Recommender System for Term Deposit Likelihood Prediction Using Cross-validated Neural Network

[1]Shawni Dutta and [2]Prof. Samir Kumar Bandyopadhyay

[1] Lecturer, Department of Computer Science, The Bhawanipur Education Society College, Kolkata, India

[2] Academic Advisor, The Bhawanipur Education Society College, Kolkata, India

Abstract-

For enhancing the maximized profit from bank as well as customer perspective, term deposit can accelerate finance fields. This paper focuses on likelihood of term deposit subscription taken by the customers. Bank campaign efforts and customer details are influential while considering possibilities of taking term deposit subscription. An automated system is provided in this paper that approaches towards prediction of term deposit investment possibilities in advance. Neural network(NN) along with stratified 10-fold cross-validation methodology is proposed as predictive model which is later compared with other benchmark classifiers such as k-Nearest Neighbor (k-NN), Decision tree classifier (DT), and Multi-layer perceptron classifier (MLP). Experimental study concluded that proposed model provides significant prediction results over other baseline models with an accuracy of 88.32% and Mean Squared Error (MSE) of 0.1168.

Keywords: **Term deposit subscription, 10-fold stratified cross-validation, Neural network, DT, MLP, k-NN**

1. Introduction

While considering socio-economic structure, banking sector plays significant role to boost up that structure. Generally, banks provide numerous products as well as services to clients. Deposits are one of products those are served to clients and from bank's perspective deposits are essential key points when bank finance topic comes into play. Bank campaign may occur either through direct marketing or mass campaigns. Mass campaigns target at general indiscriminate public and direct marketing campaigns are instigated with the target of a specific group. The problem of direct marketing is very low positive number of responses [1]. Direct marketing is not popular because of its privacy intrusion insecurity which may elicit negative attitude towards bank. Due to evolving structure of telemarketing through Computer-Telephony-Integration techniques, it became quite common and easy to generate a wide variety of reports from marketing campaigns and so to add-up other types of information available for the organizations. Bank offers several deposit schemes like term deposit, recurring deposit, fixed deposit, and deposits in savings account, current account and many more [2]. In this paper, term deposit structure is considered as one of the important investing schemes since it facilitates the bank as well as the customers with subtle amount of profit. The impact of telemarketing campaign on term deposit subscription is taken into consideration by this paper.

A recommender system is proposed in this paper that provides automatic predictions regarding the possibilities of term deposits from client side. A term deposit account is held at bank where money is locked up for definite period of time with higher interest rates than traditional saving accounts. This will definitely benefit customers with maximized profit and the bank sectors will get benefitted in terms of investment. Term deposit is often seen as an outcome of bank market campaigns. Data mining and knowledge discovery processes often play interest role while analyzing and identifying hidden patterns and/or relationship in an enormous amount of data. Bank campaign data can be

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2020 doi:10.20944/preprints202006.0360.v1

analyzed using data mining techniques and term deposit possibilities of clients may be determined beforehand. Knowing term deposit possibilities decided by clients at an early stage assist the bank sectors to look into the matter from different perspective to attract clients towards their term deposit schemes.

A NN [3] based framework is proposed in this paper for determining term deposit probabilities agreed by customers. To address the mentioned problem, NN followed by 10-fold stratified cross validation methodology is implemented as recommender system. Strength of campaign results, customer loan history, job profile, marital status etc. are considered as influential factors while identifying whether customer will place term deposit or not. All these features are given as input to the recommended system implemented in this paper. This implemented classifier model is compared with other baseline models such as k-NN [4], DT [5], and MLP [6].

2. Related Works

Using SPSS Modeler, both classification and clustering models are established by Zhuang et al [2]. In classification, boosted C5.0 model shows the best performance with highest accuracy. Clustering algorithms are applied to identify clients who have subscribed to a term deposit in order to discover and understand customers' behaviours and characteristics, social and economic context attributes. Safia Abbas [7] focused on improving the efficiency of the marketing campaigns and helping the decision makers by reducing the number of features, and predicting the deposit customer retention criteria based on potential predictive rules. By applying DT and Rough set theory (RST) classification module predictive results are obtained. Experimental results implied that because of feature reduction process, RST obtains a better summarization to the data set.

Three classification models such as DT, Naïve Bayes and Support Vector Machines (SVM) are compared in with Moro et al [1] with respect to in terms of ROC (Receiver Operating Characteristics) and Lift curve analysis. SVM obtained the best results. An analysis was applied to extract useful knowledge from this classification model [1]. A logistic regression model is constructed in by Jiang[8] considering relationship between success and other factors. The classifier model predicts the success of bank telemarketing to identify the top consumer set. To measure the effectiveness of prediction, some basic classification including Bayes, SVM, NN and DT are implemented and compared in this study. As a result, the prediction accuracy and the area under ROC curve prove the logistic regression model performs best in classifying than other models [8].

