IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF BANK CUSTOMER'S BEHAVIOUR USING ENSEMBLE TECHNIQUES

[1]Ashalata Panigrahi, [2]Sasmita Panigrahi
[1]Associate Professor, [2]Faculty of Information Technology
Master in Computer Application
Roland Institute of Technology, Berhampur, India

Abstract: The prediction of bank performance is very critical because wrong prediction can create serious problems for the bank and society. The objective of the study is to develop a predictive model to predict which customer will deposit longterm in a bank. In this study, we propose ensemble learning techniques to construct a prediction model using three classifiers, namely, adaboostM1, bagging, and dagging. Further, the most important features are selected using six rank based feature selection methods namely, one-R, symmetrical uncertainty (SU), gain ratio, information gain, chi-squared evaluator, and relief-F. After feature selection normalization procedure applied on the bank marketing dataset. Performance of different combinations of classifiers and feature selection methods are compared using the evaluation criteria Accuracy, FPR, Specificity, NPV, FNR, Error rate, and ROC.

Index Terms: Bagging, Dagging, Normalization, Ensemble learning, Chi-square, Information gain

1. INTRODUCTION

Today's business environment is highly competitive and changing rapidly due to innovations of technologies and use of internet. Collection of customer information is necessary for development of marketing strategies. Banks collect customers information using different channels such as e-mail, phones such as fixed line or mobile for sharing information about the products or services . Now-a- days the behaviour and preferences of customers are continuously changing that create a pressure on banks. Bank stores records of all the information about their customers to improve bank strategies and maintain a good relationship. Bank provides different types of facilities like short-term, long-term loans, home loans, personal loans, retirement plan etc credit cards, debit cards so that more customers will be attracted. The objective of the study is to develop a predictive model to predict which customer will deposit longterm in a bank.

The remaining of this paper is structured as follows; section 2 presents the related work. Methodology is presented in section 3. Section 4 presents different steps for the proposed approach. Section 5 deals with dataset description, evaluation metrics. Section 6 describes analysis of results and finally section 7 draws conclusion.

2. RELATED WORK

Wisaeng, 2013 compared four data mining models: SVM, J48 graft,LAD Tree, and radial basis function network. The experimental study is based on bank marketing dataset. J48 gives highest accuracy of 76.52%. Asare-Frempong and Jayabalan (2017) proposed model based on four classification techniques namely, MLPNN, Decision Tree (C4.5), Logistic Regression and Random Forest (RF). Results showed that Random Forest Classifier with an accuracy of 87% is the better predictive ability among the four classifiers. Palaniappan et al. (2017) compared the performance of three different classification techniques namely, Naïve Bayes, Random Forest, and Decision Tree. At the preprocessing stage normalization applied on the bank direct marketing dataset. The experimental result shows that Naïve Bayes has the lowest accuracy of 86.27% and Decision Tree has the highest accuracy with 90.68%. Panigrahi et al. (2020) proposed a prediction model using six neural network based classifiers, namely, SMO, SVM, RBFN, MP, SOM, and HLVQ. At the preprocessing stage three feature selection methods such as filtered attribute evaluator, one-R attribute evaluator, Relief-F attribute evaluator applied on the dataset for selection of important features and reduce processing time. The result shows that filtered attribute evaluator with multilayer perceptron gives highest accuracy of 90.0179%. Islam et al. (2019) propose prediction model using SMOTE algorithm to balance the dataset and analyse the performance using Naive Bayes algorithm. For experiment bank telemarketing dataset is used. The dataset consists of 45211 instances and 17 attributes. They achieved the best accuracy by using the Gaussian NB algorithm of 88.86%. Liu et al. (2017) proposed the prediction model using FMLP-SVM method. For experiment bank telemarketing dataset is used. Experimental result shows that fuzzy SVM algorithm outperforms the traditional SVM with 92.89 % predicting accuracy rate.

3. ENSEMBLE LEARNING CLASSIFICATION TECHNIQUES

The primary objective of ensemble learning is to improve the detection accuracy and reduce the false alarm values of predictive classifiers by combining the strengths and capabilities of various weak learners to achieve a robust, efficient, and effective classifier (Zhou et al., 2020).

3.1 Adaboost Algorithm

Adaboost algorithm (Freund et al., 1995) developed for improving the performance of weak classifiers. Starts with the original data set and uses a learning algorithm to construct a classifier. A new classifier is built using the same learning process after increasing the weights of the mistakenly classified tasks. The method is repeated several times. The classifiers are then combined using weighted voting (Danso et al., 2022)

3.2 Bagging

Bagging ensemble method (Breiman, 1996) starts with bootstrap sampling (i.e. random sampling with replacement) of the training dataset. A base classifier is then developed for each sample. Finally, the results of these multiple classifiers are then combined using majority voting.

