JKAU: Comp. IT. Sci., **Vol**. **9** No. 2, pp: 69 – 83 (1441 A.H. / 2020 A.D.)
Doi: 10.4197/Comp. 9-2.4

Predicting the Accuracy for Telemarketing Process in Banks Using Data Mining

Fawaz J. Alsolami[1], Farrukh Saleem[2] and **Abdullah AL-Malaise AL-Ghamdi[2]**

[1] Computer Science Department, and [2] Information Systems Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

falsolami1@kau.edu.sa

Abstract. Managing and analyzing the data that generates from routine business operations is one of the biggest challenges for the banking industry. Banks play a significant role in the growth of the economy as well as provide numerous financial services to the customers. In those services, telemarketing is a common business strategy applies in the bank to offer and promote new products and services to its customers. This type of campaign produces very large dataset, proper analysis of those data can support the bank in planning future strategies. Therefore, this study proposed a data mining approach, to analyze and predict using the telemarketing campaign dataset. The dataset was prepared based on the pieces of evidence collected from the customers, during the live call session organized by the bank. To implement the proposed model, we selected the real dataset generated during the telemarketing process. Logistic Regression, Decision Tree, and Multilayer Perceptron were part of conducted experiments. The cross-validation strategy applied for measuring and comparing the performances of each algorithm. The result suggested that logistic regression provides the best accuracy among the three models, recorded as 91.48%. The research is helpful for the banking industry, where the model can be used to predict the success of telemarketing calls and understand the patterns within the dataset. Moreover, the decision-makers can use the model for defining their future strategies to run the telemarketing process efficiently.

Keywords: Dimensional analysis, Dimensionless products, virus spread rate, basis and regime variables, Gauss-Jordan elimination, Temperature as energy.

1. Introduction

Telemarketing is one of the common strategies for developing a positive association between the company and customers. This type of strategy is quite a time consuming, requires hard effort from the employees to contact the customer to achieve the company's goal. The overall strategy is known as direct marketing or telemarketing executes in a specific period of time using different communication channels [1]. This type of marketing campaigns run from prearranged locations and targeted contact list. To provide some ease in this operation, researchers have proposed several strategies integrated with use of latest technologies. For example business analytics in telemarketing using artificial neural networks [2], chat-bot design on real marketing data [3], and telemarketing forecasting using SVM [4]. Therefore, it can be evident that use of current technology can reveal many facts for the organizations by extracting hidden patterns from the collected dataset. The interesting patterns like customer's choice and willingness about products and services, and to whom the organization can offer similar services with attractive offers.

This research is focused on using the bank data for predicting the success of telemarketing in banks. The banking industry still requires and open with lot of opportunities to improve their business and operational activities. Enhancing

bank's database abilities and provide sustainable business platform are some major issues targeted in previous work [5], where the banks are reluctant to provide space to the outsiders.

The main reason behind this issue can be restriction and privacy issues as bank possesses high confidential data and has strict policy in sharing of the customer's data. Another matter discussed in previous researches is the extraction of required knowledge from the bank data [6]. It is due to having numerous unrelated attributes are part of bank dataset. These were some examples highlighted in previous researches. Therefore, this research dealing with large dataset managed and generated by the bank working in Portugal. This research provides an efficient way to deal with complex issues as discussed above related with banking industry dataset.

The main problem undertaken in this research is "how data mining algorithms can improve the process of telemarketing campaign for the banks". The proposed model can assist the banks to understand the targeted customers. Subsequently, the research will consider the problem for the banks in handling the customer's data and contacting them for deposit money. The prediction of accuracy generated by different data mining algorithm to improve bank's telemarketing business model. The CRoss Industry Standard Process for Data Mining (CRISP-DM) selected due to its high applicability for finding out hidden patterns from the dataset [7]. The model has widely used for doing business analytics, forecasting and providing business intelligence to the enterprises [8]. The data mining approaches has also used for different purposes such as associating different stakeholder [9] and for building smart enterprise application using data mining [10].

This research article is further organized as follows. Section II described the related work and achieved results in this field of study. CRISP-DM step-wise methodology described in Section III. The implementation of the model using real bank data is further explained in Section IV. In the same section, all major steps discussed in detail to explain the data, model building and evaluation. Section V provides the generated results and comprehensive discussion. In addition, the results further compared with the previous work to understand the measured accuracy generated by proposed model. Finally, the research concludes by explaining all efforts performed in conducing this study, with suggestion for the future researchers.

2. Related Work

The banking industry always trying to reach to the customer using any channel to provide and offer them different services based on their characteristics. Therefore, telemarketing is one the major tools for contacting with the customers. The campaign can run for different purposes such as offering new services and promoting new product. In recent years, the telemarketing approach turn out to be a preferred tool specially in banking industry [6], [11]. The services offered by banks are different such as deposit money, saving certificates, home loan, business loan, educational scholarships and others. Overall, the purpose of direct marketing is to keep in contact with the customers and send them product's information in time. Whereas, during the process of contacting the customers, bank managed the contact list by filling the responses collected against each question during discussion.

The problem relates in this study, to assist the bank by analyzing the large dataset organized during different activities. The main issue with the bank is how to get correct prediction ratio during the direct marketing

campaign. The idea was presented in the study [2] where the researcher proposed the artificial neural net model to forecast the chances of customers who can apply for short and long term deposits in the bank. Furthermore, the research incorporated by defining the rules, which can be helpful for the decision makers to understand the potential benefits that can be achieved through the marketing campaign. The research has further proved that how integration of technology in business operations can enhance the growth of industry. The example was presented using the real case study of banking sector, whereas the model was verified through statistical tests with high accuracy.

