Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 8, Issue. 3, March 2019, pg.49 – 60

Hybrid Datamining Approaches to Predict Success of Bank Telemarketing

Anas Nabeel Falih AL-Shawi
Student nom: 173104662
Email: anoalshawi@gmail.com
Advisor: Asst.Prof. Sefer Kurnaz
GRADUATE SCHOOL OF SCIENCES ENGINEERING, ISTANBUL ALTINBAS UNIVERSITY, TURKEY

Abstract: Telemarketing is a kind of straightforward marketing in which salesman requests the consumer either face to face or telephone request and influence him to purchase the product. Telemarketing achieves most prevalence in the 20th century and still increasing it. Now, the phone has been broadly accepted. It is valued efficient and holds the consumers up to date. In the Banking area, marketing is the backbone to exchange its goods or service. Business promotion and marketing is frequently based on an exhaustive understanding of actual information about the market and the real client demands for the productive bank manner. We recommend a data mining (DM) method to foretell the achievement of telemarketing requests for contracting long-term bank deposits. A local Portuguese bank was labeled, with data gathered from 2011 to 2016, thus involving the effects of the current economic crisis. We examined a comprehensive set of 11 features associated with bank consumer, goods and social-economic characteristics. We also discuss four DM forms with the hybrid model: Naïve Bayes (NB), Decision Trees (DTs), Perceptron Neural Network (NN) and Support Vector Machine (SVM). The four types were tested and compared with proposed hybrid classification methods (Perceptron Neural Network + Decision Tree) on an evaluation set, and we are splitting data into training and testing sets using cross-validation method. The proposed hybrid classification technique presented the best results (Precision 99% and ROC = 97%).
Keywords: Data mining, Decision Tree, k-means, Support Vector Machine, bank telemarketing and neural network.

1. Introduction

Marketing is a procedure of detecting the destination consumers to buy or make a deal with a product via fitting systems. It presently promotes the process to purchase the goods or service and even assists in planning the necessary for the product and convince customers to purchase it. The overall purpose is to enhance the selling of goods and services for the industry, marketing, and commercial institutions. It also accommodates to preserve the status of the business [1].

Telemarketing is a kind of straightforward marketing. Telemarketing achieves most prevalence in the 20th century and still increasing it. Now, the phone has been broadly accepted. It is valued efficient and holds the consumers up to date [1].

1.1 Decision Support System

To obtain the best choices in organizational processes are sometimes established numerous difficulty where the quality of choice concerns. Decision Support Systems (DSS) are organized as a collection of automated events and terms that maintains the system or administration into their decision-making activities. The idea of DSS arises from stability which occupies between the data produced by the workstation and the decision of a human [2].

DSS applies statistical and mathematical processes to defeat the losses in data or knowledge and assists the decision producers to choose the best decision. Data mining (DM) represents a necessary function to maintain the DSS which are based on the data collected from the data mining forms: controls, guides, and connection. Data mining is the method of choosing, learning, and forming a large quantity of data and interpret undiscovered patterns. The purpose of data mining in DSS is to recommend a mechanism which is regularly obtainable for the market users to examine the data mining patterns [9].

A technology utilized inside the DSS is Machine learning (ML) that links data and learning on it to correctly foretelling the decisions. The basic principle of ML is to assemble the methods that can get input data and then forecast the results or outputs by applying the statistical interpretation within enough interval. ML allows the DSS to gain new experience which supports it to make the right decisions [2].

Machine Learning can be essentially distinguished into two categories, i.e. supervised training and unsupervised training. In supervised training, the producing of the algorithm is previously known, and we utilize the information data to foretell the result. Examples of supervised training are regression and classification. In opposite, unsupervised training we have input information whereas no same result variables are chosen. The form of unsupervised training is clustering.