3.3 Dagging

Dagging (Onan et al., 2016) technique integrates various classifiers on different samples of training dataset in order to enhance predictive accuracy. The Dagging ensemble generates several disjointed and stratified samples that insert each part of the data to a copy of the base classifier (Ting et al., 1997). Finally, all the classifiers are aggregated by a majority vote for classification.

4. PROPOSED MODEL

The objective of proposed model is to apply different ensemble learning algorithms to build efficient model that exhibit high accuracy and low false alarm rate. The proposed approach is as follows (Figure 1 depicts the proposed model) :

Step 1: Collect the dataset from UCI machine learning repository.

Step 2: Apply six types of rank based feature selection methods to find important features.

Step 3: Standardized the dataset.

Step 4: Apply ensemble learning classifiers on the dataset for prediction. 10-Fold cross-validation is used for training and testing purpose.

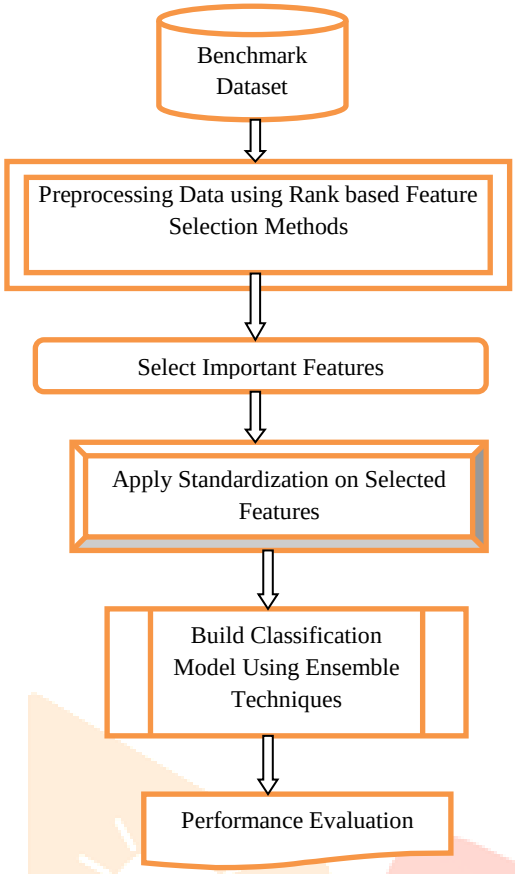

Fig. 1 Proposed Model

5. EXPERIMENTAL SETUP

5.1 Bank Marketing Dataset

The dataset is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. The objective is to promote term deposits among the customers. . The dataset publicly available in the UCI Machine Learning Repository, which can be retrieved from http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#. The dataset contains 45211 number of samples with 17 attributes without missing values.

Table No. 1: Distribution and Percentage of Instances

Class	Instances	Percentage of Class Occurrences
Yes	5289	11.7
No	39922	88.3

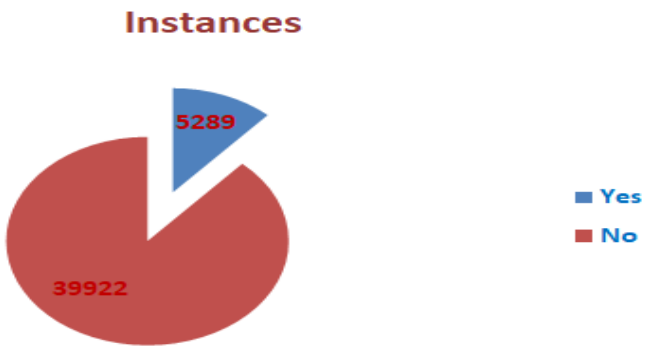

Fig. 2 Distribution of Instances

Table No.2: Attribute description of Bank Dataset

#	Attributes	Category	Attribute Description
1	Age	Numeric	The age of the customer
2	Job	Categorical	Client's occupation
3	Martial Status	Categorical	Marital status
4	Education	Categorical	The education level
5.	Default	Binary	Has credit in default
6	Balance	Numeric	Average yearly balance, in euros
7	Housing	Binary	Has housing loan
8	Loan	Binary	Has personal loan
9	Contact	Categorical	Contact communication type
10	Day	Numeric	Last contact day of the month
11	Month	Numeric	Last contact month of year
12	Duration	Numeric	Last contact duration, in seconds
13	Campaign	Numeric	Last contact day of the month
14	Pdays	Numeric	Number of days that passed by after the client was last contacted from a previous campaign
15	Previous	Numeric	Number of contacts performed before this campaign and for this client
16	Poutcome	Categorical	Outcome of the previous marketing campaign
17	Output	Categorical	Has the client subscribed a term deposit

