Research on Precision Marketing based on Big Data Analysis and Machine Learning: Case Study of Morocco

Nouhaila El Koufi, Abdessamad Belangour, Mounir Sdiq
Laboratory of Information Technology and Modeling (LTIM)
Faculty of Sciences, Ben M'sik, Hassan II University, Casablanca, Morocco

***Abstract*—With the growth of the Internet industry and the informatization of services, online services and transactions have become the mainstream method used by clients and companies. How to attract potential customers and keep up with the Big Data era are the important challenges and issues for the banking sector. With the development of artificial intelligence and machine learning, it has become possible to identify potential customers and provide personalized recommendations based on transactional data to realize precision marketing in banking. The current study aims to provide a potential customer's prediction algorithm (PCPA) to predict potential clients using big data analysis and machine learning techniques. Our proposed methodology consists of five stages: data preprocessing, feature selection using Grid search algorithm, data splitting into two parts train and test set with the ratio of 80% and 20% respectively, modeling, evaluations of results using confusion matrix. According to the obtained results, the accuracy of the final model is the highest (98.9%). The dataset used in this research about banking customers has been collected from a Moroccan bank. It contains 6000 records, 14 predictor variables, and one outcome variable.**

***Keywords*—Precision marketing; big data analysis; machine learning; potential customers prediction algorithm (PCPA)**

I. Introduction

After entering the 21st century, the world witnessed a rapid growth in internet technology which resulted in the development of online transactions and the change in consumer habits and consumption patterns. The traditional model of bank marketing seriously hinders the development of the bank sector due to the lack of understanding of clients' needs, inability to adapt to the modern market characteristics, and the lack personalized recommendations for quality customers. However, Big data analysis and machine learning techniques applied in precision marketing will provide ideal results to traditional marketing. Therefore, the banking industries need to change their old marketing model toward a precision marketing model based on big data analysis, artificial intelligence (AI), and modern information technology to realize long-term development.

The advancement of big data technology has also contributed to the acceleration of transformation from traditional banking to modern and digital banking. Meanwhile, machine learning techniques and big data analysis technology play an essential role in the precision marketing of banking services. Machine learning techniques provide the possibility to extract facts, information, and patterns about customers from the large amount of data gathered in banks [1]. Analyzing and extracting knowledge from such data can provide a decision-making foundation for banking companies.

The marketing strategy is a long-term scheme that principally studied the marketing and market environment to understand market opportunities and meet the customers' needs. It covers everything from the study of the situations confronted by enterprises marketing under current market conditions and competitions, the choice of customers, and the channel of communication between companies and clients. Whereas precision marketing refers to a marketing strategy that has a clear focus, it targets the consumers that have a great willingness to consume. Based on modern information technology, precision marketing precise the position accurately to build personalized communication between enterprises and customers so that companies can realize long-term development and maximize wealth.

Theoretically, precision marketing is a marketing strategy that aims to understand customers well and their actual needs. Practically, based on big data analysis, machine learning techniques, and modern technology, precision marketing can render the prediction reasoning nearest to customers' needs. The principle of precision marketing was first proposed back in 1999 by Lester. In 2004, the 4R rule of precision marketing was declared by Brebach and Zabin. Then, Philip Kotler gave a clear introduction to precision marketing. Jin et al. [2] have stated that companies can achieve high sales performance by adapting precision marketing.

Precision marketing can not only achieve high sales and decrease the purchase cost of customers but also help companies to build a loyal customer base. The modern bank is customer-centric. It focuses on how understanding customers' behaviors to provide clients with accurate and quality services via a digital operation technologies chain. Banks not only need to realize deals with customers but also need to maximize wealth and ensure return rates last long. Thanks to the fast development of information technologies, precision marketing realize the goal of enterprises to understand clients and provide them with the right product at the right time and via appropriate methods. Precision marketing contains four parts: target customer, right message and channel, and a good time.