The previous literature suggested that the data mining approaches has offered lot of opportunities to the business industries to expand their business using illustrative and forecast models [12]. It has been used and implemented in different scenario to identify the potential customers by analyzing the historical data, and identify the customer using predictive models [13]. Decision tree algorithm is one of the methods of data mining that used to classify the data based on supervised learning approach, and efficient for prediction [14]. In addition, neural net, logistic regression, naïve Bayes are some other classification model used commonly in business industry [15].

The CRISP-DM approach applied several times in different studies. The main purpose of this strategy is to provide business intelligence to the industry by analyzing the historical data [16]. The experiment conducted on the bank data generated from direct marketing campaign [17]. The targeted customer were contacted to deposit money for long term for the financial growth of the bank. The research applied CRISP-DM approach by using three different algorithms naïve Bayes, decision tree, and support vector machine. In addition, the research described the selected attributes where most of the female customers accepted the offer from the bank. Overall implementation of the experiment conducted using r-miner. The accuracy calculated from each model were recorded as 0.87, 0.86, and 0.93 from naïve Bayes, decision tree, and support vector machine respectively. The results suggested the high accuracy is measured through support vector algorithm.

Another idea for predicting the success of direct marketing in banking sector is presented by [18], which has been evaluated using different machine learning algorithms. The idea was established to analyze the data set collected from 2008 to 2013. During the modeling phase, feature selection technique applied to reduce the number of attributes and to select most appropriate list of attributes. The selection of attributes was statistically proved by feature selection methods. Two different feature selection strategy applied known as "no selection" and "forward selection" methods. Four different method used during model validation such as logistic regression, decision tree, neural network, and support vector machine. Overall, after successful validation process, the neural network provided the best result where 0.80 accuracy measured.

In addition, an article discussed about the importance of telemarketing that, managing the customer's data in an efficient way always provide the platform to build positive relationship and connection during promotions and new products offer [19]. The said paper applied the integrated approach of data mining and machine learning on bank data. The purpose was the same as discussed in this paper to get the optimal solution for the bank in executing telemarketing campaign. The proposed model used two different models; logistic regression and multilayer perceptron. The purpose was to differentiate the performance between classification and multilayer neuron model. During the cross validation process, different combination were

applied which provides the mixed accuracy measurement.

The related work discussed in this section, explored the main problems undertaken in this research. The previous work suggested the importance of the research idea as some of the researches are published in past two years. Therefore, this research can have significant impact on banking industry, where decision makers can used integrated computing approaches for data analysis and to have statistical proof before building new telemarketing strategy. Based on the literature review, the most common approaches extracted and used in previous work are logistic regression, decision tree, and multilayer perceptron. The most related papers and their model performances are showing in Table 1, where the same dataset used as selected in this research. The table is highlighting the lowest performance was measured through decision tree, whereas the highest performance resulted with the help of multilayer perceptron.

Table 1. The Performance Measurement in Previous Work.

Techniques / Previous Work	UCI Data, Using R Tool, [17]	UCI Data using Weka, 2019 [19]	UCI Data Using R Tool, 2014 [18]
Decision Tree	86.80%	X	83.30%
Logistic Regression	X	90.18%	90. 0%
Multilayer Perceptron	X	89.40%	92.90%

3. Methodology Based on CRISP-DM

The methodology is the critical and central part of the research paper. This section describes the stepwise approach to be taken to achieve the goal of this research. The research methodology in this study is based on CRISP-DM model, which is famous method for data mining implementation [16]. The same approach were taken in the related work selected in this study as discussed in the previous section [17]. The proposed research framework illustrated in Fig. 1.

As shown in the following figure, the process starts by selecting the data from the bank database, to understand the business environment. The data selected form the bank industry, in this step it will further identify the major business processes and operational models of the bank. The business understanding process will provide the contextual information about data and business. Data understanding is the next step as shown in the methodology. The author can identify the major attributes of the data, and how and when data generated. It can further clarify the type of process or campaign, which resulted the data creation. Both of these steps required to get the contextual information about the organization and data.

Furthermore, data preparation is the next step, which is based on CRISP-DM strategy. This is one of the critical steps in the overall research methodology. As any mistake in data or attributes selection can lead to the inappropriate results. The main procedure in this step is to apply data pre-processing techniques such as impute missing values, outlier detection, data type transformation and others. Attributes selection also very important factor, as researcher can remove those attributes, which are not related with the experiment. Proper knowledge, data management experience, and type of models play important role in attributes selection. Afterwards, the modeling phase starts to select data mining model based on the data specification. It should be clear at this step that which type of data mining techniques is applying on the selected data, and what is the purpose of implementation. Indeed, the step is important again to choose the algorithm based on data understanding.

Moving forward to the evaluation process that is the next phase in this framework. It provides the overall analysis and performances of the results generated during the experiment. The experiment conducted in

two phases that are training and testing phase. After execution of both phases, the model assessment will highlight the success or failure of the conducted experiment. The performance evaluation can be analyzed using multiple criteria such as confusion matrix and accuracy. Based on the accuracy of the model the trained model can be used and deploy for future development in organization.

4. Framework Implementation

Based on the framework presented in the previous section, this phase of the work presents the implementation of the framework using data mining algorithms. The steps of the framework discussed in the subsequent sections.

4.1 Business Understanding

The research used the real world data generated by the bank in executing their telemarketing campaign. During this campaign, the bank asked from its customers about their willingness to deposit the money in the bank or not. The data has been generated by one of the Portuguese banks, where the customer contacted by different customers. Therefore, the research in this study related to the bank, where the data mining model used to predict the customer's choice regarding bank deposit. Each record in the dataset divided into multiple attributes and can be classified based on the customer's agreement or disagreement. The data has been applied using the same scenario as described in the dataset. The research can be helpful for the banks to predict the chances of success and failure during this type of telemarketing campaign. The details of data and list of attributes are defined in the next section.

