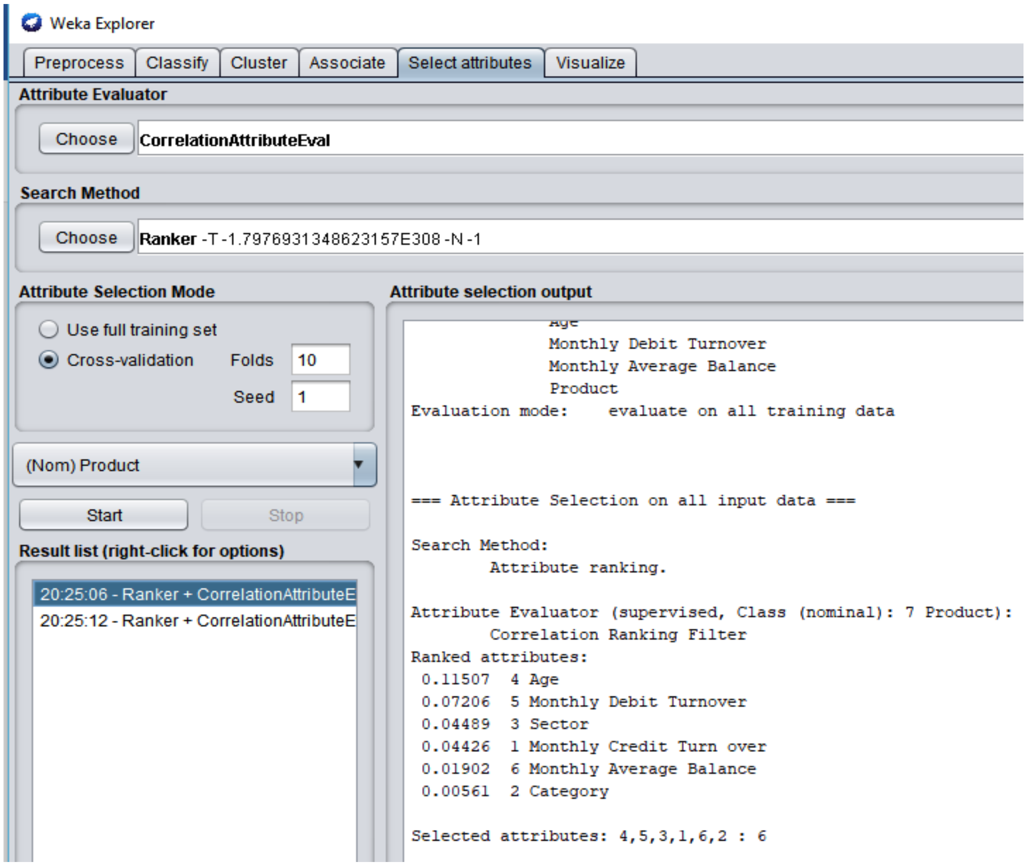

Figure 3. Shows the ranking of attributes with respect to Correlation Attribute Eval method.