The main objective of this paper is to propose a new precision marketing model based on big data analysis and machine learning techniques to identify potential customers. The rest of this paper is organized as follows: we review previous related works in the next section. In Section III, we present and discuss the proposed methodology. Section IV is reserved to illustrating and analyzing our simulation results and performance evaluation. Finally, the conclusion of our paper is in Section V.

II. Related Work

A. Literature Review

We present a summary of what has been proposed in precision marketing based on machine learning and big data by renowned researchers. In this subsection, we have reviewed various researches related to our work.

Chiu et al. [3] have proposed an Omni-channel Chabot that merges IOS, Android, and Web components. Based on convolutional neural networks (CNNs), the proposed chatbot can provide personalized service and precision marketing. In order to show the advantages of the new method, a case study of a shared kitchen is utilized, which can be employed for other consumer applications like personalized services and clothing selection.

Zhang et al. [4] developed a predictive model to forecast high potential luxury car buyers using car owners' and telecom users' data. They combined two machine learning algorithms, logistic regression, and neural network mining. A case study of a traffic management department and telecom operators in a medium-growth city in china is used to demonstrate the efficiency of the new model.

In [5], Xia Liu has compared the CNN model, LSTM model, LSTM attention model, and CNN + LSTM attention. As a result, the performance of the CNN + LSTM attention model and the LSTM attention model is better, with the highest overall accuracy of testing and training. The model is applied to precision marketing for obtaining precision consumer portraits'.

In [6], Xie et al. have used Hamming distance classification algorithm and BP Neural network to classify clients using purchase data to release the purpose of precision marketing. In order to solve the problem of the chosen transportation program, the related decision objectives, and location selection. Xiao et al. [7] built a precision marketing optimization strategy using neural network modeling and the fuzzy method.

In [8], Chong et al. adopted neural networking for consumer product demand prediction. In order to show the advantage of the proposed solution, an electronic data from Amazon.com was used. This data is about promotional marketing information and online reviews.

Tang et al. [9] proposed an advanced K-means clustering algorithm to provide an accurate customer division. Scholars have applied this new method in the precision marketing of ETC credit cards to precisely detect potential ETC users. To build a precision marketing device for forecasting the cumulative number of voice app users for China Mobile communications, Yan et al. [10] adopted ARIMA modeling. The proposed method gives a reference to understanding its market demand.

The studies in precision marketing based on Big data analysis and machine learning are fewer in the banking field. The related research of machine learning in the banking sector mainly focuses on banking crisis prediction and forecasting customer churn. In [11], Jessica et al. adopted artificial neural network modeling that predicts the short-term financial distress for Spanish banking. A predictive model based on Big data analysis and crisis Index to predict the banking crisis was proposed in [12]. Zizi et al. [13] used logistic regression and neural network modeling based on financial indicators to forecast financial distress in the bank sector.

In this study, we propose a new precision marketing model based on banking big data to forecast potential target customers and realize accurate marketing of bank housing loans. The proposed PCPA algorithm is based on big data and machine learning technology. The process of PCPA consists of five steps: data selection and understanding, data cleaning and filtering, feature selection, data modeling, and results evaluation. After a comparison of various famous machine learning methods [14], we chose XGBoost as the central algorithm of PCPA. Moreover, we have preprocessed the data and extracted the significant features using the Grid search algorithm for well understanding of the data.

III. Proposed Work

A. System Architecture

In this subsection, a pictorial representation of the whole process of PCPA is illustrated in Fig. 1. The process of PCPA consists of six phases: data selection and understanding, data cleaning and filtering, feature selection, splitting the prepared data, modeling, and results evaluation.

B. Description of Proposed Model

1) Data selection and understanding: The dataset used in this study is from a Moroccan bank, Attawfiq Micro-Finance. The data consists of 6000 records and 14 variables; it contains demographic information and information about customers' behaviors. The provided data are from 2018 to 2021. For building a suitable predictive model, a selection of important information from data was established using advanced data variance analysis and analyzing attributes correlation (using Correlation matrix).