4.2 Data Understanding

As discussed earlier, the dataset used in this study belongs to the real bank data, where the data collection was performed by Portuguese bank during 2008 to 2013. The data was taken from UCI machine repository, which was first time used by [18] to extract new knowledge from the dataset. The data was generated from the telemarketing campaign, where the main purpose of this campaign was to ask from the customers for long term deposit in the bank. The dataset has different 17 attributes and around 45 thousand transactions, and having multiple distinct values categorized under each variable. Here each attribute has been investigated by asking different questions. The distinct values under each variables are actually highlighting the answers received from the participants during the marketing campaign. Some of the common variables investigated are age, loan information, duration of the overall campaign, and outcome of earlier campaigns. In addition, each attribute has been defined using different data types such as numerical, binary, and categorical. Finally, as the dataset is supervised data, therefore it has the final output column known as class or label column. This column defined about the acceptance and rejection by the client in regard of long deposit offer asked by the bank.

4.3 Data Preparation

In this step, the data is prepared and to be ready according to the data mining models. For implementation, the rapid miner tool used, which is one of the common data mining tools has been applied several time in earlier researches [20], [21]. Therefore, for data preprocessing the provided operator in rapid miner used accordingly. Firstly, the data was selected and randomly shuffled to prepare it for implementation in rapid miner. Secondly, the missing values check was applied to know if there is any data missing in the file. The data file was further modified by converting different attributes from numerical to binary as logistic regression measure the probability based on binary numbers. For example, initially the class column has the two different values "yes" and "no", which converted into "1" and

"0" respectively. Here the "1" denotes to the client's acceptance and "0" refers to the rejection for the long term deposit in the bank. The different steps of data preparation is further shown in Fig. 2.

4.4 Modeling

During the modeling phase, the bank data imported and multiple copies created using different rapid miner operators. To understand the level of performances, multiple data mining techniques used in this research and applied for training purposes. Logistic regression, decision tree, and multilayer perceptron were selected in this experiment. The selection of algorithm is based on the previous work and performances as shown in Table 1. Overall, the modeling phase is illustrated in Fig. 3 that implemented in rapid miner. The description of each algorithm and details about modeling phase is discussed in subsequent section.

4.4.1 Logistic Regression

This algorithm is type of regression, which deals specifically when the dependent variable is in binary format. Moreover, other variables like independent variables can be of different type such as nominal or ordinal. Initially, the model was proposed by David Cox to evaluate and estimate the probability of binary number [22]. Logistic regression is common algorithm of machine learning and data mining that is also called supervised learning used to predict and measure the probability for a particular distinct values given in the dependent variable.

In this study, the selected dataset has a column known as class variable. This variable has two distinct values denoting as "0" and "1". This attribute actually the dependent variable and it also defining the final output of telemarketing campaign. For example, by asking different questions from the clients, the final output column is illustrating that, if customer agreed or not for long term deposit in the bank. Logistic regression algorithm used and applied in rapid miner to understand the association between single dependent variable and multiple independent variables. The probability in this experiment will be calculated using following formula:

$$\text{Logit(p)} = ln\left(\frac{p}{1-p}\right) = \frac{\text{probability of presence of characteristics}}{\text{probability of absence of characteristics}} \quad (1)$$

The p is the probability of presence of the characteristic.

The implementation of logistic regression in rapid miner is presented in the following figure. The first step was to import the dataset using "import operator". As the model execute three algorithm together, to check their performances in one run. Although, the experiment took long time to run, but it was good strategy and used earlier in the related work. Afterwards, in the next step researcher develop the multiple copies of data to connect it with multiple classifier. The operator that is showing in the figure named as "logistic regression" is actually the cross validation operator. This operator has the ability to run classifier inside of it. The main advantage of cross validation operator is to perform both processes together; that is training and testing of the model. The details of this operator is discussed in the next phase that is "evaluation phase".

4.4.2 Multilayer Perceptron

The second model selected in this study is related to neural network. This kind of model are used widely for different purpose such as prediction and estimation. This is a kind of machine learning model where we can train the model using the biological term called neuron [23]. Neurons are the fundamental units of neural network that organized and connected according to the architecture of the network [24]. The auto multilayer perceptron is used in this

experiment, which is known as feed-forward network. In this algorithm, in addition to the input and output layers, there can be one or more hidden layers. Each layers consists of multiple nodes which also known as perceptron [25]. During the training phase, each perceptron received information, which used to train the model.

In this kind of network, each node in a layer connected with each node of other layer with some weight. The information are sending from each node of input layer to the each node of hidden layer. Furthermore, the hidden layers connected with output layer [26]. The number of nodes in output layer can be identified based on the number of distinct values in class variable. An example of MLP is shown in Fig. 4. The figure highlighting that there is two hidden layers. It can be seen that the all nodes in a layer are connected with all nodes in forward layer.

In this study, the model implementation performed using three layers; input, hidden, and output layers. The model implemented using cross validation operator as shown in Fig. 3. First, the input layer that have multiple nodes, where each node denotes all values under each variable. The input layer is also known as list of independent variable which can create impact on decision variable after passing through hidden layer. The output layer belongs to the decision variables. In this case the output layer has two values "Yes" and "No", which representing the acceptance or rejection by the client for the offer received from the bank. The middle layer in neural network is known as hidden layer. Normally, the hidden layer can be decided according to the probability distribution and success rates. The model run 10 training cycles and 10 number of generations as well. Finally, the hidden layer used the activation function such as sigmoid, which provides the values from 0 to 1. The purpose of activation function is to generate weight, and limit the output between 0 and 1. That can further be used to predict the probability for final output layer. In this experiment the sigmoid values at Node1 located under Hidden Layer1 is shown in Fig. 5.