5.2 Model Evaluation Metrics

Many measures are used to evaluate the performance of prediction algorithms. A Confusion matrix is a T x T matrix consists of four basic variables namely TP, TN, FP, , and FN used for evaluating the performance of a classification model. The matrix value compares the actual target values with those predicted by the machine learning model.

Table No.3 : Confusion Matrix

		Predicted Class	
		Negative Class	Positive Class
Actual Class	Negative Class	True Negative (TN)	False Positive (FP)
	Positive Class	False Negative (FN)	True Positive (TP)

True Positive (TP) : Observation is positive and is predicted to be positive.
False Negative (FN) : Observation is positive and is predicted to be negative.
True Negative (TN): Observation is negative and is predicted to be negative.
False Positive (FP) : Observation is negative, but is predicted positive.

Accuracy measures the probability that an algorithm can correctly predict positive and negative examples. Accuracy is calculated as:

Accuracy = (TP + TN) / (TN + TP + FN + FP) (1)

FPR is calculated using the formula:

FPR = (FP) / (TN + FP) (2)

Specificity measures the probability that an algorithm can correctly predict negative examples. It is also called true negative rate (TNR). Specificity is calculated as

Specificity = (TN) / (TN + FP) (3)

FNR = FN / (FN + TP) (4)

NPV = TN / (TN + FN) (5)

Error Rate = (FP + FN) / (TP + TN + FP + FN) (6)

AUC-ROC curve is used to measure the quality of a classification model. The better performance means larger the area.

5.3 Data Preprocessing

Data preprocessing stage improves the effectiveness of learning techniques. Data preprocessing techniques are applied before training the model. In this paper six rank based feature selection methods namely, one-R, symmetrical uncertainty (SU), gain ratio, information gain, chi-square, and relief-F are applied on dataset to select most important features among the massive irrelevant and redundant features of the given dataset. After feature selection , standardization applied on the dataset for easy understanding of the data. Standardization applied using the formula:

$$Z = (x - \mu) / \sigma \quad (7)$$

where μ is the mean of the population

σ is the standard deviation of the population

x is the raw score

6. RESULT ANALYSIS

Different combinations of three classifiers with six rank based feature selection methods were applied on the dataset. Performance measurements were made using the criteria mentioned in sec. 5.2. 10-Fold cross-validation applied on the dataset for training and testing. Table no. 4.depict the performance of techniques based on different criteria namely, accuracy, FPR, specificity, NPV.

Table No. 4 : Accuracy, FPR, Specificity, and NPV for different Ensemble learning classifiers (The values in **boldface** represent the highest value as compared to other values)

Feature Selection Method	Test Mode	Classification Techniques	Evaluation Criteria			
			Accuracy	FPR	Specificity	NPV
One-R	10-F0ld Cross-Validation	AdaBoostM1	0.8996	0.0366	0.9633	0.926
		Bagging	0.9042	0.0433	0.9567	**0.9362**
		Dagging	0.9035	**0.0155**	**0.9845**	0.913
SU	10-F0ld Cross-Validation	AdaBoostM1	0.8936	0.0208	0.9792	0.9077
		Bagging	0.9009	0.0425	0.9575	0.9321
		Dagging	0.9016	0.0178	0.9822	0.913
Gain Ratio	10-F0ld Cross-Validation	AdaBoostM1	0.8919	0.05	0.9499	0.9293
		Bagging	0.9009	0.0423	0.9577	0.932
		Dagging	0.9032	0.0183	0.9817	0.9149
Information Gain	10-F0ld Cross-Validation	AdaBoostM1	0.8955	0.0505	0.9495	0.9332
		Bagging	0.9042	0.0425	0.9575	0.9355
		Dagging	0.9038	0.0181	0.9819	0.9153
ChiSquared Evaluator	10-F0ld Cross-	AdaBoostM1	0.8942	0.051	0.9494	0.932
		Bagging	**0.9048**	0.0425	0.9575	0.9361

	Validation	Dagging	0.903	0.0163	0.9838	0.9132
Relief-F	10-F0ld Cross-Validation	AdaBoostM1	0.8953	0.049	0.9501	0.9326
		Bagging	0.9035	0.0387	0.9612	0.9316
		Dagging	0.9037	0.017	0.983	0.9144

Bagging technique with chi-squared evaluator gives highest accuracy of 0.9048. Dagging technique with one-R feature selection gives lowest false positive rate of 0.0155 and highest specificity value of 0.9845.