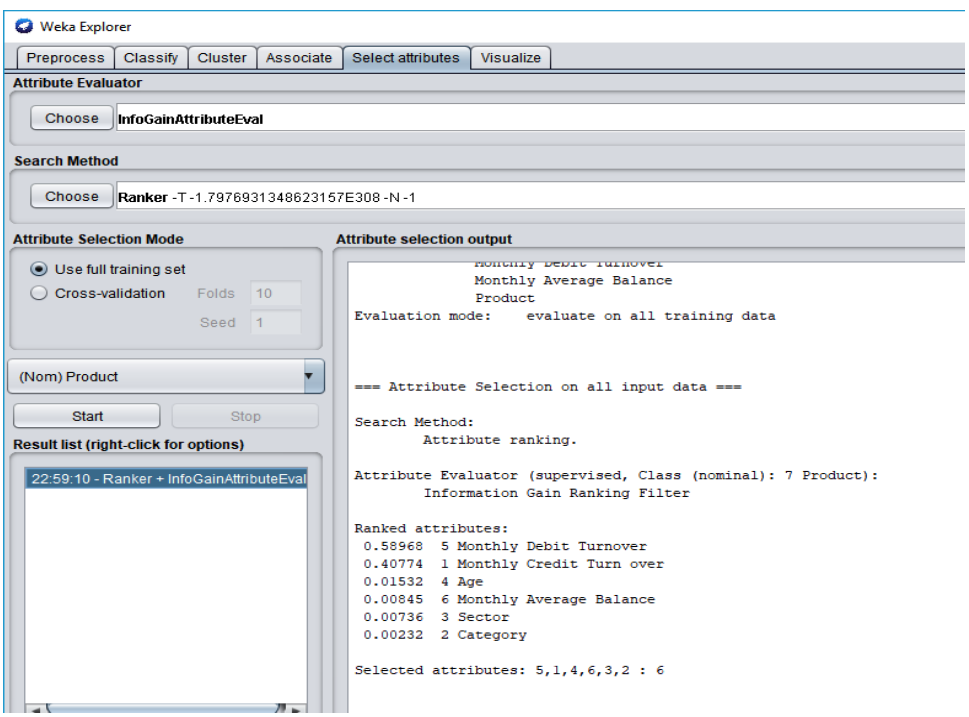

Figure 4. Shows the ranking of attributes with respect to InfoGainAttributeEval.

CONCLUSION

In this research study, a Pakistani bank's corpus was used to predict the best Customer-Product Pair that helps in streamlining the marketing strategies of a bank. This Customer-Product pairing not only helps in minimizing the wastage of bank's resources but also saves time of the marketing people by helping them contacting the customers with the most appropriate products according to respective customer's past month transactional and historical pattern. With the help of machine learning algorithms, Data Analysis of different features (Monthly Credit & Debit turnover, average balance etc.) have been performed to identify the best Term-Deposit Category suitable for a customer. Classification algorithms that were used are, Decision Tree (Cart 4.5) and Random Forest. The classification modelling was done with the help of WEKA tool. Random Forest provides a much better accuracy as compared to Decision Tree. The experimental results clearly show the accuracy of Decision Tree to be 90.5% whereas, of the Random Tree was 93.3%. The main key features identified contributing towards the Term Deposit subscription are Age and Monthly Debit Turnover. For future research, many other different banking products could be analyzed with more features along with other Machine Learning techniques to achieve more accuracy and improve, plan and design the marketing strategies through the help of machine learning.

REFERENCES

1. C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Inf. Sci. (Ny)., vol. 275, pp. 314–347, 2014.
2. X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97–107, 2013.
3. J. C. Paradi and H. Zhu, "A survey on bank branch efficiency and performance research with data envelopment analysis," Omega, vol. 41, no. 1, pp. 61–79, 2013.
4. F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," Big data, vol. 1, no. 1, pp. 51–59, 2013.
5. C. S. Yiu, K. Grant, and D. Edgar, "Factors affecting the adoption of Internet Banking in Hong Kong—implications for the banking sector," Int. J. Inf. Manage., vol. 27, no. 5, pp. 336–351, 2007.
6. H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," IEEE access, vol. 2, pp. 652–687, 2014.
7. Z. Allam and Z. A. Dhunny, "On big data, artificial intelligence and smart cities," Cities, vol. 89, pp. 80–91, 2019.
8. I. Lee and Y. J. Shin, "Machine learning for enterprises: Applications, algorithm selection, and challenges," Bus. Horiz., vol. 63, no. 2, pp. 157–170, 2020.
9. H. A. Elsalamony, "Bank direct marketing analysis of data mining techniques," Int. J. Comput. Appl., vol. 85, no. 7, pp. 12–22, 2014.
10. A. Hashmi, "Application of data mining techniques for occupational stress management," SHRI JAGDISHPRASAD JHABARMAL TIBREWALA UNIVERSITY.
11. S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," Decis. Support Syst., vol. 62, pp. 22–31, 2014, doi: 10.1016/j.dss.2014.03.001.
12. G. Guido, M. I. Prete, S. Miraglia, and I. De Mare, "Targeting direct marketing campaigns by neural networks," J. Mark. Manag., vol. 27, no. 9–10, pp. 992–1006, 2011.
13. J. Asare-Frempong and M. Jayabalan, "Predicting customer response to bank direct telemarketing campaign," 2017 Int. Conf. Eng. Technol. Technopreneurship, ICE2T 2017, vol. 2017-Janua, pp. 1–4, 2017, doi: 10.1109/ICE2T.2017.8215961.
14. C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman, and F. Provost, "Machine learning for targeted display advertising: Transfer learning in action," Mach. Learn., vol. 95, no. 1, pp. 103–127, 2014.
15. R. Ramesh, "Predictive analytics for banking user data using AWS Machine Learning cloud service," Proc. 2017 2nd Int. Conf. Comput. Commun. Technol. ICCCT 2017, no. May 2008, pp. 210–215, 2017, doi: 10.1109/ICCCT2.2017.7972282.
16. A. De Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," Eur. J. Oper. Res., vol. 269, no. 2, pp. 760–772, 2018.
17. B. Kalantar, B. Pradhan, S. A. Naghibi, A. Motevalli, and S. Mansor, "Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN)," Geomatics, Nat. Hazards Risk, vol. 9, no. 1, pp. 49–69, 2018.
18. S. Moro, P. Cortez, and P. Rita, "Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns," Neural Comput. Appl., vol. 26, no. 1, pp. 131–139, 2015.

