ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

Evaluation and Comparison of Data Mining Techniques Over Bank Direct Marketing

Niharika Sharma[1], Arvinder Kaur[2], Sheetal Gandotra [3], Dr Bhawna Sharma[4]

B.E. Final Year Student, Department of Computer Engineering, Government College of Engineering & Technology, Jammu, India[1]

B.E. Final Year Student, Department of Computer Engineering, Government College of Engineering & Technology, Jammu, India[2]

Assistant Professor, Department of Computer Engineering, Government College of Engineering & Technology, Jammu, India[2]

Assistant Professor, Department of Computer Engineering, Government College of Engineering & Technology, Jammu, India[2]

ABSTRACT: All bank marketing campaigns are dependent on customers' huge electronic data. It is impossible for a human analyst to come up with interesting information that will help in the decision-making process due to the size of these data sources. Data mining models are completely helping in the performance of these campaigns. Data mining techniques have good prospects in their target audiences and improve the likelihood of response. In this paper, performance of various data mining techniques is compared.The business goal is to find a model that can explain success of a contact, i.e., if the client subscribes the deposit. Such model can increase campaign efficiency by identifying the main characteristics that affect success, helping in a better management of the available resources (e.g. human effort, phone calls, time) and selection of a high quality and affordable set of potential buying customers.

KEYWORDS:Data Mining Techniques, Bank Direct Marketing, ROC Curve, Confusion Matrix, Machine Learning.

I. INTRODUCTION

Data mining[3][4] is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions.

1. Bank direct marketing[1][6]There are two main approaches for enterprises to promote products and/or services: through mass campaigns, targeting general indiscriminate public or directed marketing, targeting a specific set of contacts . Directed marketing focuses on targets that assumable will be keener to that specific product/service, making this kind of campaigns more attractive due to its efficiency. This paper aims at bringing an improvement in efficiency: making lesser contacts, but achieving a better number of successes(clients subscribing the deposit).

2. Machine learning and classifiers[17] The area of machine learning constitutes a number of paradigms and algorithms for classification and learning. A simple and general explanation of these algorithms is that they are used to learn from data to be able to classify instances of data into different categories (classes).A classifier is built by letting a learning algorithm generalize from a set of data (often referred to as training data). The training data consists of a number of instances. A particular instance is described by a set of attribute values. There exist several types of values and classification types. Attributes can consist of numerical values, Boolean values or other types of values. One of the attributes is often referred to as the target attribute. In order words, a classifier should be able to predict the value of the target attribute of an instance, given the values of some or all of the other attributes of the instances

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

3. Classifier comparison One way to find a good solution for a classifier learning problem is to compare the performance of different classifiers on the same data. A simple comparison could be made by training a number of classifiers on the same data set and comparing their accuracy over the test data.

II. DATA MINING TECHNIQUES

The brief description of various machine learning classification techniques used for comparison in this paper for bank direct marketing is given below.

1.Neuralnetworks[19][22]. An artificial neural networks ANN model emulates a biological neural network Neural network can be trained to perform a particular function by adjusting the values of the connections (weights) between elements.

2.Logistic regression[10][12]**.** LR uses maximum probability estimation LR is an approach to learning functions of the form: F: A →Y, or P (Y|A) in the case where Y is discrete-valued as a target, and A = (A1, A2,…, An) is any attribute containing discrete, flag or continuous independent attributes.

3.Discriminant analysis.[19] Discriminate Analysis is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).This is generalized linear type model that uses statistical analysis to predict an event based on known factors .A Discriminate Analysis can make predictions about whether a customer will buy a product based on age, gender, geography and other demographic data.

4.Decision trees.[9] Decision Trees are considered to be one of the most popular approaches for representing classifiers. Researchers from various disciplines such as statistics, machine learning, pattern recognition, and Data Mining have dealt with the issue of growing a decision tree from available data. Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values.