4.4.3 Decision Tree

Decision tree is the third algorithm used in this research to train, test, and predict the telemarketing campaign data. This algorithm generate different nodes, which connected to each other and look like a tree. Each node in the tree work for a specific attribute and work on the bases of splitting rule. The splitting rule helps the classifier to collect the sample fulfilling the node's criteria [27]. The classifier is popular in data mining and machine learning area of research, whereas the classification of data in a tree-like nodes is used to predict the new dataset. This tree like structure is easy to implement and feasible to interpret the small tree structure, which make this algorithm fast implementation than other.

The selected dataset applied using decision tree algorithm to predict the accuracy of the telemarketing campaign. During the training phase the example dataset used to prepare the machine using tree structure. The criteria "gain ratio" used for splitting the attributes. Gain ratio is a common criteria apply on each attribute to create the uniformity of the attributes [28]. An example of splitting attribute can be 50>age>50.

4.5 Evaluation

The evaluation phase is an important part of the experiment conducted in this study. In the modeling phase, all the algorithm logistic regression, MLP, and decision tree applied and trained based on the selected data. The evaluation phase applied using cross validation procedure to run the evaluation. The cross validation operator is a common procedure in machine learning, which used to validate the model. It further provides the statistical performance of the learning algorithm. It is a

nested operator that divides into two part known as training and testing phase. We used 10 fold cross validation to divide the number of examples into subset for training purposes. It will run the process 10 times to train the model [29]. In the training phase, the selected classifier applied in order to train the model. At this stage, we can select the particular criteria or parameter for the classifier. Furthermore, the train model forward its specification and learning output to the testing port. In the testing phase, normally two operator used, the first and mandatory operator is known as "apply model". Apply model receives the trained model from one port, whereas at the second port it received the testing dataset. The last step in evaluation phase is to test the performance of the model. Therefore, the "performance" operator is used, which connected with "apply model" port to receive the results after model validation. This performance port will deliver the overall performance of the classifier in the form of confusion matrix. Figure 6 is representing the evaluation phase for logistic regression classifier. This screenshot is showing the nested view of cross validation operator.

4.6 Deployment

After conducting the overall experiment in which step by step different procedure applied. Starting from business understating till model evaluation. It is the phase where the researcher will evaluate the performance of the experiment. The confusion matrix generated for each model, which has different criteria such as precision, recall, and overall accuracy of the model. The performance of the classifier is explained in the next "result and discussion" section. Here, based on the performance of the model the bank can deploy the model for new dataset and can take the decision for new business campaign.

5. Result and Discussion

This section describes the result generated from the experiment conducted in this study using CRISP-DM approach. Overall, there were three model applied and executed as discussed earlier. The model successfully trained using the selected dataset. Finally the model validation applied using 10fold cross validation method. The obtained results are showing in Table 2. The table is illustrating the results using different criteria mentioned as precision, recall, and overall accuracy. There are different purpose of each criteria, for example the precision values generated the percentage based on correct predictions divided by total predictions. On the other side, the recall criteria is used to understand by dividing the number of correct prediction results divided by total number of results should be returned. Finally, the last and most important criteria used in this research is to understand the results is known as accuracy. The accuracy is the measuring criteria, which highlights the percentage of dataset that predicted correctly [30]. Finally, the results shown in the table is also illustrating the comparison of accuracy between different models and model's reliability so that it can be used in future or not.

Table 2. The Performance Metrics for All Models.

Classifier	Class	Precision	Recall	Overall Accuracy
Logistic Regression	0	93.31%	97.39%	91.48%
	1	68.61%	44.99%	
Decision Tree	0	93.79%	96.97%	89.91%
	1	58.82%	57.92%	
Multilayer Perceptron	0	91.90%	97.43%	90.10%
	1	61.55%	32.39%	

The success ratio of the model is presented based on the decision variable that was about the client acceptance of rejection on

the offer from the bank. This class has two distinct values; "0" and '"1" as can be seen in the above table. The class "0" is representing that customer has not accepted the bank offer, whereas class "1" is denoting the agreed group of customers. The result showing the correct prediction level, rather discussing about the chances of customer agreement. All three approaches has shown acceptable performances of the model for predicting the new dataset. Overall, in all three classifiers, for class "0" the precision and recall values are better than class "1". As discussed above all criteria such as precision, recall, and accuracy performances are measured through original and predicted data values. In the dataset, most of the data belonged to the class "0", as during this telemarketing campaign, most of the customers refused the offer. The possible reason behind low precision and recall values of class "1" may be the less number of dataset available in training phase. The maximum precision values were measure for class "0" by decision tree (93.79%), whereas the least is measured by MLP (91.90%). Looking at the recall values, the maximum measured for class "0" by MLP that is 97.43%. Whereas, 96.97% is the recall value calculated by decision tree algorithm for class "0".

The overall accuracy measured by each algorithm is much closed to each other, which highlights the importance of each algorithm. The main purpose of this research is to propose a machine learning framework that can train and predict the instances efficiently. The model can be used further analysis of the new dataset. The implementation of the model used the cross validation process, which can help to avoid overfitting. Apart from the recall and precision values where the mixed high and low values generated from each algorithm. Here, the overall accuracy is important to highlights the best performance among all. It can be evident from the Table 2 that best promotion efficiency is provided by logistic regression where the accuracy measured as 91.48%. It shows that this algorithm can provide measure benefits to the bank while executing the same kind of campaign. The model has trained and run efficiently using the telemarketing data. The other two performances are also much closed to each other, like decision tree (89.91%) and MLP (90.10%). The model has shown some better performance as compare to related work.

