A Comparative Study to the Bank Market Prediction

Soumadip Ghosh[1], Arnab Hazra[2], Bikramjit Choudhury[3](✉), Payel Biswas[4], and Amitava Nag[3]

[1] Department of IT, Academy of Technology, Aedconagar, Hooghly 712121, West Bengal, India
soumadip.ghosh@gmail.com

[2] Department of CSE, Academy of Technology, Aedconagar, Hooghly 712121, West Bengal, India
arnabhazra08@gmail.com

[3] Department of IT, Central Institute of Technology, Kokrajhar, BTAD 783370, Assam, India
b.choudhury@cit.ac.in, amitavanag.09@gmail.com

[4] Department of Computer Science, Jogesh Chandra Chaudhuri College, Kolkata 700033, West Bengal, India
payel.biswas.cs@gmail.com

Abstract. Bank market prediction is an important area of data mining research. In the present scenario, we are given with huge amounts of data from different banking organizations, but we are yet to achieve meaningful information from them. Data mining procedures will help us extracting interesting knowledge from this dataset to help in bank marketing campaigns. This work introduces analysis and applications of the most important techniques in data mining. In our work, we use Multilayer Perception Neural Network (MLPNN), Decision Tree (DT) and Support Vector Machine (SVM). The objective is to examine the performance of MLPNN, DT and SVM techniques on a real-world data of bank deposit subscription. The experimental results demonstrate, with higher accuracies, the success of these models in predicting the best campaign contact with the clients for subscribing deposit. The performance is evaluated by some well-known statistical measures such as accuracy, Root-mean-square error, Kappa statistic, TP-Rate, FP-Rate, Precision, Recall, F-Measure and ROC Area values.

Keywords: Data mining · Classification
Multilayer Perception Neural Network · Decision tree · Support Vector Machine

1 Introduction

Banks keep huge amount of data about their customers. This data can be used to create and keep clear relationship and connection with the customers in order to target them individually for definite products or banking offers. Usually, the selected customers are

P. Perner (Ed.): MLDM 2018, LNAI 10934, pp. 259–268, 2018.
https://doi.org/10.1007/978-3-319-96136-1_21

contacted directly through personal contacts, telephone cellular, mail, and email or any other contacts to advertise the new product/service or give an offer, this kind of marketing is called direct marketing. In fact, direct marketing is in the main a strategy of many of the banks and insurance companies for interacting with their customers. Data mining [1] has gained popularity for illustrative and predictive applications in banking processes. Three techniques will apply to the data set on the bank direct marketing. The Multilayer perception neural network (MLPNN) is one of these techniques, which have their roots in the artificial intelligence. MLPNN is a mutually dependent group of artificial neurons that applying a mathematical or computational model for information processing using a connected approach to computation.

Another technique of data mining is the decision tree approach. Decision tree provides powerful techniques for classification [2] and prediction. There are many algorithms to build a decision tree model [3, 4]. It can generate understandable rules, and to handle both continuous and categorical variables. One of the famous techniques of the decision tree is CART, which will be applied in this work. The third technique is Support Vector Machine (SVM) [5], an algorithm for solving the quadratic programming (QP) problem that arises during the training of support vector machines.

2 Related Works

Osuna et al. [6] proved a theorem which suggests a whole new set of QP algorithms for SVMs. By the virtue of this theorem a large QP problem can be broken down into a series of smaller QP sub-problems to converge to the global optimum. This decomposition algorithm can be used to train SVM on larger dataset.

Moro et al. [7] worked with a large dataset, collected over 2008 to 2013 from a Portuguese retail bank, was addressed which includes the recent financial crisis. They analyzed a large set of 150 features related with bank client, product and social-economic attributes. Because of a semi-automatic feature selection explored in the modelling phase of their method, performed with the data prior to July 2012, the data set was reduced to 22 features (which we are using in our approach). They also compared four DM Models (logistic regression, decision trees (DT), neural network (NN) and support vector machine) using two metrics (area of the receiver operating characteristic curve (AUC) and area of the LIFT cumulative curve (ALIFT)) out of which NN presented the best results (AUC = 0.8 and ALIFT = 0.7), allowing to reach 79% of the subscribers by selecting the half better classified clients.

Hu [8] applied data mining techniques to help retailing banks for attrition analysis to identify a set of customer having high probability to attrite. He has used decision tree (DT), boosted naive Bayesian network, selective Bayesian network, neural network as data mining model.

