Performance Evaluation for Competency of Bank Telemarketing Prediction using Data Mining Techniques

Md. Rashid Farooqi, Naiyar Iqbal

Abstract*: In today's market there is cut throat competition in the banks and struggling hard to gain competitive advantage over each other. The banking industry has undergone tremendous changes in the way business conducted. They realizes the needs and techniques of data mining which is helpful tool to gather, store, capture data and convert into knowledge. The application of data mining enhances the performance of telemarketing process in banking industry. It also provide an insight how these techniques effectively used in banking industry to make the decision making process easier and productive. This work describes a data mining approach to extract valuable knowledge and information from a bank telemarketing campaign data. At this time, the potential of five data mining methods was explored for forecasting of term deposit subscription. The presentation of these techniques was evaluated on fourteen different classifier parameters. The overall better performance achieved by J48 decision tree which reported 91.2% correctly classified with sensitivity, specificity and lowest error rate of 53.8, 95.9 and 8.8 % respectively.

Keywords: Bank telemarketing, direct marketing, decision support, data mining, classification.

I. INTRODUCTION

Telemarketing is one of the powerful means of direct marketing as mobile brought revolution in the field of communication. It is really beneficial for direct marketing at much lower cost. The present era can be described as mobile information society that is why mobile became a household name [4]. The boom in mobile sale and multiple application of it has indeed affected the telemarketing services [8, 9].

Telemarketing is one of the important tools of direct marketing. The most successful telemarketing is to rely upon the quality of prospect data. By using data mining technique one may predict the expected customer that have a greater chance to use the services [7,9]. In order to understand the behaviours of customer many commercial banks using different predictive models based upon the data mining to predict or forecast the customer for segmenting the customer before offering any services by banks [6].

In case of commercial banks and financial institutions decision making becomes ever more critical. There is a large quantity of data, incomplete knowledge and availability of various alternatives increasing complexity in decision making. A good decision support system is very helpful to extract the best alternative. For organization like commercial banks, a computer based information system is very helpful and provide a quick decision.

A well designed decision support system is an interactive software based system a business intelligence system that means to helps support decision maker [10]. It complies and select useful and meaning information form a huge set of raw data, documents, knowledge or business model in order to make concrete decision.

Data mining refers to extracting valuable knowledge from huge amount of data. It is a powerful new technology with great potential to analyze important information from the data warehouse [3,5,12]. Besides, statute of these techniques has been evaluated on various classification performance indicators for forecast of term deposit subscription. For the consideration, five data mining algorithms have been applied namely as J48 decision tree, Sequential minimal optimization, Artificial neural network, Naive bayes and k-nearest neighbour.

II. OBJECTIVES

The main objective of the study is to predict the success or effectiveness of telemarketing for banking sector. A commercial Portuguese banking institution was considered. A data mining approach is proposed the competency of telemarketing for selling bank long term deposit.

Therefore many classifier methods can be analyzed with respect to performance. Target of this research additionally incorporates the correlation of various classification techniques with the assistance of charts on the basis of the dataset. All the methods have been measured by the use of WEKA data mining tool.

III. MATERIAL & TOOL

In this research work, bank marketing dataset [8] is used which is available at UCI machine learning repository [1].

Revised Manuscript Received on 30 July 2019.

* Correspondence Author

Md. Rashid Farooqi*, Department of Management Studies, Maulana Azad National Urdu University, Hyderabad, Telangana, India

Naiyar Iqbal, Department of Computer Science and Information Technology, Maulana Azad National Urdu University, Hyderabad, Telangana, India