Feature selection is the method of choosing the subset of relevant variables from the total features or patterns. It recognizes the most influential properties which accommodate to foretell the output. By using this method, we can decrease the dimensionality of properties, limit the scheme from overfitting and decrease the training time. In this process, the parsimonious form can be accomplished with a smallest amount of parameter with good performance in minimum time [3].

1.2 Paper Contributions

Inside our paper, we have focused on data mining classification methods these can forecast a certain consequence based on a specified input. We have utilized four classifiers and create comparative study to analyze a bank telemarketing dataset that recorded previously to take a decision. The main contribution of this study, apply hybrid techniques (Neural network and Decision tree) in a proposed framework which had a highest accuracy to classify client's records.

The full model which is used in this study consists of 20 variables. Feature selection approach has been used to select the best subsets of variables and then different type of classification algorithms have been utilized to check their accuracy and performance. A number of trials have been constructed to compare the accuracy of the implemented classifiers on a different size full training dataset with 11 attributes. Results showed that hybrid classification in proposed framework outperforms other classifiers with an accuracy rate of 97%, which provided a more effective and comprehensive classification mechanism than other classification techniques.

1.3 Paper Structure

In the first section, introduction of decision support systems and data mining. The literature review which state the previous researches in the field of classification is presented in section II. Section III discusses the research methodology, data and variables. Comparative study introduced in section IV. The results and discussions are obtainable into section V. The final conclusions including later works are offered in section VI.

2. Related Work

This part describes the earlier study work which has been previously made in classification using ML methods. The data which is applied in this investigation work is the data of consumers of a Portuguese banking organization. In manuscript [5], this research proposed to obtain the form that can enhance the achievement rate of telemarketing for the bank. The statistical and analytical procedures of data mining which have been applied in their study are SVM, DT, and Naive Bayes. The achievement of these procedures was examined through the Receiver Operator Characteristics (ROC) curve. Between all these analytical procedures, DT comes up with the various useful effects.

In study [6], the investigation aimed to foretell the achievement of bank telemarketing. The data set which they applied in their study was included of 150 properties and is a comprehensive dataset of the term from 2008 to 2013. They examine four data mining procedures, i.e. Logistic Regression (LR), SVM, and ANN. The ANN achieved the highest result.

In the study [7], authors analyzed the fault diagnosis system for reciprocating compressors. Data was taken from oil corporation (5 years of operational data) and utilizes the SVM to examine it. They come up with the outputs that SVM correctly foretells the 80% correct classification of the possible mistakes in the compressor.

In a study [8], authors researched to foretell the growing of bankruptcy by using the SVM and RF techniques. The input was received from the Salomon Center database regarding North American from the period 1985 to 2013 with notes of more than 10,000. After implementing SVM and RF procedures, they analyze the outputs using discriminant analysis and logistic regression.

3. Proposed Framework

3.1 Methodology

In this part, the block design of the recommended mechanism is presented in Figure 1. We moved within two stages toward developing the recommended mechanism: data preprocessing and data classification processing. A separate subsection is dedicated to each stage.