Ling and Li [9] used data mining techniques for direct marketing in three datasets form three different sources. The first dataset was for a loan product promotion in Canada. Second dataset was from a major life insurance company and third dataset was from a company which runs a particular “bonus program”. Two learning algorithms (ADA-boosted Naive Bayes and ADA-boosted C4.5 with CF) that also produce probability had been used.

According to Turban et al. [10] business intelligence includes architectures, tools, databases, applications and methodologies with the goal of using data to support decisions of business managers. Data mining is a business intelligence technology that uses data-driven models to extract useful knowledge i.e., patterns from complex and large dataset [11].

Chitra and Subashini [12] employed some data mining algorithms for customer retention, automatic credit approval, fraud detection, marketing and risk management in banking sector. They have identified some procedures and models to improve customer retention and to fraud detection.

Rafiqul Islam and Ahsan Habib [13] applied DT approach to predict prospective business sector for lending in retail banking. They have used data from different branches of a bank and analysed borrowers’ transactional behavioural data.

3 Dataset Description

The present work is related with direct marketing campaigns of a Portuguese banking institution. We have taken this dataset from the University of California at Irvine (UCI) Machine Learning Repository (as given below in Table 1). The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (yes) or not (no) subscribed. There are four datasets:

1. “bank-additional-full.csv” with all examples (41188) and 20 inputs, ordered by date (from May 2008 to Nov 2010), very close to the data analyzed.
2. “bank-additional.csv” with 10% of the examples (4119), randomly selected from dataset 1 as mentioned above), and 20 inputs.
3. “bank-full.csv” with all examples and 17 inputs, ordered by date (older version of this Dataset with fewer inputs).
4. “bank.csv” with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

The smallest datasets are provided to test more computationally demanding machine learning algorithm (e.g. SVM). The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y). This UCI dataset is used to evaluate the performances of the multilayer perception neural network (MPLNN), decision tree and SVM classification model. The description of each of the attributes in the dataset is given in Table 1.

Table 1. Attribute description

Sl. no.	Attribute	Meaning	Type
1	age	age of the person	Numeric
2	job	type of job	Categorical
3	marital	marital status	Categorical
4	education	highest education achieved	Categorical
5	default	has credit in default? (categorical: 'yes', 'no')	Binary (Categorical)
6	housing	has housing loan? (categorical: 'yes', 'no', 'unknown')	Binary (Categorical)
7	loan	has personal loan? (categorical: 'yes', 'no', 'unknown')	Binary (Categorical)
8	contact	contact communication type (categorical: 'cellular', 'telephone')	Binary (Categorical)
9	month	last contact month of year	Categorical
10	day_of_week		Numeric
11	duration	last contact duration, in seconds	Numeric
12	Campaign	number of contacts performed during this campaign and for this client	Numeric
13	pdays	number of days that passed by after the client was last contacted from a previous campaign	Numeric
14	previous	number of contacts performed before this campaign and for this client	Numeric
15	poutcome	outcome of the previous marketing campaign	Categorical
16	emp.var.rate	employment variation rate- quarterly indicator	Numeric
17	cons.price. idx	consumer price index- monthly indicator	Numeric
18	cons.conf.idx	consumer confidence index- monthly indicator	Numeric
19	euribor3 m	euribor 3 month rate - daily indicator	Numeric
20	nr.employed	number of employees - quarterly indicator	Numeric
21	Output	has the client subscribed a term deposit? (binary: 'yes', 'no')	Binary (Categorical)

4 Proposed Method

Three techniques namely MLPNN, DT and SVM are applied to the data set for bank marketing prediction. The detailed procedure is divided into two major steps. First one is data pre-processing and then data classification. Figure 1 below depicts the proposed methodology of our system.