This dataset contains 41188 records with 4640 records with term deposit subscribed (yes) and 36548 records without term deposit subscribed (no). For forecasting of term deposit subscription, 20 distinct attributes have been considered which is tabulated below (table 1).
This dataset was contributed by Portuguese banking institution. Dataset records consisted of total of 41188 records with 20 attributes for each record. Records with presence of term deposit subscription were treated as yes class and records with the absence of term deposit subscription were treated as no class for purpose of analysis. The correlation between twenty attributes of yes and no records shows the high correlation between the attributes of the two classes of samples as stated in Figures 1 and 2.
Chunk was chosen from this dataset which was dealt as training set and tested this dataset on WEKA data mining tool. ARFF is the file format of datasets that is accepted by WEKA data mining tool. WEKA is a well-known set for data mining software coded by JAVA programming, developed at the University of Waikato, New Zealand. Our fundamental concentration is on term deposit subscription that whether a client is subscribed term deposit or not by applying some attributes. On the premise of outcomes, it will demonstrate accuracy of classifiers and then compare on different performance parameters. WEKA is one of the prominent data mining tool for the classification of accuracy measurement by applying the distinctive methods.

IV. DATA MINING TECHNIQUES

Data mining is the proficient detection of previously obscure, substantial, potential, usable, distinguishable patterns in huge datasets. The investigation of observational data sets to discover unsuspected connections and to condense the data in new ways that are both distinguishable and usable [12].

Data mining phases are as follow:

- Step 1: Problem identification
- Step 2: Formulation hypothesis
- Step 3: Data collection
- Step 4: Data pre-process
- Step 5: Model estimation
- Step 6: Model Interpret and draw conjecture

V. METHODOLOGY

In this research study, five prominent data mining techniques have been used for predicting client term deposit subscription. These forecasts have been accomplished for the goal of classification and accuracy by applying diverse data mining methods. The edge used for this objective in paper is Explorer Interface. Explorer Interface is the strand applied for this purpose in research work. Accuracy can be seen by choosing the accompanying data mining techniques: DT, SMO, ANN, NB and k-NN.

Attribute Name	Data type	Remark
1. Age	Numeric	age of client
2. Job	Categorical	type of job
3. Marital	Categorical	marital status
4. Education	Categorical	has credit in default?
5. Default	Categorical	has credit in default?
6. Housing	Categorical	has housing loan?
7. Loan	Categorical	has personal loan?
8. Contact	Categorical	contact communication type
9. Month	Categorical	last contact month of year
10. Day of Week	Categorical	last contact day of the week
11. Duration	Numeric	last contact duration, (in seconds)
12. Campaign	Numeric	number of contacts performed during this campaign and for this client
13. pdays	Numeric	number of days that passed by after the client was last contacted from a previous campaign
14. Previous	Numeric	number of contacts performed before this campaign and for this client
15. Poutcome	Categorical	outcome of the previous marketing campaign
16. Emp_Var_Rate	Numeric	employment variation rate quarterly indicator
17. cons_price_idx	Numeric	consumer price index monthly indicator
18. cons_conf_idx	Numeric	consumer confidence index monthly indicator
19. euribor3m	Numeric	euribor 3 month rate
20. nr_employed	Numeric	number of employees quarterly indicator
y (desired target)	Binary	has the client subscribed a term deposit? (Yes/No)

5.1 Decision Tree

Decision tree is a tree like model that describes the data in hierarchy based by each node, and branch has a explicit coexisting with result. Selection of root of tree is appointed commonly through information gain, entropy and gain calculation [7,8].
Attribute selection measure by information gain is described as:

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

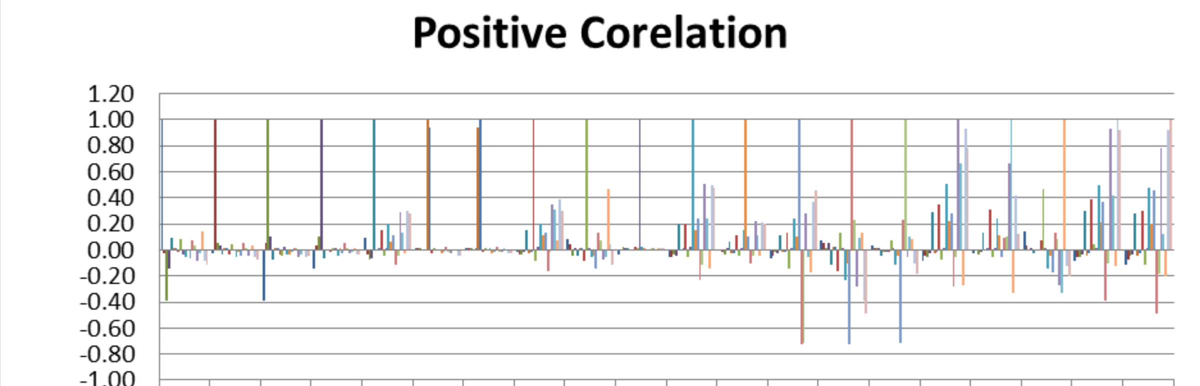
The entropy or requisite information required to classification of objects in overall sub-trees is calculated as:

$$Entropy(A) = \sum_{i=1}^{v}\frac{p_i+n_i}{p+n}I(p_i+n_i)$$

The encoded information can be achieved by branching on A:

$$Gain(A) = I(p,n) - E(A)$$

Here A is Attribute, I is Information gain, p and n are element of class P and N respectively [2].
In this experiment, the dataset is analyzed on WEKA data mining tool with J48 model and its outcomes is illuminated in the (Figure 4, Table 4).



	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
age	1.00	-0.02	-0.39	-0.14	0.09	0.00	0.00	-0.01	0.08	-0.03	-0.06	-0.01	-0.06	0.07	0.03	-0.08	-0.02	0.14	-0.09	-0.11
job	-0.02	1.00	0.06	0.04	-0.04	0.01	0.01	-0.04	0.04	0.00	-0.06	-0.01	-0.04	0.06	0.00	-0.05	0.00	0.04	-0.06	-0.07
marital	-0.39	0.06	1.00	0.10	-0.07	0.00	0.00	-0.04	-0.04	0.03	-0.04	-0.03	-0.01	0.01	0.01	-0.06	-0.04	0.00	-0.06	-0.04
education	-0.14	0.04	0.10	1.00	-0.06	0.00	0.00	-0.02	0.01	0.01	-0.05	0.01	-0.03	0.05	0.00	-0.03	-0.01	0.02	-0.03	-0.03
default	0.09	-0.04	-0.07	-0.06	1.00	0.00	0.00	0.15	-0.04	0.01	0.20	0.06	0.11	-0.12	-0.04	0.29	0.13	-0.03	0.30	0.28
housing	0.00	0.01	0.00	0.00	0.00	1.00	0.95	-0.03	0.00	-0.01	-0.01	-0.02	-0.01	0.03	-0.02	-0.03	-0.01	-0.01	-0.04	-0.04
loan	0.00	0.01	0.00	0.00	0.00	0.95	1.00	-0.01	0.01	-0.01	0.01	-0.02	-0.01	0.02	-0.02	-0.01	0.01	0.00	-0.03	-0.03
contact	-0.01	-0.04	-0.04	-0.02	0.15	-0.03	-0.01	1.00	-0.08	0.02	0.20	0.11	0.13	-0.16	-0.02	0.35	0.31	0.07	0.39	0.30
month	0.08	0.04	-0.04	0.01	-0.04	0.00	0.01	-0.08	1.00	0.02	-0.05	-0.05	-0.14	0.13	0.07	-0.08	-0.06	0.47	0.05	-0.11
day_of_week	-0.03	0.00	0.03	0.01	0.01	-0.01	-0.01	0.02	0.02	1.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
duration	-0.06	-0.06	-0.04	-0.05	0.20	-0.01	0.01	0.20	-0.05	0.03	1.00	0.16	0.24	-0.23	-0.11	0.50	0.24	-0.14	0.50	0.48
campaign	-0.01	-0.01	-0.03	0.01	0.06	-0.02	-0.02	0.11	-0.05	0.00	0.16	1.00	0.11	-0.10	-0.04	0.22	0.12	-0.04	0.21	0.20
pdays	-0.06	-0.04	-0.01	-0.03	0.11	-0.01	-0.01	0.13	-0.14	0.00	0.24	0.11	1.00	-0.73	-0.72	0.28	-0.06	-0.17	0.36	0.45
previous	0.07	0.06	0.01	0.05	-0.12	0.03	0.02	-0.16	0.13	0.00	-0.23	-0.10	-0.73	1.00	0.23	-0.28	0.09	0.13	-0.39	-0.49
poutcome	0.03	0.00	0.01	0.00	-0.04	-0.02	-0.02	-0.02	0.07	0.00	-0.11	-0.04	-0.72	0.23	1.00	-0.06	0.10	0.08	-0.11	-0.18
emp.var.rate	-0.08	-0.05	-0.06	-0.03	0.29	-0.03	-0.01	0.35	-0.08	0.00	0.50	0.22	0.28	-0.28	-0.06	1.00	0.66	-0.27	0.93	0.79
cons.price.idx	-0.02	0.00	-0.04	-0.01	0.13	-0.01	0.01	0.31	-0.06	0.00	0.24	0.12	-0.06	0.09	0.10	0.66	1.00	-0.33	0.41	0.12
cons.conf.idx	0.14	0.04	0.00	0.02	-0.03	-0.01	0.00	0.07	0.47	0.02	-0.14	-0.04	-0.17	0.13	0.08	-0.27	-0.33	1.00	-0.12	-0.20
euribor3m	-0.09	-0.06	-0.06	-0.03	0.30	-0.04	-0.03	0.39	0.05	0.00	0.50	0.21	0.36	-0.39	-0.11	0.93	0.41	-0.12	1.00	0.92
nr.employed	-0.11	-0.07	-0.04	-0.03	0.28	-0.04	-0.03	0.30	-0.11	0.00	0.48	0.20	0.45	-0.49	-0.18	0.79	0.12	-0.20	0.92	1.00