Table 3. The Performances Comparison with Related Work.

Techniques / Previous Work	**This Study**	**Related Work**		
		UCI Data, Using R Tool, [17]	UCI Data using Weka, 2019 [19]	UCI Data Using R Tool, 2014 [18]
Decision Tree	89.91%	86.80%	X	83.30%
Logistic Regression	91.48%	X	90.18%	90.0%
Multilayer Perceptron	90.10%	X	89.40%	92.90%

Finally, the comparison of accuracy measured in this study with previous work is showing in Table 3. The proposed framework applied in this study used three different machine learning algorithms. In previous work different combination were used as mentioned in above table. In this study, most of the model performed well and measured better accuracy than related work except MLP. The comparison report provides the better understating and clarification regarding the improvement in current work. For example, the decision tree performance recorded in this study is 89.91%, which is higher than the accuracy evaluated and presented in first (86.80%) and third (83.30%) related work. In addition, the performance of MLP is very close to second related work, but to compare with third related work the performance is less by 2.8%. The reason behind the low performance can be related with model training or selection of data during training and testing phase. Altogether, the best model predicted in this study is logistic regression

with accuracy higher among all three model that is 91.48%. Comparatively, the model also outperformed in all related work (90.18% and 90.0%) as shown in the table. According to the results and measured accuracy, the logistic regression can be ideal model for predicting the consent of the customers regarding long term deposit offer from the bank.

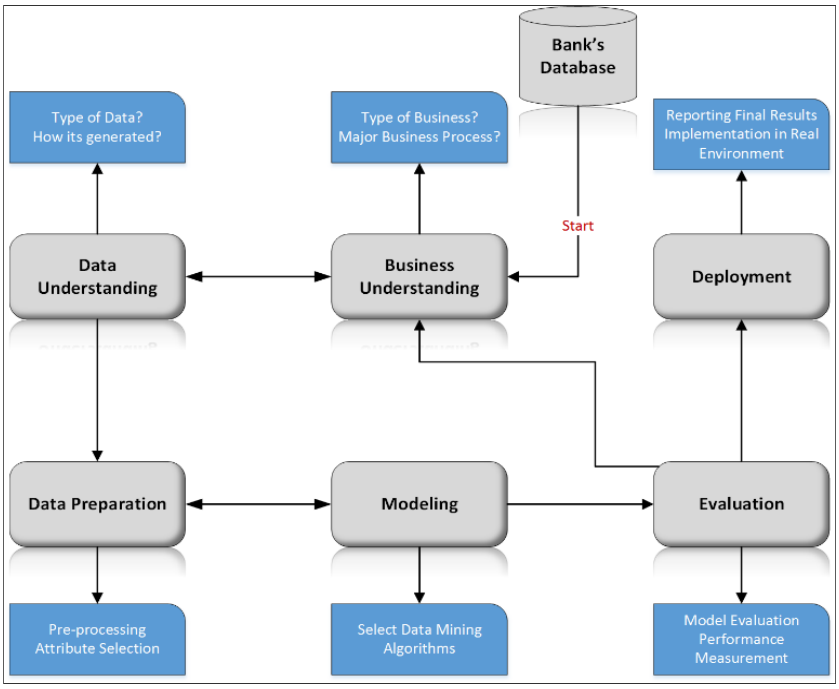

Fig. 1. The Research Model Adapted from CRISP-DM.

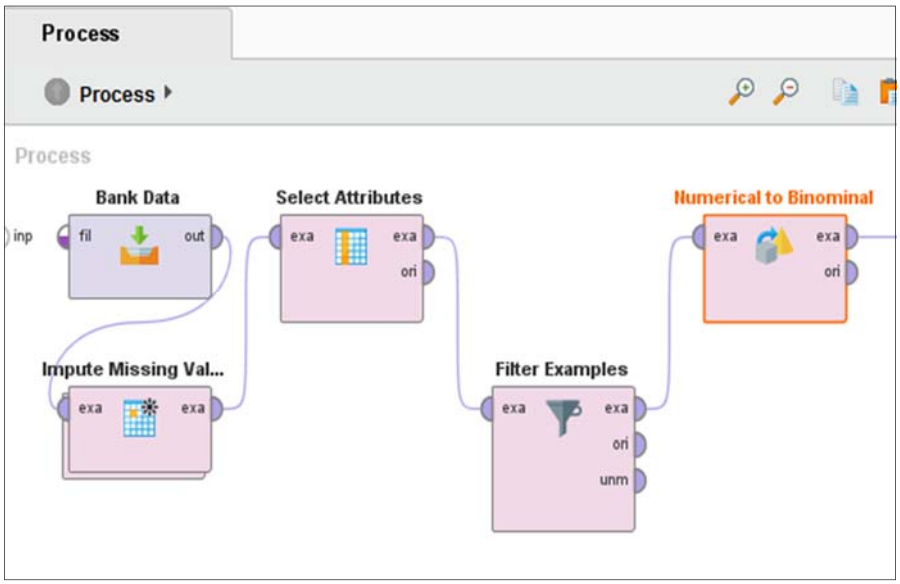

Fig. 2. Data Preparation Phase.