Each case from the final dataset is represented by thirteen variables used as input to the proposed model and one categorical variable that represents whether the customer has applied for a housing loan (Housing_credit) as output. Table I represents the attributes of the data. We represent the coded of some indicators in Table II.

2) Data cleaning and filtering: The data cleaning and filtering stage consists of cleaning the selected data from missing values, outliers, noise, etc. This phase is eliminatory for reducing the dimensions of the dataset and reducing the time of required computation. In order to extract deeper insights, we have considered the data visualization in this methodology.

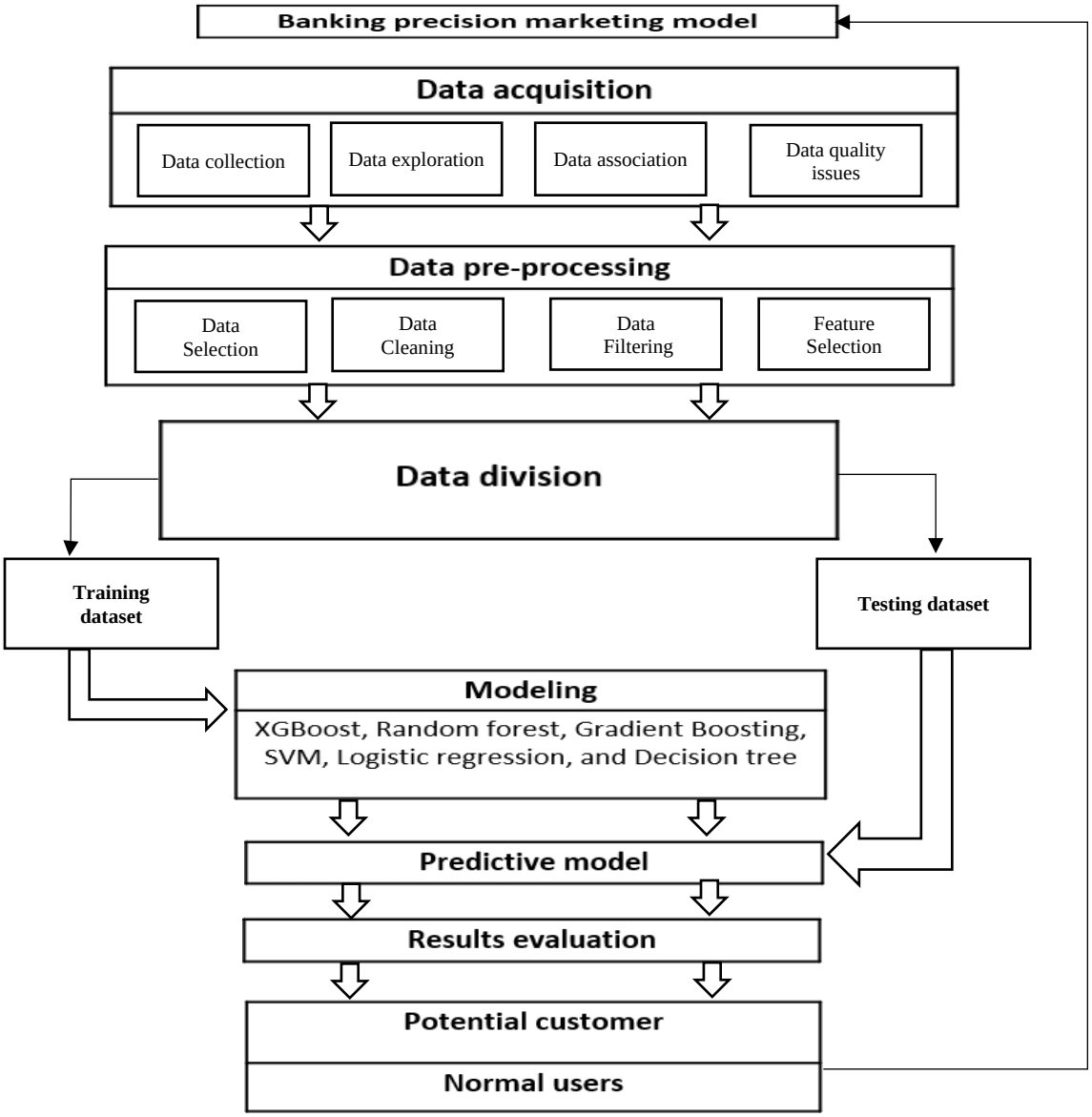

Fig. 1. PCPA Architecture.

3) Feature selection: The feature selection stage plays an essential role in extracting the critical features and removing the non-significant attributes from the dataset. Also, it contributes to the improvement of model performance and reduces the time of training and validation. Firstly a uni-variate selection is applied then, Grid search (GS) method is used to choose the optimal parameter.

4) Modeling: As we declared above, our proposed PCPA is based on machine learning techniques and big data. In this phase, we have applied five machine learning algorithms, namely, XGBoost, Random forest, Gradient Boosting, SVM, Logistic regression, and Decision tree. Then after comparison of the accuracy and performance of the five models, we choose XGBoost as the central algorithm of PCPA because it illustrates the most excellent accuracy and performance. The results of all the models are presented in Section IV.

5) Results evaluation:

a) Confusion matrix: In order to evaluate and measure the performance of the predictive models for forecasting the potential customers correctly, we have applied different metrics: namely, precision, recall, accuracy, and F–measure. The calculation of these four measures depends on information extracted using the confusion matrix. After the prediction, the confusion matrix analyzes the value of accurate and wrong predictions. In Table III, the representation of the confusion matrix is presented.