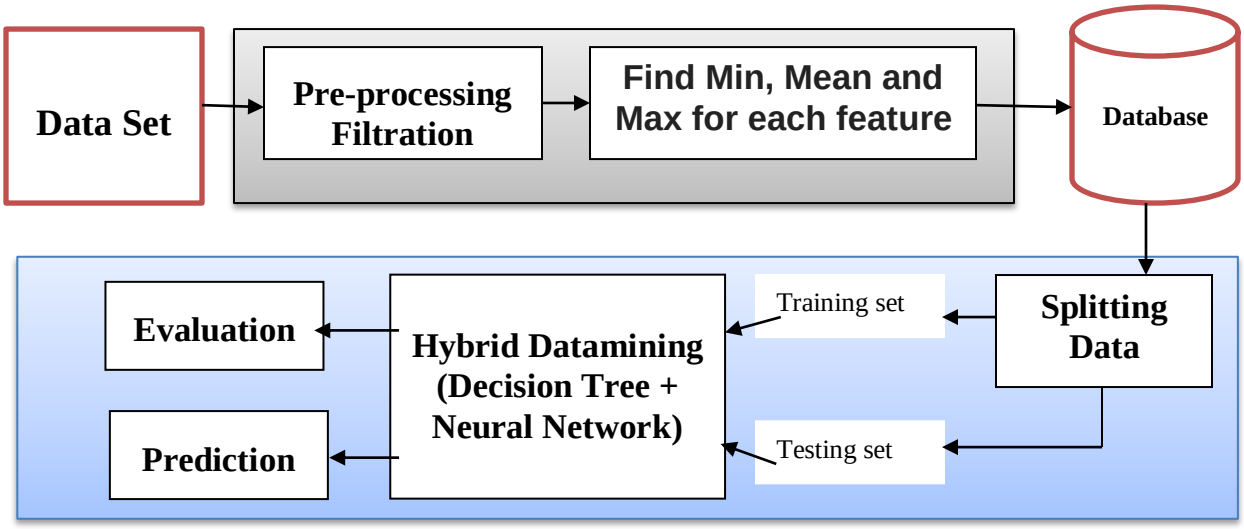

Figure 1 - The block diagram of the proposed tool

Preprocessing (Steps)

1. Filter data (extract each feature) and remove incomplete records
2. Find minimum, average and maximum for each feature
3. Insert each record in database

Data processing (Steps)

1. Divide data into training set and test set
2. Apply different datamining techniques using cross validation
3. Apply hybrid of two best techniques according to accuracy

3.2 Dataset description

A local Portuguese bank was labeled, with data gathered from 2011 to 2016, thus involving the effects of the current economic crisis. We examined a comprehensive set of 11 features associated with bank consumer, goods and social-economic characteristics.

3.3 Attribute Information:

1. Age (numeric)
2. Job: type of job (employed','services','student','technician','unknown', ...)
3. Marital: marital status ('divorced','married','single','unknown';)
4. Education ('high. school','illiterate','professional,'university,'unknown'
5. Default: has credit in default? (categorical: 'no','yes','unknown')
6. Housing: has housing loan? (categorical: 'no','yes','unknown')
7. Loan: has personal loan? (categorical: 'no','yes','unknown')
8. Contact: contact communication type (categorical: 'cellular','telephone')
9. Month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10. Day_of_week: last contact day of the week ('mon','tue','wed','thu','fri')
11. Duration: last contact duration, in seconds (numeric). # other attributes:

3.4 Hybrid Classification Processing Stage

A. Sampling or Splitting

Various classifiers elective involves distributing the training data in decreased equal subsections from data using the cross-validation procedure. Cross-validation, seldom called rotation evaluation, or out-of-sample measurement is any of several related form validation procedures for evaluating how the outcomes of the mathematical interpretation will conclude to an independent data set [10].

B. Perceptron Neural Network

Perceptron is a method for supervised training of binary classifiers. A binary classifier is a purpose which can determine whether an input, described by a vector of integers, refers to some particular class. It is a kind of linear classifier, i.e., a classification method that presents its forecasts based on a linear predictor function linking a set of measurements with the feature vector. [3].

C. Decision Tree

In the decision tree method, we need to pick the excruciating feature that reduces the value from entropy and exploiting the Information Gain. To recognize an excruciating feature from the Decision Tree, should compute the Information Gain to every feature also then choose a feature that exploits an Information Gain.

$$E = \sum_{i=1}^{k} -P_i \, log_2 \, P_i \qquad 1)$$

Where

- k is that value from classes of this objective feature
- Pi is a value of incidences from class i separated via the whole value from occurrences

4. Comparative Study

4.1 Naïve Bayes

Table 1 - Naïve Bayes Result

Result	Values
Correctly Classified Instances	946
In Correct Classified Instances	54
Precision	94 %
Recall	95 %
F-measure	94.5 %
ROC Area	75 %
Time	0.03 second

In the table1, we discuss the Naïve Bayes algorithm's results. We applied cross validation for splitting records. We noticed that corrected classified records 949 and incorrect classified records 54 from total 1000 records. Also, we noticed that the accuracy ratio is very low with 75% but its very fast. It is applied in 0.03 second.