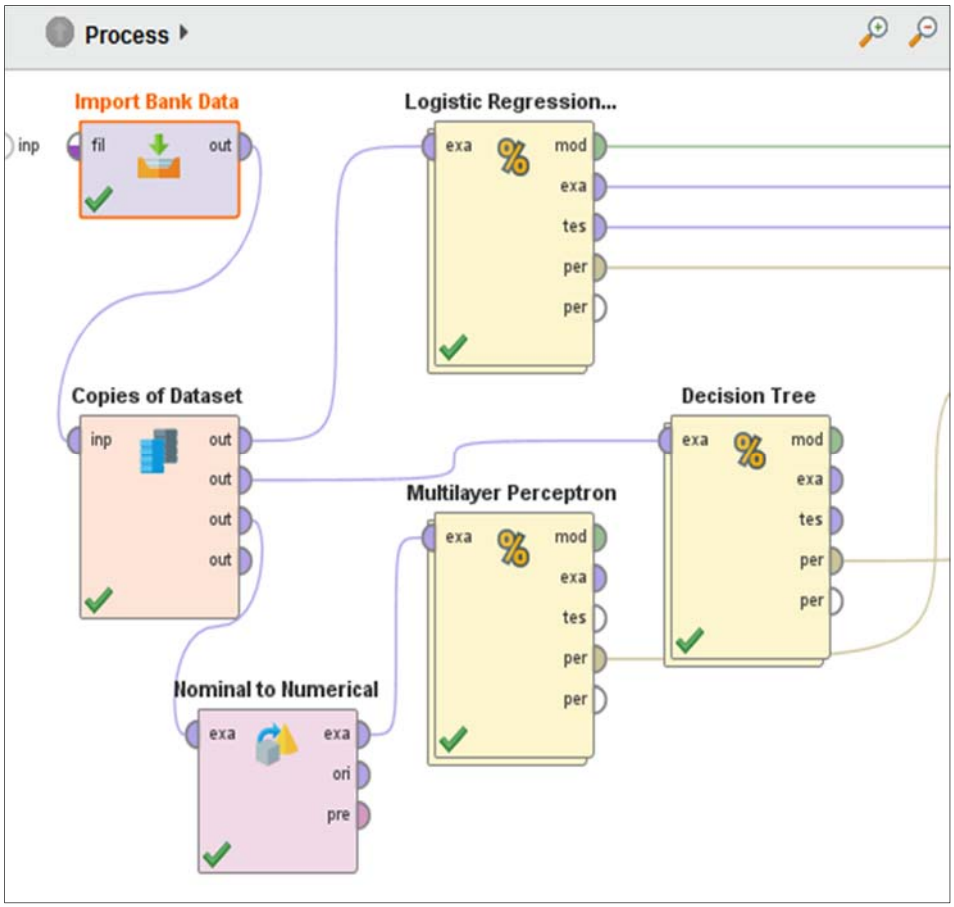

Fig. 3. Modeling Phase - Using All Classifiers.

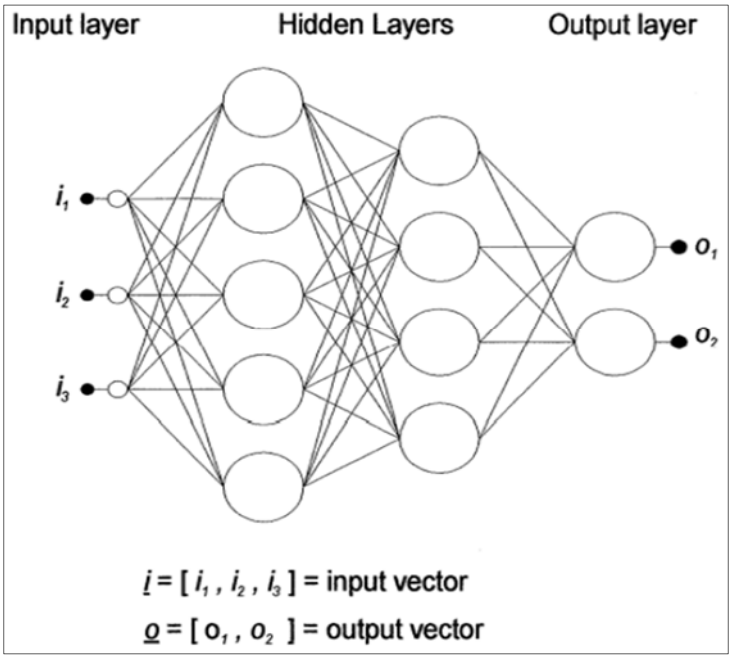

Fig. 4. MLP with Two Hidden Layers [26].

```
AutoMLPImprovedNeuralNet

Hidden 1
========

Node 1 (Sigmoid)
----------------
job = housemaid: 2.553
job = services: 3.146
job = admin.: 1.246
job = blue-collar: 2.595
job = technician: 1.332
job = retired: -0.978
job = management: -0.087
job = unemployed: -0.534
job = self-employed: 2.534
job = unknown: -0.673
job = entrepreneur: 0.849
job = student: -1.226
marital = married: -1.865
marital = single: 0.290
marital = divorced: 3.918
marital = unknown: -0.126
education = basic.4y: 1.176
education = high.school: -0.149
education = basic.6y: 0.585
education = basic.9y: -0.149
```

Fig. 5. Sigmoid Values at Nod-1, Hidder Layer-1.