TP: These are the correctly predicted negative values which mean that the value of the actual class is potential, and the value of the predicted class is also potential customers.

TN: These are the correctly predicted negative values which mean that the value of the actual class is non-potential and the value of the predicted class is also non-potential customers.

FP: The number of non-potential customers, but the predictive model has forecasted them incorrectly (as potential).

FN: The number of potential customers, but the predictive model has forecasted them incorrectly (as non-potential).

TABLE I. CONFUSION MATRIX

	Potential	***Non-Potential***
Potential	TP	FN
Non-Potential	FP	TN

b) Performance indicators

Accuracy: The proportion of the number of all right predictions is known as accuracy. It is calculated by the following formula:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (1)$$

Precision: It is the proportion of true positives to all positives. It is calculated by the following formula:

$$\Pr ecision = \frac{TP}{TP + FP} \quad (2)$$

Recall: It is used to calculate the real positive rate. It is calculated by the following formula:

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad (3)$$

F1-measure: The weighted harmonic average of precision and recall is called F1-measure.

$$F1 - measure = \frac{(2 \times \Pr ecision \times \mathrm{Re}call)}{(\Pr ecision + \mathrm{Re}call)} \quad (4)$$

TABLE II. CHARACTERISTICS OF BANK CUSTOMERS

Attribute	***Possible Value***	***Type***
Age	*20-60 year*	*Numerical*
Marital status	*0-3*	*Categorical*
Income(dh)	*1-5*	*Numerical*
Education level	*0-6*	*Categorical*
Housing	*0-3*	*Categorical*
Family_numbrs	*1-6*	*Numerical*
Gender	*0-1*	*Categorical*
Experience	*0-42*	*Numerical*
Lodgment_situation	*0-1*	*Categorical*
Credit Card	*0-1*	*Categorical*
ASCC	*0-9,30*	*Numerical*
Securities_account	*0-1*	*Categorical*
Online_bank_service	*0-1*	*Categorical*
Housing_credit	*0-1*	*Categorical*