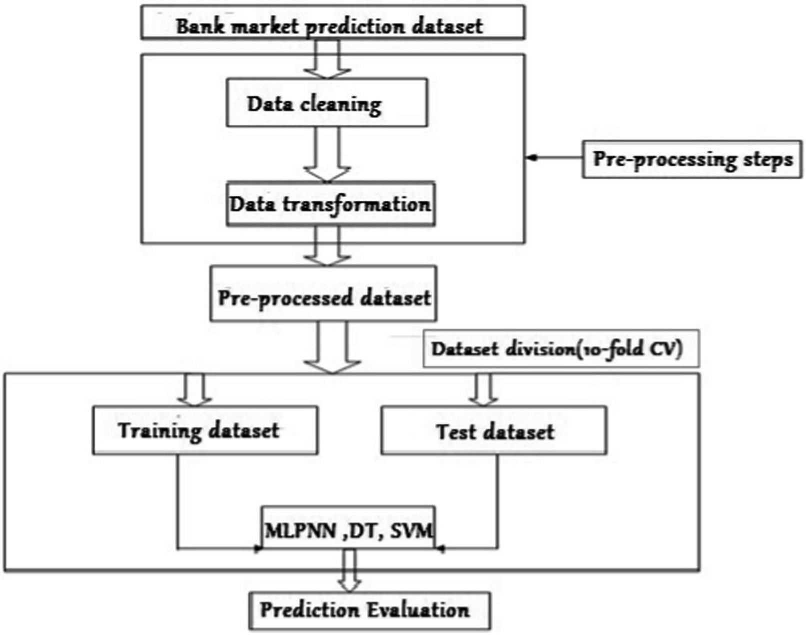

Fig. 1. Proposed methodology of the system using different classifiers

Step 1: Data pre-processing

Data pre-processing techniques are applied to the original dataset before the data classification procedure. It may involve different techniques such as data cleaning and data transformation.

(1a) Data cleaning: Data cleaning denotes the pre-processing of data for removing or reducing noise and the handling the missing values. A missing value is typically replaced by the mean value for that attribute based on statistics. This step is not required in our work as there are no missing or inconsistent values present.

(1b) Data transformation: In data transformation step, the dataset is normalized as because the ANN based technique requires distance measurements in the training phase. It transforms attribute values to a small-scale range like −1.0 to +1.0.

Step 2: Data classification

After pre-processing steps are over, the original data set is divided into two sub-sets namely the training data set and the test data set. We apply 10-fold cross-validation technique for data distribution so as to generate training and test datasets separately. In classification step, firstly the mathematical model of the classifiers is initialized with default control parameters. After initialization is over, they are trained using the training tuples of training dataset. And after the training phase, they are tested with unknown tuples of test dataset as test input to obtain predicted class label. This label is compared with the actual class label to estimate the accuracy of the classifiers being used. The configuration parameters of MLPNN, DT and SVM are given below.

For MLPNN classifier we have,

$$\mathrm{H} = \sqrt{\mathrm{I}\,.\,\mathrm{O}} \tag{1}$$

Where H, I and O denotes the number neurons in the hidden layer, number of input and output attributes respectively. Table 2 describes different metrics of MLPNN model.

Table 2. Different metrics of the MLPNN model

Metric	Value
Number of hidden layers	One
Number of neurons in input layer	Number of input attributes
Number of neurons in output layer	Data classes present
Learning rule	Gradient descent with momentum
Transfer function used	Tan-sigmoid

After construction of the tree, minimal cost complexity pruning algorithm to be used is a post-pruning approach. This algorithm produces a decision tree classifier with minimum cost complexity. Table 3 describes different metrics used in the CART model. All these metrics have their usual meanings.

Table 3. Different metrics of the CART model

Metric	Value
Attribute selection measure	Gini index
Minimal number of instances at terminal nodes	2
Pruning approach used	Post-pruning approach
Pruning algorithm name	Minimal cost complexity pruning
Number of folds used	5
random seed number	1

Different possible combinations like the number of folds used, value of random seed, and different kernel based techniques are investigated here for developing an SVM classifier. Then an SVM model with a Gaussian radial-basis function (RBF) kernel is selected. A non-linear version of SVM can be represented by using a kernel function K as:

$$K\left(x_i, x_j\right) = \phi(x_i)\,.\,\phi\left(x_j\right) \tag{2}$$

Here $\varphi(x)$ denotes non-linear mapping function employed to map the training instances. An SVM model with a Gaussian RBF kernel is defined as:

$$K\left(x_i \cdot x_j\right) = e\frac{-\left\|x_i - x_j\right\|^2}{2\sigma^2} \tag{3}$$

Table 4 below provides different metrics used in the given SVM model. All these metrics have their usual meanings.

Table 4. Different metrics of the SVM model

Metric	Value
Type of kernel used	Non-linear
Kernel name	Gaussian radial-basis function (RBF)
Cache size	250007
Value of σ	0.01
Complexity parameter	1.0
Number of folds used	−1
Random seed value	1
Epsilon value for round-off error	1.0e−12
Tolerance Parameter	0.001

5 Results and Discussion

Here three classifiers namely Multilayer Perceptron Neural Network, Decision Tree and Support Vector Machine are applied to the UCI machine learning repository data set for investigation and performance analysis. Here we have divided the data set into training purpose and testing purpose. The results described here are exclusively based on the simulation experiment that we have taken. We have done several comparison of these classifiers based on some performance measures like classification accuracy, root-mean square error (RMSE) [14], kappa statistic [15] values. And also performed detail accuracy by each class for three classifiers using True Positive Rate (TP-Rate) or Recall, False Positive Rate (FP-Rate), Precision, F-Measure and ROC area values derived from the confusion matrix [16] of each classifier. Three classifiers (MLPNN, DT and SVM) are applied to a test set for classification after completion of the training phase. Firstly, we perform comparisons of these classifiers which are based on said performance measure as shown below in Table 5.