4.2 Support Vector Machine

Table 2 - SVM Result

Result	Values
Correctly Classified Instances	967
In Correct Classified Instances	33
Precision	97 %
Recall	95 %
F-measure	96 %
ROC Area	83 %
Time	0.55 second

In the table2, we discuss Support Vector Machine algorithm's results. We applied cross validation for splitting records. We noticed that corrected classified records 967 and incorrect classified records 33 from total 1000 records. Also, we noticed that the accuracy ratio is low with 83% and it's fast. It is applied in 0.55 second.

4.3 Decision Tree

Table 3 – Decision Tree Result

Result	Values
Correctly Classified Instances	959
In Correct Classified Instances	41
Precision	95 %
Recall	96 %
F-measure	95.5 %
ROC Area	79 %
Time	0.40 second

In the table3, we discuss Support Vector Machine algorithm's results. We applied cross validation for splitting records. We noticed that corrected classified records 959 and incorrect classified records 41 from total 1000 records. Also, we noticed that the accuracy ratio is low with 79% but it's fast. It is applied in 0.4 second.

4.4 Perceptron Neural Network

Table 4 – Perceptron Neural Network Result

Result	Values
Correctly Classified Instances	978
In Correct Classified Instances	22
Precision	97 %
Recall	98 %
F-measure	97.5 %
ROC Area	90 %
Time	12 second

In the table4, we discuss Support Vector Machine algorithm's results. We applied cross validation for splitting records. We noticed that corrected classified records 978 and incorrect classified records 22 from total 1000 records. Also, we noticed that the accuracy ratio is low with 83% but it's very slow. It is applied in 12 second.

5. Experiments and Results

Table 5 - Tools and device used to preform proposed framework

Metric	**Values**
CPU	Intel core i7
RAM	4G
Operating system	Windows 10
Programming Language	PHP v4
Server Platform	Apache server

For training and testing the data sets, we use ten-fold cross validation technique. This technique splits the dataset to 10 portions 9 portions are then applied to training and that tenth fragment is applied for testing. This is recurring, applying the alternative portion to the test section. Individually the data portion is utilized 1 for testing and 9 events for training. This is recurrent 10 events, including a novel portion doing the testing part. The average outcome is produced from the 10 runs.

The accuracy of the applied procedures must be evaluated applying under titles about rightly classified instances, wrongly classified instances, recall, precision, CPU Time and accuracy. We used several measures to evaluate the methods used on the heart diseases dataset as conferred below:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100\% \quad 1)$$

$$\text{Specificity} = \frac{TN}{FP + TN} \times 100\% \quad 2)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad 3)$$

Table 6 - Comparative Result between Proposed hybrid algorithm and other algorithms

Classifier	**Sensitivity**	**Specificity**	**Accuracy**	**Time**
Decision Tree	95%	96%	79%	0.4
Naïve Bayes	94%	95%	75%	0.03
SVM	97%	95%	83%	0.55
Perceptron	97%	98%	90%	12
Hybrid Perceptron + Decision Tree (iteration=100)	99%	98%	97%	0.8