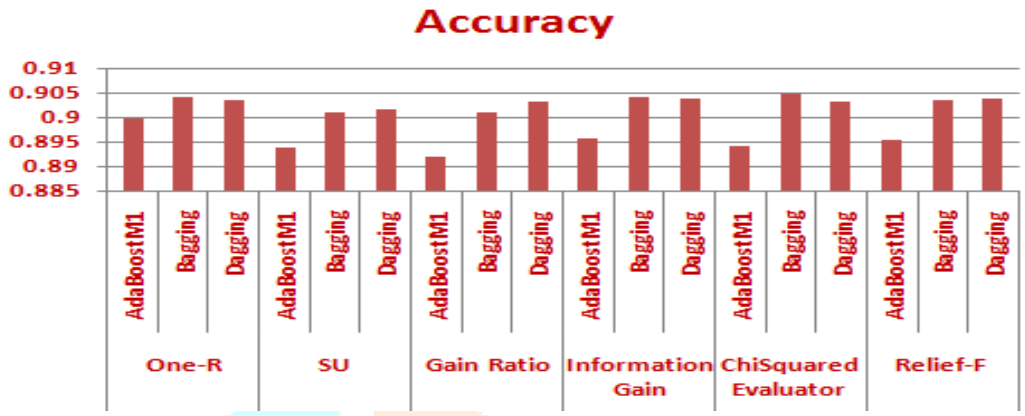

Fig. 3 Comparison of Accuracy

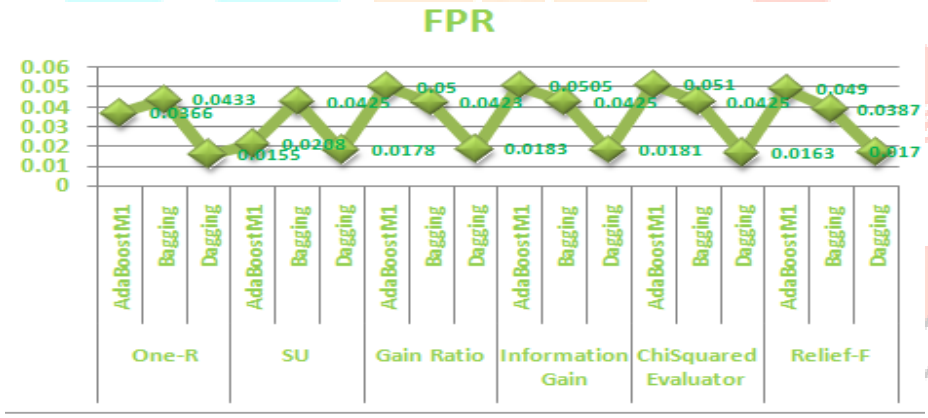

Fig. 4 Comparison of FPR

Table No. 5: FNR, Error rate, and ROC for different Ensemble learning classifiers (The values in **boldface** represent the highest value as compared to other values)

Feature Selection Method	Test Mode	Classification Techniques	Evaluation Criteria		
			FNR	Error Rate	ROC
One-R	10-F0ld Cross-Validation	AdaBoostM1	0.5814	0.1004	0.8752
		Bagging	**0.4916**	0.0957	0.9278
		Dagging	0.7077	0.0965	0.93
SU	10-F0ld Cross-Validation	AdaBoostM1	0.7519	0.1063	0.8605
		Bagging	0.53	0.0991	0.9167
		Dagging	0.7064	0.0984	0.9213
Gain Ratio	10-F0ld Cross-Validation	AdaBoostM1	0.5455	0.108	0.888
		Bagging	0.5273	0.0991	0.9161
		Dagging	0.6893	0.0968	0.9204
InfoGain	10-F0ld	AdaBoostM	0.512	0.1045	0.9

	Cross-Validation	1			
		Bagging	0.498	0.0958	0.9272
		Dagging	0.6856	0.0962	0.9292
Chisquared	10-F0ld Cross-Validation	AdaBoostM1	0.5228	0.1058	0.8977
		Bagging	0.4932	**0.0952**	**0.98**
		Dagging	0.71	0.097	0.93
ReliefF	10-F0ld Cross-Validation	AdaBoostM1	0.5184	0.1047	0.8983
		Bagging	0.5322	0.0965	0.9267
		Dagging	0.6945	0.0963	0.93

Bagging with ChiSquared evaluator gives lowest error rate of 0.0952 and highest ROC area of 0.98.