Table 5. Performance comparison of three classifiers

Classifier	Accuracy	Kappa Statistics	RMSE
MLPNN	87.2%	0.7842	0.3018
DT	92.8%	0.8269	0.2678
SVM	86.3%	0.7736	0.3127

From Table 5 we could see that, accuracy of MLPNN, DT and SVM are 87.2%, 92.8% and 86.3% respectively. So it is clear that accuracy wise DT has performed better than MLP and SVM here. Based on the result, DT comes out first with an RMSE value of 0.2678 and a kappa statistic value of 0.8269; followed by MLPNN having an RMSE value of 0.3018 and a kappa statistic value of 0.7842 and SVM stands last with the highest RMSE value 0.3127 and the lowest kappa statistic value 0.7736. Therefore, with regard to the performance measures such as classification accuracy, RMSE and kappa statistic, the DT classifier has performed the best. Figure 2 shows a performance comparison among three classifiers.

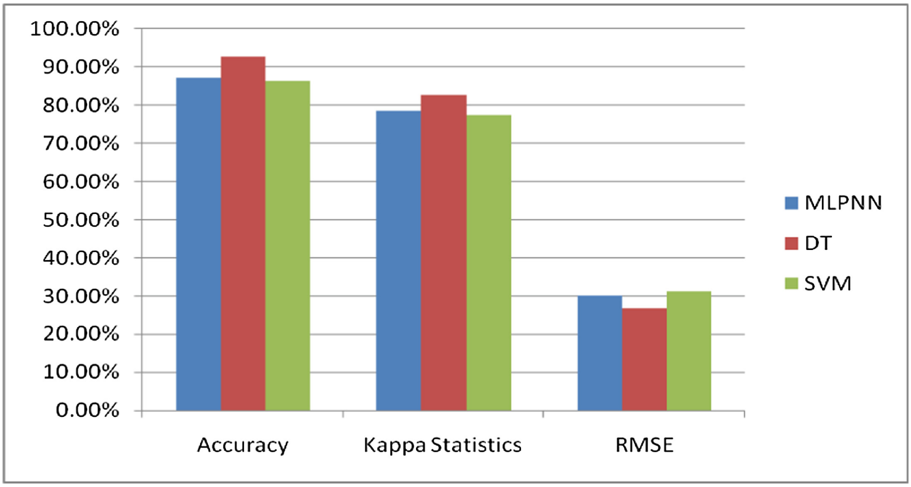

Fig. 2. Performance comparison of three classifiers

After that we have compared these models based on the TP-Rate (or Recall), FP-Rate, Precision, F-Measure and ROC area values derived from the confusion matrix of individual with respect to the test data set.

Table 6. Detailed accuracy by each of the three classifiers

Classifier	TP-Rate/Recall	FP Rate	Precision	F-Measure	ROC Area
MLPNN	87.2%	12.8%	87.2%	87.2%	0.867
DT	92.8%	9.2%	92.8%	92.8%	0.931
SVM	86.3%	13.9%	86.3%	86.3%	0.845

From Table 6 we could discover that the weighted average values of TP-Rate (or Recall), FP-Rate, Precision, F-Measure, and ROC Area for MLPNN classifier are 87.2%, 12.8%, 87.2%, 87.2% and 0.867 respectively; whereas for DT classifier the values are 92.8%, 9.2%, 92.8%, 92.8% and 0.931 respectively. For SVM these values are 86.3%, 13.9%%, 86.3%, 86.3%, and 0.845 respectively. Detail accuracy by the three classifiers is also shown in 2-D column chart in Fig. 3.