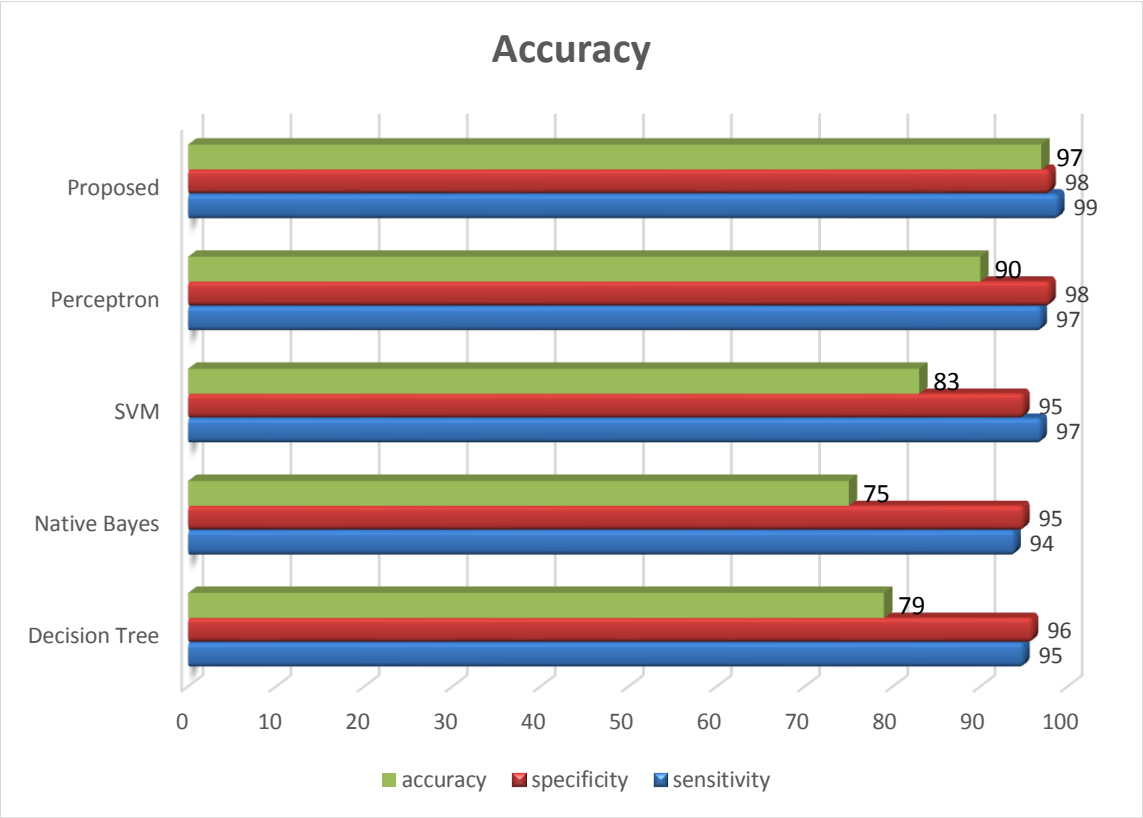

Figure 2- Accuracy Diagram of Comparative Study between proposed hybrid classification algorithm and other algorithms