TABLE III. DESIGNATION OF SOME INDICATORS

Attribute	***Attribute Value***
Marital status	*0—Single* *1—Married* *2—Divorced*
Income(dh)	*1—[0dh-3000dh]* *2—[3001dh-5000dh]* *3—[5001dh-8000dh]* *4—[8001dh-10000dh]* *5—10000dh>>*
Education level	*0—Baccalaureate* *1— Baccalaureate+2* *2— Baccalaureate+3* *3— Baccalaureate+5* *4— Baccalaureate+8* *5—No diploma*
Property_Area	*0—Urban* *1—Semi-Urban* *2—Rural*
Gender	*0 — Female* *1— Male*
Lodgment_situation	*0 — Personal house* *1— Parent house* *3— House renter*
Credit Card	*0 — No* *1— Yes*
Securities_account	*0 — No* *1— Yes*
Online_bank_service	*0 — No* *1— Yes*
Housing_credit	*0 — No* *1— Yes*

C. PCPA Algorithm

Algorithm 1: Proposed algorithm for Potential Customers prediction

Input: The training dataset consisting of input features such as xi and output label y;
Output: Predicted labels; Potential or not potential
Procedure;

1. **Data selection and understanding;**
2. **Data cleaning and filtering;**
3. **Feature selection using Grid search algorithm;**
4. **Data modeling using XGBoost, Random forest, Gradient Boosting, SVM, Logistic regression, Decision tree;**
5. **Results Evaluation (confusion matrix);**

IV. RESULTS AND DISCUSSION

The PCPA is based on machine learning technology and big data. First, we split the data into two parts train with a proportion of 80% and a test set with a ratio of 20%. We tested the pre-processed data on various famous classification algorithms that we have chosen based on our previous study [14], such as XGBoost, Random forest, Gradient Boosting, SVM, Logistic regression, and Decision tree. Then we compared the performance of the selected machine learning methods based on performance indicators (Accuracy, Precision, Recall, F-measure, and Cross-validation score). We presented the findings in Table IV. According to the comparison of the performance of the six machine learning methods illustrated in Table IV and Fig. 2, the RF and GBDT have the highest accuracy (98, 9%). RF achieved good precision with 93, 6%, Recall of 93,5%, and an F-measure of 93,5%. Another model which illustrates good results is

XGBoost. It reached a good accuracy with 98,1%, F-measure with 90,8%, Cross-validation score of 90,1%, and highest Recall of 95,4%. Also, the SVM gave significant results, registered a good accuracy of 97, 1%, Recall of 88,7%. We can notice that the GBDT shows a better performance amount the other models in terms of Accuracy, Precision, Recall, F-Measure, and Cross-validation score. It accomplished the highest accuracy compared to other i.e. 98, 9%, precision with 96,3%, F-measure 92,9%, Cross-validation of 94,2% with good Recall 89,7%. Hence Gradient Boosting outperforms the other ML methods used in our work. This is why, we choose Gradient Boosting as the central algorithm of PCPA.