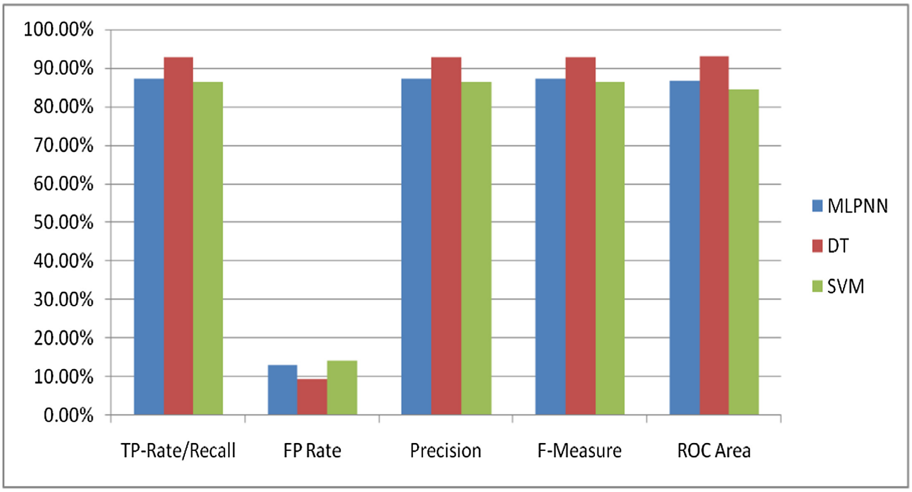

Fig. 3. Detailed accuracy by each of the three classifiers

The DT model has the highest weighted average values for TP-Rate, Precision and F-Measure and the lowest weighted average value for FP-Rate. Indeed, the classification accuracy value of DT model is considerably better (more than 5%) compared to the other models.

6 Conclusion

Bank market prediction is needed for bank deposit subscription, customer relationship management, fraud detection and building marketing strategies. This kind of prediction is certainly helpful for running the business successfully. This paper uses procedures to evaluate and compare the classification performance of three different data mining techniques using models such as MPLNN, Decision tree and SVM on the bank direct marketing dataset. The purpose is to increase the effectiveness of the bank marketing campaigns by identifying the main characteristics that affect the success. These classifiers mainly aim to indicate the deposit money subscribed by the clients. The classification performances of these three models have been evaluated using several useful statistical measures. Finally, experimental results have shown the effectiveness of DT model compared to MLPNN and SVM models. In fact, the accuracy value of DT model is significantly higher (more than 5%) compared to the other classification models used here. So, the DT model developed using CART technique could be very much helpful for direct bank market prediction.

References

1. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, Burlington (2000)
2. Pujari, A.K.: Data Mining Techniques, 1st edn. Universities Press (India) Private Limited, Hyderabad (2001)
3. Quinlan, J.R.: Simplifying decision trees. Int. J. Man Mach. Stud. **27**(3), 221–234 (1987)
4. Breiman, L., Freidman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees Belmont. Wadsworth, Belmont (1984)
5. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
6. Osuna, E., Freund, R., Girosi, F.: An improved training algorithm for support vector machines. In: IEEE NNSP 1997, Amelia Island, FL, pp. 24–26, September 1997
7. Moro, S., Laureano, R., Cortez, P.: Using data mining for bank direct marketing: an application of the CRISP-DM methodology. In: Novais, P., et al. (Eds.), Proceedings of the European Simulation and Modelling Conference – ESM 2011, Guimarães, Portugal, pp. 117–121, October 2011
8. Hu, X.: A data mining approach for retailing bank customer attrition analysis. Appl. Intell. **22**(1), 47–60 (2005)
9. Ling, C.X., Li, C.: Data mining for direct marketing: problems and solutions. In: Proceedings of the 4th KDD Conference, pp. 73–79. AAAI Press (1998)
10. Turban, E., Sharda, R., Delen, D.: Decision Support and Business Intelligence Systems, 9th edn. Prentice Hall Press, Upper Saddle River (2010)
11. Witten, I., Frank, E.: Data Mining – Practical Machine Learning Tools and Techniques, 2nd edn. Elsevier, New York (2005)
12. Chitra, K., Subashini, B.: Data mining techniques and its applications in banking sector. Int. J. Emerg. Technol. Adv. Eng. **3**(8), 219–226 (2013). ISSN 2250-2459, ISO 9001:2008 Certified Journal
13. Rafiqul Islam, M., Ahsan Habib, M.: A data mining approach to predict prospective business sectors for lending in retail banking using decision tree. Int. J. Data Min. Knowl. Manag. Process (IJDKP) **5**(2), 13–22 (2015)
14. Armstrong, J.S., Collopy, F.: Error measures for generalizing about forecasting methods: empirical comparisons. Int. J. Forecast. **8**, 69–80 (1992)
15. Carletta, J.: Assessing agreement on classification tasks: the Kappa statistic. Comput. Linguist. **22**(2), 249–254 (1996). MIT Press, Cambridge
16. Stehman, S.V.: Selecting and interpreting measures of thematic classification accuracy. Rem. Sens. Environ. **62**(1), 77–89 (1997)