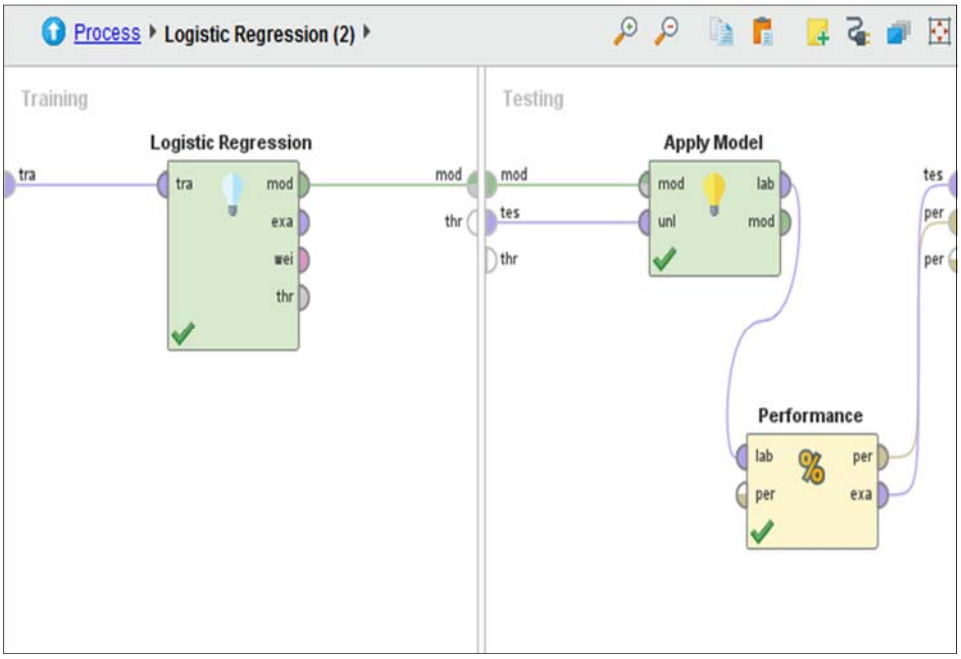

Fig. 6. Evaluation Phase – Logistic Regression Classifier.

6. Conclusion and Future Work

The telemarketing campaign are supposed to be beneficial for the banking industry to create positive relationship with the customers. This type of campaign offers several seasonal and attractive promotions to their customer by using different communication channels. This study used one of this kind of data maintained and launched by Portuguese bank. The data was created during telemarketing campaign, where they asked different questions from the customers to get their consent on long term deposit money in the bank. The data analysis and machine learning algorithms applied in this study to predict the customer's level of agreement. The study provides the performances of multiple algorithms. The performance were than compared with previous work to better understand the implications of this research work. Overall, among three models selected in this study the best accuracy measured by logistic regression (91.48%). Furthermore, the same algorithm were checked in previous work, and it proved the best with high performance among three related work who used logistic regression. The framework can be applied using other machine learning algorithm by combining with different validation processes.

References

[1] **Kotler, P.** and **Keller, K. L.,** Framework for *Marketing Management*, 5th Editio. 2012.

[2] **Ghatasheh, N., Faris, H., AlTaharwa, Harb, I., Y.** and **Harb, A., "**Business Analytics in Telemarketing: Cost-Sensitive Analysis of Bank Campaigns Using Artificial Neural Networks," *Appl. Sci*., **10**(7), 2020.

[3] **Zhang, J., Zhang, J., Ma, S., Yang, J.** and **Gui, G.,** "Chatbot Design Method Using Hybrid Word Vector Expression Model Based on Real Telemarketing Data," *KSII Trans. Internet Inf. Syst*., **14**(4) 2020.

[4] **Che, J., Zhao, S., Li, Y.** and **Li, K.,** "Bank Telemarketing Forecasting Model Based on t-SNE-SVM.," *J. Serv. Sci. Manag*., **13**(3) 2020.

[5] **Ajah, I. A.** and **Nweke, H. F.,** "Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications," *Big Data Cogn. Comput*, **32**(3), 2019.

[6] **Moro, S., Cortez, P.** and **Rita, P.,** "divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing," *Expert Syst*., **35**, 2018.

[7] **Han, J., Pei, J.** and **Kamber, M.,** *Data mining: Concepts and techniques*, 3rd ed. Elsevier, 2012.

[8] **Spruit, M., Vroon, R.** and **Batenburg, R.,** "Towards healthcare business intelligence in long-term care: An explorative case study in the Netherlands," *Comput. Human Behav*., **30**: 698–707, 2014.

[9] **Saleem, F.** and **Malibari, A.,** "DATA MINING COURSE IN INFORMATION SYSTEM DEPARTMENT–CASE STUDY OF KING ABDULAZIZ UNIVERSITY," in *3rd International Congress on Engineering Education, 2011*.

[10] **AL-Ghamdi, A. A.-M.** and **Saleem, F.,** "Enterprise application integration as a middleware: Modification in data & process layer," in *Proceedings of 2014 Science and Information Conference, SAI 2014, 2014*, pp. 698–701.

[11] **Koç, A.** and **Yeniay, Ö. A.,** "Comparative Study of Artificial Neural Networks and Logistic Regression for Classification ofMarketing Campaign Results," *Math. Comput. Appl,* **18**, 2013.

[12] **H. A. Elsalamony** and **A. M. Elsayad,** "Bank Direct Marketing Based on Neural Network," *Int. J. Eng. Adv. Technol*., **2**(6), 2013.

[13] **U. M. Fayyad, G. Piatetsky-Shapiro,** and **P. Smyth,** "*From Data Mining to knowledge Discovery: An Overview, Advances in Knowledge Discovery and Data Mining*," AAAI Press., 1996.

[14] **Thimm, G., Fiesler, E., Furuhashi**, and **Takeshi,** "Neural Network Pruning and Pruning Parameters," in *1st Workshop on Soft Computing*, 1996.

[15] **Dalipi, F., Imran, A. S.** and **Kastrati3, Z.,** "MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges," in *EDUCON 2018*, 2018.