Figure 1: Linear positive correlation

5.2 Sequential Minimal Optimization

Sequential minimal optimization is a model for solving the quadratic programming problem which arises at the training time of Support Vector Machine (SVM). SMO is extensively applied for training of SVM [13]. This model is applied for dividing the data based on dataset. After processing this model the outcome of classifier can be estimated by modified estimations to make forecast for every single event of bank advertising dataset. In this experiment, the dataset is analyzed on WEKA data mining tool with SMO model [11] and its outcomes is illuminated in the (Figure 5, Table 5).

5.3 Artificial Neural Network

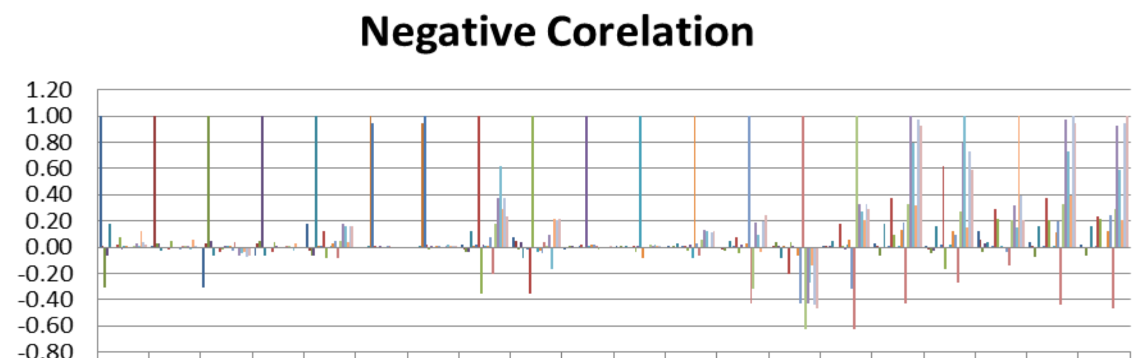
Artificial neural network (ANN) methods effort to influence the arrangement and functions of biological networks of neurons. Fundamental development of every ANN is artificial neuron, to be appropriate and a basic numerical capacity [8,13]. Summation, multiplication and then activation functions are the three basic principles of the ANN model. At the time of initial processing of artificial neuron, the inputs are weight value and which that each input value is multiply different weight. At the time of internal processing, the summation function sums every weight bias and input values. And then at the time of outcome of artificial neuron, the sum of previous weighted inputs and bias is gone via activation function [2,3].