19. K. H. Kim, C. S. Lee, S. M. Jo, and S. B. Cho, "Predicting the success of bank telemarketing using deep convolutional neural network," Proc. 2015 7th Int. Conf. Soft Comput. Pattern Recognition, SoCPaR 2015, pp. 314–317, 2016, doi: 10.1109/SOCPAR.2015.7492828.
20. G. Cui, M. L. Wong, and H.-K. Lui, "Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming," Manage. Sci., vol. 52, no. 4, pp. 597–612, 2006, doi: 10.1287/mnsc.1060.0514.
21. H. Zhongyi, S. Liangzhong, and C. Shengkai, "An improved apriori-based personal recommendation algorithm for E-commerce," 2008 3rd Int. Conf. Pervasive Comput. Appl. ICPCA08, vol. 1, pp. 60–64, 2008, doi: 10.1109/ICPCA.2008.4783649.
22. D. P. Wall, R. Dally, R. Luyster, J.-Y. Jung, and T. F. DeLuca, "Use of artificial intelligence to shorten the behavioral diagnosis of autism," PLoS One, vol. 7, no. 8, p. e43855, 2012.
23. S. A. Ali, A. Khan, and N. Bashir, "Analyzing the Impact of Prosodic Feature (Pitch) on Learning Classifiers for Speech Emotion Corpus," Int. J. Inf. Technol. Comput. Sci., vol. 7, no. 2, pp. 54–59, 2015, doi: 10.5815/ijitcs.2015.02.07.
24. L. Rokach, "Decomposition methodology for classification tasks: a meta decomposer framework," Pattern Anal. Appl., vol. 9, no. 2–3, pp. 257–271, 2006.
25. M. S. Kumar and B. Indrani, "Frequent rule reduction for phishing URL classification using fuzzy deep neural network model," Iran J. Comput. Sci., pp. 1–9, 2020.
26. J. Tanha, M. van Someren, and H. Afsarmanesh, "Semi-supervised self-training for decision tree classifiers," Int. J. Mach. Learn. Cybern., vol. 8, no. 1, pp. 355–370, 2017.
27. Y. Beggah, N. Tabet, and R. J. Tarento, "Random Forests and Decision Trees," IJCSI Int. J. Comput. Sci., vol. 9, no. 5, pp. 272–278, 2012, doi: 10.1016/0921-5107(94)90307-7.
28. A. Haldorai and U. Kandaswamy, "Supervised Machine Learning Techniques in Intelligent Network Handovers," in Intelligent Spectrum Handovers in Cognitive Radio Networks, Springer, 2019, pp. 135–154.
29. D. L. Gupta, A. K. Malviya, and S. Singh, "Performance Analysis of Classification Tree Learning Algorithms," Int. J., vol. 55, no. 6, pp. 39–44, 2012, doi: 10.5120/8762-2680.
30. "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." .
31. S. Gnanambal, "Classification Algorithms with Attribute Selection : an evaluation study using WEKA," vol. 3644, pp. 3640–3644, 2018.
32. U. Loobuyck, "Constructiveness-Based Product Review Classification." 2020.
33. "How to Perform Feature Selection With Machine Learning Data in Weka." .
34. T. T. S. Nguyen and P. M. T. Do, "Classification optimization for training a large dataset with Na{\"\i}ve Bayes," J. Comb. Optim., pp. 1–29, 2020.