Moro et al [9] approached a concept of LTV to improve the return and investment about bank marketing. Several parameters including recency, frequency and monetary value are considered for this purpose. The provided results in [9] are particularly useful for contact center companies with an improved predictive performance.

A comparative study is drawn by Moro et al [10] among four models such as logistic regression, DT, NN and SVM in terms of two metrics AUC and ALIFT. Concluding study states that the NN presented the best results with AUC=0.8 and ALIFT=0.7. Knowledge extraction confirmed the obtained model as credible and valuable for telemarketing campaign managers.

3. Proposed Methodology

The objective of this study is to determine customer term deposit subscription behaviors in advance. In this context, supervised classification algorithms assist in establishing predictive model by learning

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2020 doi:10.20944/preprints202006.0360.v1

and discovering the relationship between a set of feature variables and a target variable. The feature variables include dominant reasons such as customer's age, job profile, marital status, education field, taken personal loan or not, taken home loan or not, has credit details or not, contact details, related details with the last contact of current campaign in terms of day, month, contact duration, related details with contact details, number of days passed, outcome of previous campaign. The above factors are acquired for identifying customer term deposit subscription which is in turn the target variable of the classification. The framework implemented in this paper proceeds through following series of steps.

3.1. Dataset Used

In order to fulfill the objective of the study, Portugal bank marketing campaigns results are obtained from kaggle [11] as a collection of 45211 numbers of records and each of having 17 attributes. The attributes infer the related factors that affect campaign results. The target variable identifies whether a customer place term deposit or not. Hence a binary classification problem is addressed in this paper. Histogram representations of the attributes present in the dataset are provided in Figure 1. Table1 provides details of attributes present in the dataset in terms of types of attributes and usage of them.

Figure1. Representation of Portugal bank marketing campaigns dataset

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2020 doi:10.20944/preprints202006.0360.v1

Input Variables	Categorical Attributes	job, marital, education, default (has credit in default?), housing (has housing loan?), loan (has personal loan?), contact (contact communication type), month (last contact month of year), day*of*week(last contact day of the week), poutcome(outcome of the previous marketing campaign)
	Numeric Attributes	Age, duration(last contact duration, in seconds), campaign(number of contacts performed during this campaign and for this client), pdays(number of days that passed by after the client was last contacted from a previous campaign), previous(number of contacts performed before this campaign and for this client),
Output Variable	Categorical Attributes	y (has the client subscribed a term deposit?)

Table1: Detailed description of attributes present in the dataset.

Collected data are preprocessed and a multistep procedure is followed for obtaining a balanced dataset. Pre-processing techniques include missing values handling such as unknown values. In order to fit the data into classifier, non-numeric data is transformed into numeric data. This will be followed by scaling values of every feature with large set of data points. Feature scaling will assist the classifier to work using normalized data with an enhanced efficiency. Once this feature scaling operation is performed, feature vector is fitted to classifier model as training purpose.

3.2. Methodology

Classification procedure is applied in this framework that is applied on the Portugal bank marketing campaigns results dataset in order to obtain term deposit subscription prediction in advance. Classification strategy is implemented by designing neural network model followed by 10-fold cross-validation structure.

NN mimics human brain like operations in order to inferring complex problem-solving approach. It recognizes underlying relationships in a set of data which the provision of necessary adaptation of changing input in order to generate the best possible result without altering the output criteria [12]. NN proposed in this paper is comprised of several neurons. Each of these neurons will accept necessary parameters and apply some activation functions in order to produce outputs. Activation functions [13] are useful to perform diverse computations and produce outputs within a certain range. In other words, activation function is a step that maps input signal into output signal. Among several types of activation function, sigmoid and relu are two popular activation functions. A brief description of the functions is discussed as follows-

- Sigmoid activation function [13] transforms input data in the range of 0 to 1 and it is shown in equation (1).

$$f(x) = 1/(1 + exp^{-x}) \quad (1)$$

- ReLu activation function [13] is a faster learning Activation function which is the most successful and widely used function. It performs a threshold operation to each input element where values less than zero are set to zero whereas the values greater or equal to zeros kept as intact and it is shown in equation (2).

$$f(x) = \max(0, x) = \begin{cases} \text{Xi}, & \text{if Xi} \geq 0 \\ 0, & \text{if Xi} < 1 \end{cases} \quad (2)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2020 doi:10.20944/preprints202006.0360.v1

After configuring this neural model, training process is executed. The training process goes through one cycle known as an epoch where the dataset is partitioned into smaller sections. An iterative process is executed through a couple of batch size that considers subsections of training dataset for completing epoch execution.