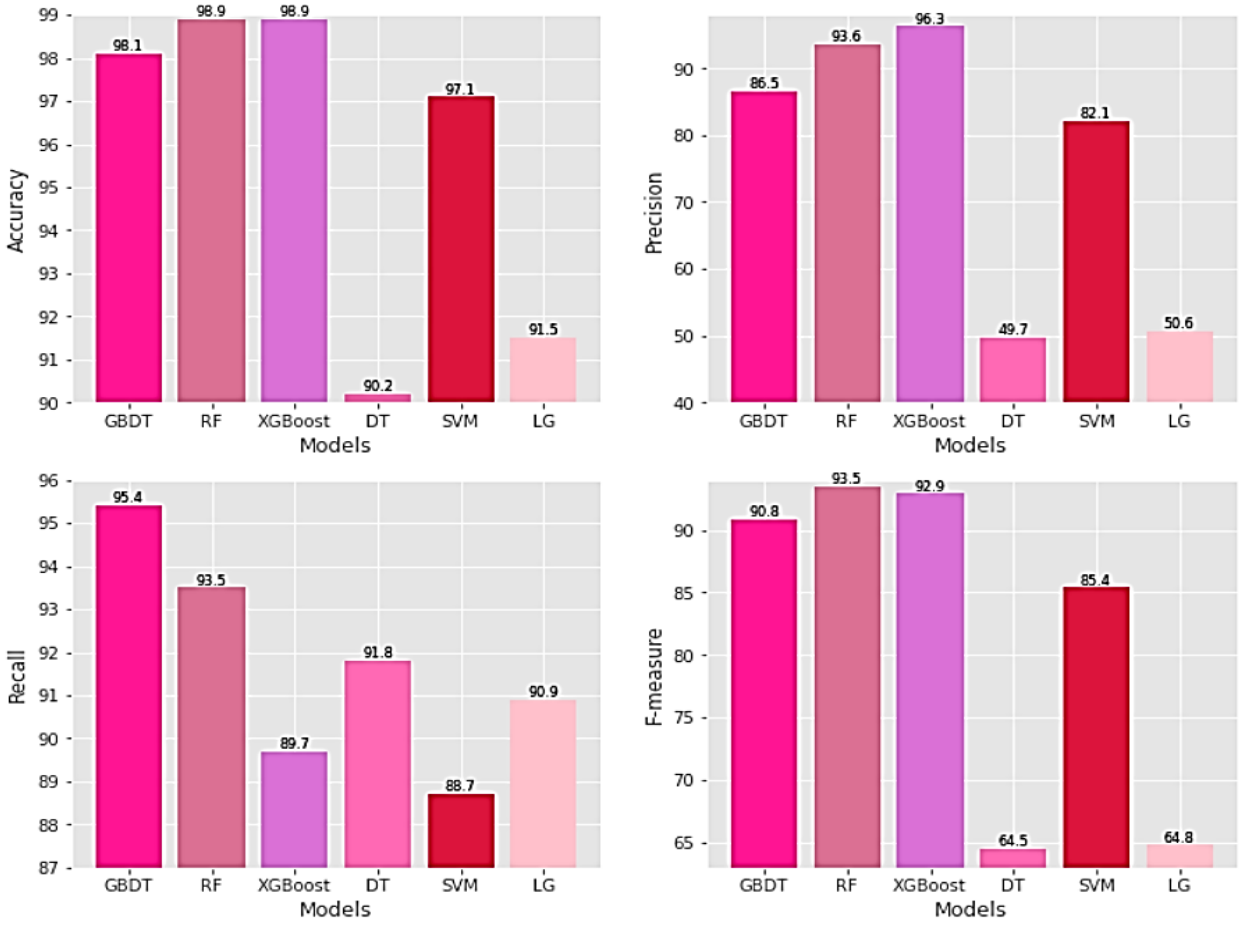

Fig. 2. Evaluation of Models on Performance Indicators (Accuracy, Recall, Precision, F-measure).

TABLE IV. COMPARISON OF MACHINE LEARNING MODELS

Model	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F-Measure (%)*	*Cross Validation (%)*
XGBOOST	**98,1**	**86,5**	**95,4**	**90,8**	**90,1**
RF	**98,9**	**93,6**	**93,5**	**93,5**	**91,8**
GBDT	**98,9**	**96,3**	**89,7**	**92,9**	**94,2**
DT	**90,2**	**49,7**	**91,8**	**64,5**	**61,2**
SVM	**97,1**	**82,1**	**88,7**	**85,4**	**88,1**
LG	**91,5**	**50,6**	**90,9**	**64,8**	**65,3**

V. CONCLUSION

In this study, we have proposed a PCPA algorithm for precision marketing based on machine learning and Big data analysis. We have used a dataset collected from Moroccan banking. Our model predicts potential housing loan customers from whole banking users. This prediction can help the marketing department to target quality customers at a low cost and fast time. At the same time, we compared the efficiency of different famous machine learning approaches, which are: XGBoost, Random forest, Gradient Boosting, SVM, Logistic regression, and Decision tree in terms of accuracy, F-measure, recall, precision. The results illustrate that Gradient Boosting achieved better performance amount the other ML methods used in terms of accuracy, F-measure, recall, precision, and cross-validation score. Hence we choose Gradient Boosting as the central algorithm of PCPA.