5.SVM.[20]Support vector machine are basically binary classification algorithms. The basic support vector machine takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. An SVM classifies data by finding the best hyper plane that separates all data points of one class from those of the other class.

6.Tree bagger.[17]Bagging stands for bootstrap aggregation. Every tree in the ensemble is grown on an independently drawn sample of input data. To compute prediction for the ensemble of trees, Tree Bagger takes an average of predictions from individual trees (for regression) or takes votes from individual trees (for classification). Ensemble techniques such as bagging combine many weak learners to produce a strong learner.

7. Nearest neighbours.[21]Categorizing query points based on their distance to points in a training dataset can be a simple yet effective way of classifying new points. Various distance metrics such as Euclidean, correlation, hamming or your own distance metric may be used.

III. METHODOLOGY

The learning process[2] used in this paper is comprised of a **data preprocessor** and a **learning algorithm[24]**. First dataset is divided into training and test set then training set it fed to classifier.The trained classifier is then used to make a prediction on the test dataset. Predicted values are compared with actual values to compute the confusion matrix. Confusion matrix is used to visualize the performance of a machine learning techniques. This paper analyzes the performance of different classification techniques to select the one with the most accurate results for classification of bank direct marketing dataset.

DATA DESCRIPTION: Proposed work extracts the datasets of bank direct marketing from UCI repository[7]. It has dimensions of 16 attribute and 45,211 instances. For purpose of training and testing, only 40% of the overall data is used for training and testing the accuracy of the selected classification algorithms. The detailed description of the data set is summarized as follows-

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

The data set contains 45211 observations capturing following 16 attributes/features.

1. age (numeric)
2. job: type of job (categorical: "admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services")
3. marital: marital status (categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed)
4. education (categorical: "unknown", "secondary", "primary", "tertiary")
5. default: has credit in default? (binary: "yes","no")
6. balance: average yearly balance, in euros (numeric)
7. housing: has housing loan? (binary: "yes","no")
8. loan: has personal loan? (binary: "yes","no")
9. contact: contact communication type (categorical: "unknown","telephone","cellular")
10. day: last contact day of the month (numeric)
11. month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12. duration: last contact duration, in seconds (numeric)
13. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
15. previous: number of contacts performed before this campaign and for this client (numeric)
16. poutcome: outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success")

Output variable (desired target):

y: has the client subscribed a term deposit? (binary: "yes","no")

IV. EXPERIMENTAL RESULTS

The performance of each classification model is evaluated using three statistical measures: classification accuracy, sensitivity and specificity[15]. These measures are defined by a confusion matrix that contains information about actual and predicted classifications done by a classification system. The statistical measures are calculated using true positive (TP), true negative (TN),false positive(FP)and false negative(FN) of confusion matrix.The percentage of Correct/Incorrect classification is the difference between the actual and predicted values of variables. Table 1shows the confusion matrix for a two-class classifier.

Table1. The confusion matrix for a two-class classifier

Classifier / Predicted class	TRUE CLASS	
	p (positive)	n (negative)
Y	True Positives	False Positives
N	False Negatives	True Negatives
Total	P	N

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

Table2 following gives the formulation of required statistical measures which form the basis of comparison

Table.2 Formulas

$$TPR = \frac{TP}{P} = Recall,\ FPR = \frac{FP}{N}$$

$$Precision = \frac{TP}{TP + FP},\ Accuracy = \frac{TP + TN}{P + N}$$

$$Sensitivity = Recall,\ Specificity = 1 - FPR$$

Classifier's prediction results. Table3 shows the confusion matrices generated from various data mining techniques i.e. Neural network, logistic regression, discriminant analysis,k-nearest neighbours, naïve bayes,support vector machine, decision trees and tree bagger respectively.