It is not possible here the use of survey method over different banks in terms of term deposits by customers. If it is possible then the real nature of prediction can be obtained. In practical situation it is not possible for researchers to go through the survey regarding term deposit in various banks since this may need to collect huge data which is impossible to tackle for prediction.

3.3. Implementation

While designing this model it is necessary to tune hyper-parameters in order to achieve maximized efficiency. This section describes specification of the model along with its hyper-parameters. This model consists of three Dense layers with 32,16,1 number of nodes respectively. In this context, sigmoid and relu activation functions are applied in each of these specified layers. The first two layers apply relu as activation function and the final layer applies sigmoid activation function.

Finally these aforementioned layers are compiled using adam solver [14] through 30 epochs and with a batch size of 10. Adjustment of the hyper-parameters assists the model to obtain best predictive result. The NN receives a total of 1,089 parameters and trains those parameters in order to obtain prediction. Components of the model in terms of layers, shape of output data from each layer, and number of parameters received in each layer are described in Figure 2.

Layer (type)	Output Shape	Param #
dense_144 (Dense)	(None, 32)	544
dense_145 (Dense)	(None, 16)	528
dense_146 (Dense)	(None, 1)	17

Figure2. Summary of the model

This implementation is followed by 10-fold cross-validation method [15] for estimating the skill of the model. It is a resampling methodology where the dataset is partitioned into 10 groups and in each iteration one group is considered as the test data and the remaining nine folds are considered as training data. The above-mentioned model is fitted into the training dataset and it is evaluated against the test dataset during each fold separately. Later evaluation scores for each of these iterations are accumulated and mean score is calculated. Rather than splitting data randomly using k-fold validation method, stratified k-fold mechanism [15] is employed in this framework. In stratified k-fold cross validation method, class distributions are managed in such a way that each fold approximately matches the proportion of all labels as the original data.

The proposed NN with 10-fold stratified cross validation methodology as well as DT, k-NN classifiers, MLP classifier are implemented and evaluated in terms of some pre-defined metrics. These metrics will support the comparison platform while inferring the best problem-solving approach.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2020 doi:10.20944/preprints202006.0360.v1

Performance Measure Metrics

In order to justify performance skill of a model, it is necessary to employ metrics to estimate the evaluation. For this purpose, following metrics are taken into consideration in order to identify the best relevant problem-solving approach.

1. Accuracy [16] is a metric that detects the ratio of true predictions over the total number of instances considered. Mathematically, accuracy can be defined as follows with given True Positive, True Negative, False Positive, False Negative as TP, TN, FP, FN respectively-

   Accuracy= TP+TN/(TP+FP+TN+TP)

2. MSE [16] is another evaluating metric that measures absolute differences between the prediction and actual observation of the test samples. It can be defined as follows-

   MSE= $(\sum_{i=1}^{N}(Xi - Xi')^2 / N)$ where Xi is the actual value and Xi' is the predicted value.

A model that exhibits lower MSE value and higher accuracy result turns out to be the best problem-solving approach.

3.4. Results and Discussion

The proposed model is evaluated in terms of performance measure metrics. All the scores obtained during each fold of 10-fold cross-validation are acquired and mean is calculated. The mean score is the final result exhibited by the proposed model. The results are summarized along with specified baseline classifiers such as K-NN, DT, and MLP in Table 1. This indicates that proposed method provides better performance with respect to other classifiers. Accuracy and MSE obtained in each fold with respect to training and testing dataset is shown in Figure 3.

Table1: Performance Summarization of specified classifiers

Models	Performance Measure Metrics	Accuracy	MSE
Proposed Model	NN with 10-fold cross-validation method.	88.32%	0.1168
Baseline Classifier	DT Classifier	83.67%	0.16
	K-NN Classifier	87.16%	0.13
	MLP Classifier	87.13%	0.13

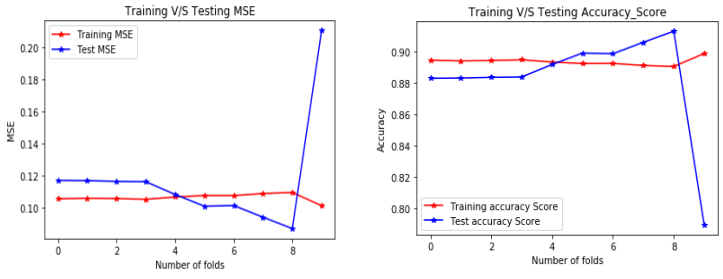

Figure3. Training and Testing accuracy and MSE shown in each iteration of cross-validation