REFERENCES

[1] C. X. Yu, Z. X. Min, M. Ying and G. Feng, “Research Progress and Trend of the Machine Learning based on Fusion,” International Journal of Advanced Computer Science and Applications(IJACSA), vol. 13, pp. 1-7, July 2022.

[2] H. Jin, C. Chi and X. Gao, “Strategic Research on Accurate Marketing to Enhance Consumer Experience of Social Media Users,” 2nd International Conference on Economic Development and Education Management (ICEDEM 2018), Atlantis Press, vol. 290, pp. 455-458, December 2018.

[3] M. C. Chiu and K. H. Chuang, “Applying transfer learning to achieve precision marketing in an omni-channel system–a case study of a sharing

kitchen platform," International Journal of Production Research, vol. 59, pp. 7594-7609, January 2021.

[4] H. Zhang, L. Zhang, X. Cheng and W. Chen, "A novel precision marketing model based on telecom big data analysis for luxury cars," International Symposium on Communications and Information Technologies (ISCIT), IEEE, Qingdao, China, vol. 59, pp. 307-311, November 2016.

[5] X. Liu, "E-Commerce Precision Marketing Model Based on Convolutional Neural Network," Scientific Programming, vol. 2022, pp. 1-11, March 2022.

[6] Y. Xie, X. Liu, Y. Wen and Y. Xiao, "Precision Marketing Based on Hamming Distance Classification Algorithm," Computer Science and Application, vol. 9, pp. 1403-1406, September 2018.

[7] K. Xiao and X. Hu, "Study on maritime logistics warehousing center model and precision marketing strategy optimization based on fuzzy method and neural network model," Polish Maritime Research, vol. 24, pp. 30-38, September 2017.

[8] X. Tang, C. Cheng and L. Xu, "Research and Application of Precision Marketing Algorithms for ETC Credit Card Based on Telecom Big Data," Signal and Information Processing, Networking and Computers. Springer, Singapore, vol. 677, pp. 1075-1084, December 2020.

[9] K. Xiao and X. Hu, "Research and Application of Precision Marketing Algorithms for ETC Credit Card Based on Telecom Big Data Signal and Information Processing, Networking and Computers. Springer, vol. 677, pp. 1075-1084, December 2021.

[10] B. Yan and Z. Chen, "A prediction approach for precise marketing based on ARIMA-ARCH Model: A case of China Mobile," Communications in Statistics-Theory and Methods, vol. 47, pp. 4042-4058, January 2018.

[11] J. Paule-Vianez, M. Gutiérrez-Fernández and J. L. Coca-Pérez, "Prediction of financial distress in the Spanish banking system: An application using artificial neural networks," Applied Economic Analysis, vol. 28, pp. 69-87, December 2019.

[12] M. Musdholifah, U. Hartono and Y. Wulandari, "Banking crisis prediction: emerging crisis determinants in Indonesian banks," International Journal of Economics and Financial, vol. 10, pp. 124, February 2020.

[13] Y. Zizi, A. Jamali-Alaoui, B. El Goumi, M. Oudgou and A. El Moudden, "An optimal model of financial distress prediction: A comparative study between neural networks and logistic regression," Risks, vol. 9, pp. 200, November 2021.

[14] N. El Koufi, A. Belangour, A. El Koufi and M. Sadiq, "A systematic literature review of machine learning techniques applied to precision marketing," Technical and Physical Problems of Engineering (IJTPE), vol. 14, pp. 104-110, September 2022.