Table3. Confusion matrices for different data mining techniques

```
C_nn =

   96.7742    3.2258
   69.8980   30.1020

C_glm =

   97.5806    2.4194
   67.8571   32.1429

C_da =

   90.2605    9.7395
   52.5510   47.4490

C_knn =

   94.7891    5.2109
   72.4490   27.5510

C_nb =

   90.8809    9.1191
   50.0000   50.0000

C_svm =

   97.6427    2.3573
   87.7551   12.2449

Elapsed time is 0.650927 seconds.

C_t =

   92.6799    7.3201
   59.6939   40.3061

C_tb =

   93.6104    6.3896
   51.5306   48.4694
```

Calculations are performed using the formulas given in Table 2 on the results of various confusion matrices obtained. The comparison for different techniques on calculated measures is carried out in the given Table 4.TREE BAGGER (ENSEMBLE METHOD) comes out to be the best method.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

Table 4.Comparison table for performance analysis

	ACCURACY	SENSITIVITY	SPECIFICITY	PRECISION
NEURAL NETWORKS	0.63	0.58	0.90	0.97
LOGISTIC REGRESSION	0.64	0.59	0.93	0.98
DISCRIMINANT ANALYSIS	0.69	0.63	0.83	0.90
K-NEAREST NEIGHBORS	0.61	0.57	0.84	0.95
NAÏVE BAYES METHOD	0.70	0.65	0.85	0.91
SVM	0.54	0.52	0.84	0.98
DECISION TREES	0.66	0.60	0.85	0.93
TREE BAGGER (ENSEMBLE METHOD)	**0.71**	**0.65**	**0.88**	**0.94**
REDUCED FEATURE SET(TB)	**0.77**	**0.73**	**0.82**	**0.86**

Comparison plot for comparing all techniques comes out in the form of a bar graph that visualises the confusion matrix as shown in Fig 1.**ROC curves**: ROC[18] stands for receiver operating characteristics. Fig 2 shows the ROC curve generated for the ensemble method (tree bagger), the method showing the best results for statistical measures.

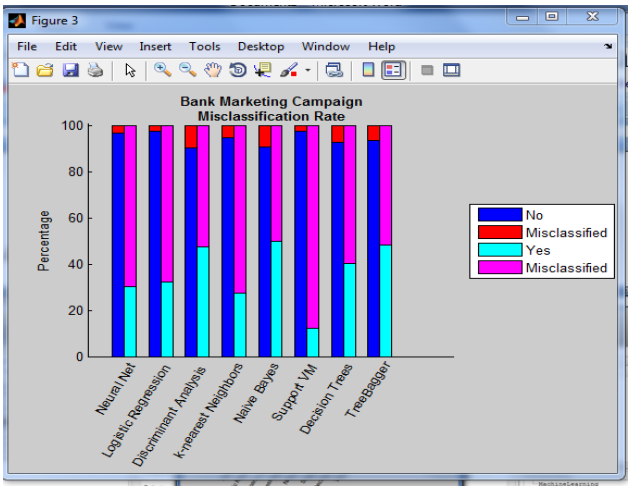

Fig.1 Comparison plot

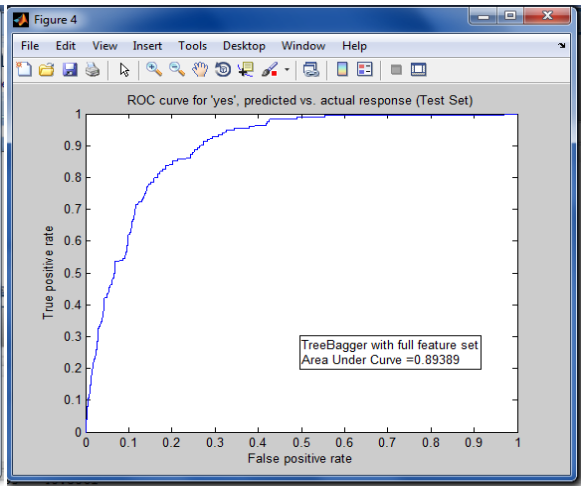

Fig.2 ROC curve for tree bagger

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

Most relevant features which affects the response are also selected by applying feature selection. As shown in Fig 3 the feature with most importance comes out to be 'month of last contact', then 'contact duration' and so on as given in plot. Such knowledge can be used by managers to enhance campaigns (e.g. by asking agents to increase the length of their phone calls or scheduling campaigns to specific months).