[16] **P. Chapman et al**., "CRISP-DM 1.0 - Step-by-step data mining guide," in *CRISP-DM Consortium*, 2000.

[17] **Moro, S., Laureano, R. and Cortez, P.,** "Using data mining for bank direct marketing: An application of the crisp-dm methodology.," in *Proceedings of European Simulation and Modelling Conference, 2011*, pp. 117–121.

[18] **Moro, S., Cortez, P.** and **Rita, P.,** "A data-driven approach to predict the success of bank telemarketing," *Decis. Support Syst.,* **62**: 22–31, 2014.

[19] **Mauritsius, T., Braza, A. S.,** and **Fransisca,** "Bank Marketing Data Mining using CRISP-DM Approach," *Int. J.* Adv. *Trends Comput. Sci. Eng*., **8**(5) 2019.

[20] **Tripathi, P., Vishwakarma, S. K.** and **A. Lala,** "Sentiment analysis of english tweets using rapid miner," in Computational Intelligence and Communication Networks (CICN), *2015 International Conference on, 2015*, pp. 668–672.

[21] **Angra, S.** and **Ahuja, S.,** "Implementation of Data Mining Algorithms on Student's Data using Rapid Miner," in *International Conference On Big Data Analytics and computational Intelligence (ICBDACI),* 2017, pp. 387–391.

[22] **DW Jr, H., S, L.** and **RX, S.,** *Applied Logistic Regression*, 3rd ed. New Jersey: John Wiley & Sons, 2013.

[23] **Kim, K. H., Lee, C. S., Jo, S. M.** and **Cho, S. B.,** "redicting the success of bank telemarketing using deep convolutional neural network," in: *7th International Conference of Soft Computing and Pattern Recognition (SoCPaR), IEEE, 2015*, pp. 314–317.

[24] **Haykin, S.,** *Neural Network and Learning Machines*, 3rd ed. Pearson, 2009.

[25] **Churchland, P. S.,** *Toward a unified science of the mind/brain,* Reprint ed. Bradford: A Bradford Book, 1986.

[26] **Gardner, M. W.** and **Dorling, S. R.,** "Artificial Neural Networks (The Multilayer Perceptron**) -** A Review of Applications in the Atmospheric Sciences," *Atmos. Environ*., **32**(14): 2627–2636, 1998.

[27] **Delen, D.,** "Predicting student attrition with data mining methods," *J. Coll. Student Retent. Res. Theory Pract.,* **13**(1), 2011.

[28] **Documentation, R. M.,** "*Information Gain*." [Online]. Available: https://docs.rapidminer.com/8.0/studio/operators/modeling/predictive/trees/parallel_random_forest.html. [Accessed: 01-Apr-2019].

[29] **Rapid Miner**, "*Cross Validation Operator*." [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/validation/cross_validation.html. [Accessed: 01-Apr-2020].

[30] **Olson, D. L.** and **Delen, D.,** "Performance evaluation for predictive modeling. Advanced data mining techniques," in *Advanced Data Mining Techniques, 2008*, pp. 137–147.

توقع دقة عملية التسويق عبر الهاتف في البنوك باستخدام استخراج البيانات

فواز السلمي[1]، و فرخ سليم[2]، وعبدالله المليص الغامدي[2]

[1]قسم علوم الحاسبات، و[2]قسم نظم المعلومات، كلية الحاسبات وتقنية المعلومات، جامعة الملك عبد العزيز، جدة، المملكة العربية السعودية

falsolami1@kau.edu.sa

المستخلص. تعد إدارة وتحليل البيانات الناتجة عن العمليات التجارية الروتينية أحد أكبر التحديات التي تواجه الصناعة المصرفية. تلعب البنوك دورًا مهمًا في نمو الاقتصاد بالإضافة إلى تقديم العديد من الخدمات المالية للعملاء. في هذه الخدمات، يعد التسويق عن بعد استراتيجية عمل مشتركة مطبقة في البنك لتقديم وترويج منتجات وخدمات جديدة لعملائه. ينتج عن هذا النوع من الحملات مجموعة بيانات كبيرة جدًا، ويمكن أن يدعم التحليل المناسب لهذه البيانات البنك في التخطيط للاستراتيجيات المستقبلية. لذلك، اقترحت هذه الدراسة نهجًا لاستخراج البيانات لتحليل والتنبؤ باستخدام مجموعة بيانات حملة التسويق عبر الهاتف. تم إعداد مجموعة البيانات بناءً على الأدلة التي تم جمعها من العملاء، خلال جلسة المكالمة المباشرة التي نظمها البنك. لتنفيذ النموذج المقترح، اخترنا مجموعة البيانات الحقيقية التي تم إنشاؤها أثناء عملية التسويق عبر الهاتف. كان الانحدار اللوجستي وشجرة القرار وPerceptron متعدد الطبقات جزءًا من التجارب التي أجريت. تم تطبيق استراتيجية التحقق المتبادل لقياس ومقارنة أداء كل خوارزمية. أشارت النتائج إلى أن الانحدار اللوجستي يوفر أفضل دقة بين النماذج الثلاثة، مسجلة بنسبة 91.48٪. يعد البحث مفيدًا للصناعة المصرفية، حيث يمكن استخدام النموذج للتنبؤ بنجاح مكالمات التسويق عبر الهاتف وفهم الأنماط داخل مجموعة البيانات. علاوة على ذلك، يمكن لصانعي القرار استخدام النموذج لتحديد استراتيجياتهم المستقبلية لتشغيل عملية التسويق عبر الهاتف بكفاءة.

الكلمات المفتاحية: الصناعة المصرفية، استخراج البيانات، التعلم الآلي، التسويق عبر الهاتف.