Portugal bank marketing campaigns results are taken from kaggle [11] as a collection of 45211 numbers of records and each of having 17 attributes. The proposed method is entirely based on the dataset. It is not possible to test the method on other dataset since no other dataset is available. It is the main limitation of the method.

 doi:10.20944/preprints202006.0360.v1

4. Conclusions

This study applies data mining techniques to forecast customers' term deposit subscription behaviours and comprehend customers' features to improve the effectiveness and accuracy of bank marketing. NN along with 10-fold cross-validation methodology is implemented under a single platform that determines term deposit subscription behaviours. This implementation is incorporated with necessary parameter tuning as well as data oriented operations. However, this method is compared with baseline classifiers such as MLP, k-NN and DT. The comparative study concludes that implemented method indicates superior result with an accuracy of 88.32% and MSE 0.1168. Predictive results provided by the proposed method assist bank financial sectors to take informed decision in customer attraction process towards term deposit subscription.

5. Abbreviations-

Neural network(NN)

k-Nearest Neighbor (k-NN)

Decision tree classifier (DT)

Multi-layer perceptron classifier (MLP)

Mean Squared Error (MSE)

Rough set theory (RST)

Support Vector Machines (SVM)

ROC (Receiver Operating Characteristics)

lifetime value (LTV)

6. Declarations section

Availability of supporting data-	Through Internet (Open Source)
Competing interests-	There is no Competing interest of any author
Funding-	There is no funding
Authors' contributions	All authors are equally works for making this research paper.
Acknowledgements	University of Calcutta

7. References

[1] S. Moro, P. Cortez, and R. M. S. Laureano, "A Data Mining Approach for Bank Telemarketing Using the rminer Package and R Tool," vol. 2014, no. September, pp. 1–23, 2013.

[2] Q. R. Zhuang, Y. W. Yao, and O. Liu, "Application of data mining in term deposit marketing," *Lect. Notes Eng. Comput. Sci.*, vol. 2, pp. 14–17, 2018.

[3] M. R. M. Veera Manickam *et al.*, "Research study on applications of artificial neural networks

and e-learning personalization," *Int. J. Civ. Eng. Technol.*, vol. 8, no. 8, pp. 1422–1432, 2017.

[4] P. Cunningham and S. J. Delany, "K -Nearest Neighbour Classifiers," *Mult. Classif. Syst.*, no. May, pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6.

[5] H. Sharma and S. Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining," *Int. J. Sci. Res.*, vol. 5, no. 4, pp. 2094–2097, 2016, doi: 10.21275/v5i4.nov162954.

[6] F. Murtagh, "Multilayer perceptrons for classification and regression," *Neurocomputing*, vol. 2, no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925-2312(91)90023-5.

[7] S. Abbas, "Deposit subscribe Prediction using Data Mining Techniques based Real Marketing Dataset," *Int. J. Comput. Appl.*, vol. 110, no. 3, pp. 1–7, 2015, doi: 10.5120/19293-0725.

[8] Y. Jiang, "Using Logistic Regression Model to Predict the Success of Bank Telemarketing," *Int. J. Data Sci. Technol.*, vol. 4, no. 1, pp. 35–41, 2018, doi: 10.11648/j.ijdst.20180401.15.

[9] P. R. S´ergio Moro, Paulo Cortez, "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.

[10] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decis. Support Syst.*, vol. 62, no. February 2014, pp. 22–31, 2014, doi: 10.1016/j.dss.2014.03.001.

[11] Sharan MK (2018, December). Bank Customers Survey - Marketing for Term Deposit,Version 2. Retrieved on June 1,2020 from https://www.kaggle.com/sharanmk/bank-marketing-term-deposit

[12] S. Harvey and R. Harvey, "An introduction to artificial intelligence," *Appita J.*, vol. 51, no. 1, 1998.

[13] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation Functions: Comparison of trends in Practice and Research for Deep Learning," pp. 1–20, 2018.

[14] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

[15] R. H. Kirschen, E. A. O'Higgins, and R. T. Lee, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Am. J. Orthod. Dentofac. Orthop.*, vol. 118, no. 4, pp. 456–461, 2000, doi: 10.1067/mod.2000.109032.

[16] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.