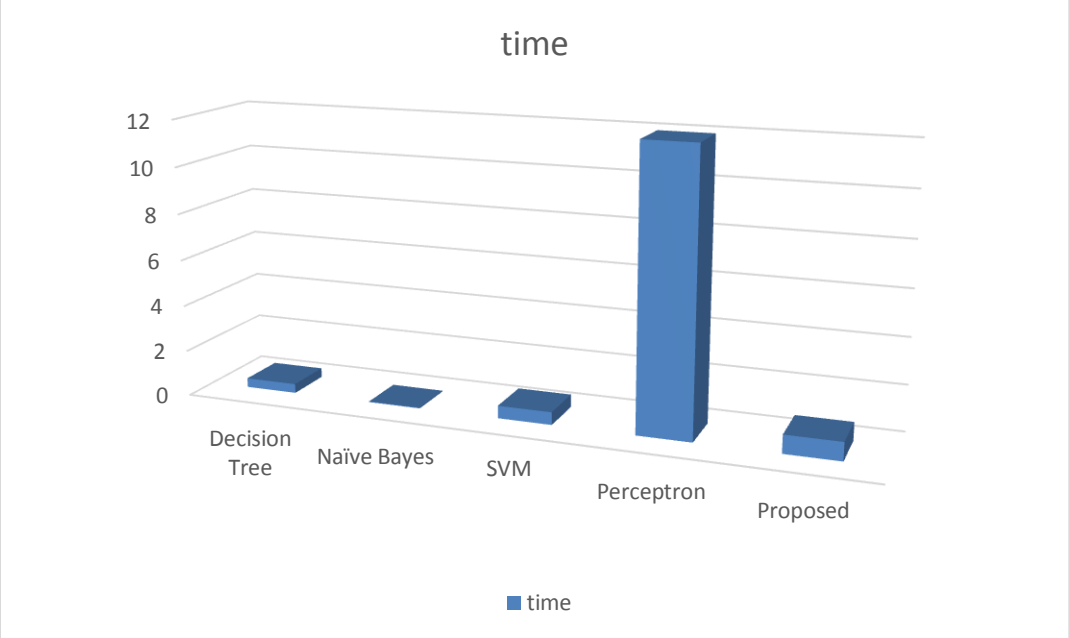

Figure 3- Time Consuming Diagram of Comparative Study between proposed hybrid classification algorithm and other algorithms

6. Conclusion

Within the banking enterprise, optimizing targeting for telemarketing is a vital problem, under increasing stress to improve earnings and decrease losses. Inappropriate, Portuguese banks were constrained to develop capital elements (e.g., by catching extra long-term deposits). In this research, we recommend a particular and smart DSS that applies a data mining (DM) procedure for the determination of bank telemarketing consumers. We examined a current and sizeable Portuguese bank dataset, with a whole of 1000 records. We picked a standardized set of 11 related features. Also, four DM procedures were analyzed: NB, DT, NN and SVM. These models were compared with proposed hybrid classification methods (Decision Tree + Neural) using four metrics, Precision, recall, ROC and time. For both parameters and phases, the valid and high outcomes were achieved by the hybrid techniques, which resulted in a ROC of 97%.

References

1. Sadaf Hossein Javaheri, Mohammad Mehdi Sepehri, Babak Teimourpour, Response modeling in direct marketing: a data mining based approach for target selection, Data Mining Applications with R, Elsevier, 2014, pp. 153–178.
2. Rupnik, R. & Kukar, M. (2007), `Decision support system to support decision processes with data mining', Journal of information and organizational sciences 31(1), 217-232.
3. Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. (2010), `Variable selection using random forests', Pattern Recognition Letters 31(14), 2225-2236.
4. Izetta, J., Verdes, P. F. & Granitto, P. M. (2017), `Improved multiclass feature selection via list combination', Expert Systems with Applications 88, 205-216.
5. Moro, S., Cortez, P. & Rita, P. (2014), `A data-driven approach to predict the success of bank telemarketing', Decision Support Systems 62, 22-31.
6. Moro, S., Laureano, R. & Cortez, P. (2011), Using data mining for bank direct marketing: An application of the crisp-dm methodology, in `Proceedings of European Simulation and Modelling Conference-ESM'2011', Eurosis, pp. 117-121.
7. Qi, G., Zhu, Z., Erqinhu, K., Chen, Y., Chai, Y. & Sun, J. (2018), `Fault-diagnosis for reciprocating compressors using big data and machine learning', Simulation Modelling Practice and Theory 80, 104-127.
8. Barboza, F., Kimura, H. & Altman, E. (2017), `Machine learning models and bankruptcy prediction', Expert Systems with Applications 83, 405-417.
9. Le, H. H. & Viviani, J.-L. (2017), `Predicting bank failure: An improvement by implementing machine learning approach on classical financial ratios', Research in International Business and Finance.
10. Rohani, A., Taki, M. & Abdollahpour, M. (2018), `A novel soft computing model (gaussian process regression with k-fold cross validation) for daily and monthly solar radiation forecasting (part: I)', Renewable Energy 115, 411-422.