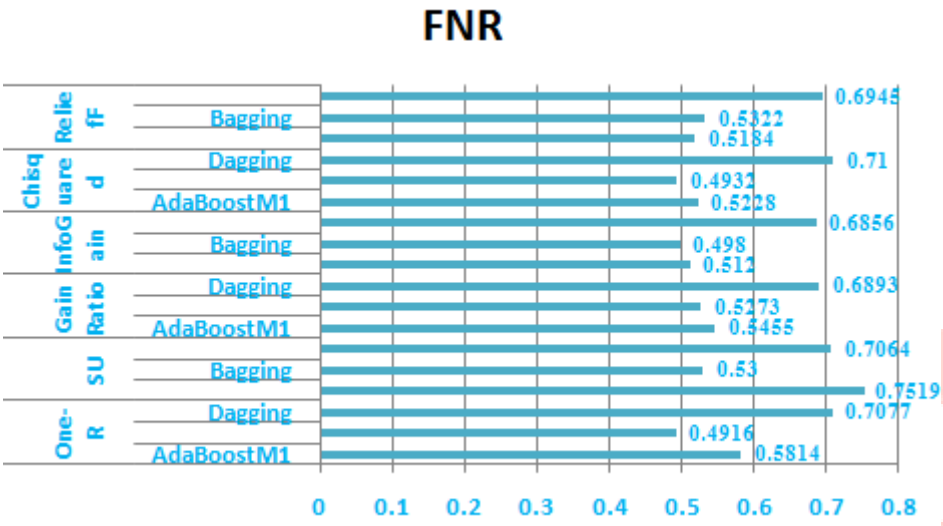

Fig. 5 Comparison of FNR

7. CONCLUSION

In this study the analysis is based on three different ensemble classifiers with six rank based feature selection methods. Bagging with ChiSquared evaluator gives lowest error rate of 0.0952 and highest ROC area of 0.98. Bagging technique with chi-squared evaluator gives highest accuracy of 0.9048. Bagging classifier gives better result as compared to other two methods.

REFERENCES

[1] Asare-Frempong, J and Jayabalan. M. 2017. Predicting Customer Response to Bank Direct Telemarketing Campaign. The International Conference on Engineering Technologies and Technopreneurship (ICE2T 2017), IEEE.

[2] Breiman, L. 1996. Bagging predictors. Machine Learning. 24(2) : 123-140.

[3] Danso, P.K. , Neto, E.C.P., Dadkhah, S., Zohourian, A., Molyneaux, H., and Ghorbani, A.A. 2022. Ensemble-based intrusion detection for internet of things devices. In Proceedings of the 2022 IEEE 19th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET), IEEE. pp. 034–039.

[4] Freund, Y., Schapire, R.E. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In Proceedings of the European Conference on Computational Learning Theory, Barcelona, Spain. pp. 23–37.

[5] Liu, M., Yan, Y., He, Y. 2017. A fuzzy support vector machine algorithm and its application in telemarketing. In Quantitative Logic and Soft Computing 2016; Springer: Cham, Switzerland. pp. 671–679.

[6] Islam, S., Arifuzzaman, M. 2019. SMOTE approach for predicting the success of bank telemarketing. In Proceedings of the 2019 4th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON), Bangkok, Thailand. pp. 1–5.
[7] Palaniappan, S., Mustapha, A., Foozy, C. F. M. , and Atan, R. 2017. Customer profiling using classification approach for bank telemarketing. International Journal on Informatics Visualization, 1(4-2) : 214–217.
[8] Onan, A., Korukoğlu, S., Bulut, H. 2016. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Systems with Applications.* *62*: 1–16.
[9] Panigrahi, A., Patnaik, M. 2020. Customer deposit prediction using neural network techniques. International Journal of Applied Engineering Research. 15(3) : 253-258.
[10] Ting, K.M.; Witten, I.H. *Stacking Bagged and Dagged Models*; University of Waikato: Hamilton, New Zealand, 1997. [11] [11] Wisaeng, K. 2013. A Comparison of Different Classification Techniques for Bank Direct Marketing. International Journal of Soft Computing and Engineering (IJSCE). 3: 116–119.
[12] Zhou, Y., Cheng, G., Jiang, S., Dai, M. 2020. Building an efficient intrusion detection system based on feature selection and ensemble classifier. Computer Networks. 174:1-21.