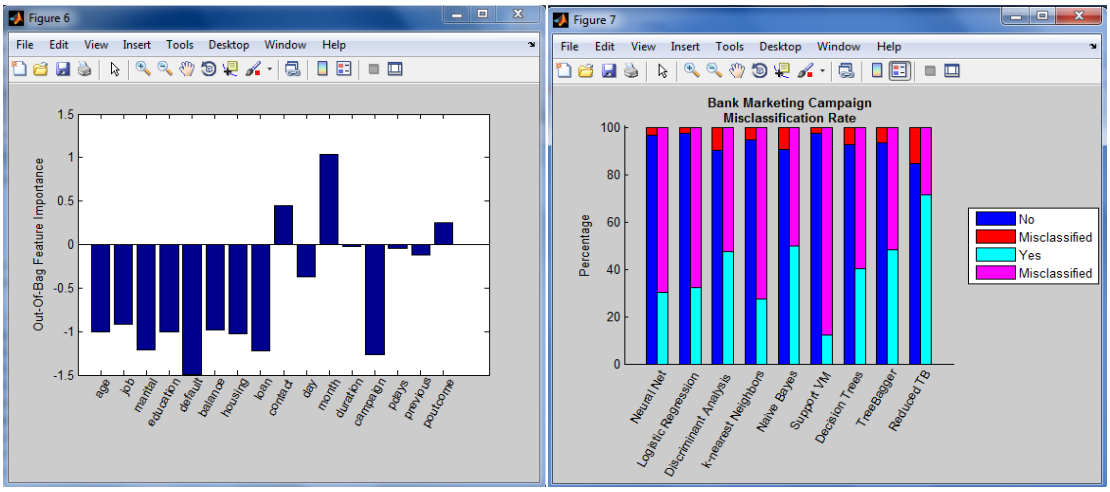

Fig.3. Feature importance plot.Fig.4.comparison plot with reduced feature set

Top five features determined are then included to reduce the number of combinations to be tried by sequentials. The **comparison plot** involving the tree bagging with reduced feature set is also plotted as given in Fig.4. This is used to Fig 5. Shows **the ROC curve** for tree bagger with reduced feature set. And it is evident from there that AUC is not affected much by reducing the feature set.

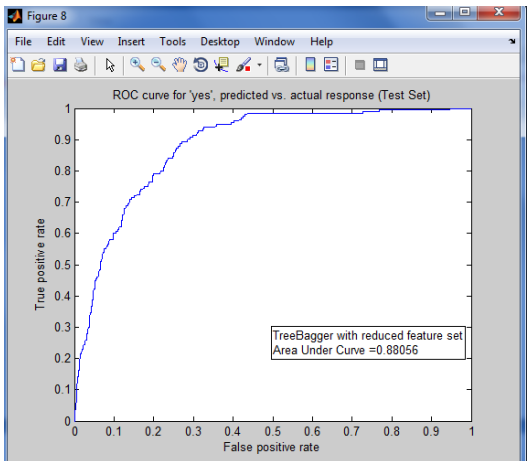

Fig.5 ROC for tree bagger with reduced feature set

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

V. CONCLUSION

Bank direct marketing and business decisions are more important than ever for preserving the relationship with the best customers. For success and survival in the business, there is a need for customer care and marketing strategies. Data mining and predictive analytics can provide help in such marketing strategies. These applications are influential in almost every field containing complex data and large procedures. It has proven the ability to reduce the number of false positives and false-negative decisions. Proposed work evaluatesand compares the classification performance of various data mining techniques models on the bank direct marketing data set to classify for bank deposit subscription. The purpose is increasing the campaign effectiveness by identifying the main characteristics that affect the success (the deposit subscribed by the client). The classification performances of the models are measured using statistical measures: Classification accuracy, sensitivity and specificity. Experimental results have shown the effectiveness of models. **ENSEMBLE METHOD(TREE BAGGING)** achieves slightly better performance than rest of the models.

REFERENCES

[1] C. X. Ling and C. Li, "Data Mining for Direct Marketing: Problems and Solutions Proceedings of International Conference on Knowledge Discovery from Data (KDD 98), New York City, 27-31 August 1998, pp. 73-79.
[2] S. Moro, R. Laureano and P. Cortez, Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology, In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modeling Conference - ESM'2011, pp. 117-121, 2011.
[3] Berson, "Data Warehousing, Data-Mining & OLAP", TMH
[4] AdemKarahoca, DilekKarahoca and MertŞanver, "Data Mining Applications in Engineering and Medicine", ISBN 978-953-51-0720-0, In Tech, August 8, 2012.
[5] A. Trnka, "Market Basket Analysis with Data Mining Methods," International Conference on Networking and Information Technology (ICNIT), 2010, pp. 446-450.
[6] Bult, J.R. (1993). Target Selection for Direct Marketing. Ph.D. Thesis, Rijksuniversiteit Groningen, the Netherlands.
[7] http://archive.ics.uci.edu/ml (The UCI Machine Learning Repository is a collection of databases)
[8] http://en.wikipedia.org/wiki/Naïve_Bayes_classifierWikipedia has a tool to generate citations for particular articles related to Naïve Bayes classifier.
[9] Baik, S. Bala, J. (2004), A Decision Tree Algorithm for Distributed Data Mining: Towards Network Intrusion Detection, Lecture Notes in Computer Science, Volume 3046, Pages 206 – 212.
[10] D.G. Kleinbaum and M. Klein, Logistic Regression, Statistics for Biology and Health, DOI 10.1007/978-1-4419-1742-31, Springer Science Business Media, LLC 2010.
[11] Derrig, Richard A., and Louise A. Francis, "Distinguishing the Forest from the TREES: A Comparison of Tree-Based Data Mining Methods," Variance 2:2, 2008, pp. 184-208.
[12] Domínguez-Almendros S. "LOGISTIC REGRESSION MODELS". AllergolImmunopathol (Madr). 2011. doi:10.1016/j.aller.2011.05.002.
[13] Eniafe Festus Ayetiran, "A Data Mining-Based Response Model for Target Selection in Direct Marketing", I.J.Information Technology and Computer Science, 2012, 1, 9-18.
[14] Harry Zhang., Liangxiao Jiang., Jiang Su. "Augmenting Naïve Bayes for Ranking". Published in: Proceeding of ICML '05 Proceedings of the 22nd international conference on Machine learning Pages 1020 – 1027, 2005
[15] Bouckaert, R. (2003). Choosing between two learning algorithms based on calibrated tests. Proc 20th IntConf on Machine Learning, pp. 51-58. Morgan Kaufmann.
[16] Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984) Classification and Regression Trees, Wadsforth International Group.
[17] Breiman, L., Bagging Predictors. Machine Learning, 24 (1996) 123-140.
[18] Fawcett, T. 2005. "An introduction to ROC analysis", Pattern Recognition Letters 27, No.8, 861–874.
[19] Bishop, C.M. (1995). Neural Networks for Pattern Recognition. Clarendon Press, Oxford.
[20] Zahavi, J. and N.Levin (1995), Issues and Problems in Applying Neural Computing to Target Marketing, Journal of Direct Marketing, 9(3), 33-45.
[21] Velmurugan T., T. Santhanam(2010), performance evaluation of k-means & fuzzy c-means clustering algorithm for statistical distribution of input data points., European Journal of Scientific Research, vol 46
[22] TIAN YuBo, ZHANG XiaoQiu, and ZHU RenJie. "Design of Waveguide Matched Load Based on Multilayer Perceptron Neural Network". Proceedings of ISAP, Niigata, Japan 2007.