Artificial neuron calculates the output y_k as a prescribed function of net_k value:

$$y_k = f(net_k)$$

here x and y are input and output signals respectively, w_{kj} is synaptic weight, j is synapse and f is activation function.

In this experiment, the dataset is analyzed on WEKA data mining tool with MLP (Multilayer Perceptron) model and its outcomes is illuminated in the (Figure 6, Table 6).



	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
age	1.00	0.00	-0.31	-0.06	0.18	0.00	0.00	0.02	0.07	-0.02	0.00	0.01	-0.01	0.00	0.01	0.03	0.01	0.12	0.04	0.02
job	0.00	1.00	0.03	0.03	-0.03	0.00	0.00	-0.02	0.04	-0.01	-0.01	0.00	-0.02	0.01	0.01	0.01	-0.01	0.05	0.01	0.00
marital	-0.31	0.03	1.00	0.05	-0.06	0.00	0.00	-0.04	-0.02	0.01	0.00	0.00	-0.03	0.03	-0.01	-0.06	-0.05	-0.04	-0.07	-0.07
education	-0.06	0.03	0.05	1.00	-0.07	0.00	-0.01	-0.03	0.04	0.01	-0.01	0.00	-0.01	0.01	0.00	-0.01	-0.02	0.03	-0.01	-0.01
default	0.18	-0.03	-0.06	-0.07	1.00	0.00	0.01	0.12	-0.09	0.00	0.00	0.02	0.05	-0.08	0.05	0.17	0.16	0.04	0.16	0.16
housing	0.00	0.00	0.00	0.00	0.00	1.00	0.95	0.00	0.00	0.00	-0.01	0.00	0.00	0.01	-0.01	-0.01	-0.01	0.00	-0.01	-0.01
loan	0.00	0.00	0.00	-0.01	0.01	0.95	1.00	0.02	-0.01	0.00	-0.01	0.00	0.01	0.00	0.00	0.00	0.02	0.01	0.01	0.00
contact	0.02	-0.02	-0.04	-0.03	0.12	0.00	0.02	1.00	-0.35	0.02	0.01	0.01	0.08	-0.21	0.17	0.38	0.62	0.29	0.38	0.23
month	0.07	0.04	-0.02	0.04	-0.09	0.00	-0.01	-0.35	1.00	-0.01	-0.03	-0.03	-0.05	0.03	0.01	0.10	-0.16	0.22	0.21	0.21
day_of_week	-0.02	-0.01	0.01	0.01	0.00	0.00	0.00	0.02	-0.01	1.00	0.00	0.02	0.02	0.00	-0.02	0.00	0.00	0.00	0.00	0.00
duration	0.00	-0.01	0.00	-0.01	0.00	-0.01	-0.01	0.01	-0.03	0.00	1.00	-0.08	-0.01	0.00	0.02	0.00	0.02	0.00	0.01	0.00
campaign	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.01	-0.03	0.02	-0.08	1.00	0.03	-0.07	0.06	0.13	0.12	-0.01	0.12	0.13
pdays	-0.01	-0.02	-0.03	-0.01	0.05	0.00	0.01	0.08	-0.05	0.02	-0.01	0.03	1.00	-0.43	-0.32	0.19	0.09	-0.03	0.20	0.24
previous	0.00	0.01	0.03	0.01	-0.08	0.01	0.00	-0.21	0.03	0.00	0.00	-0.07	-0.43	1.00	-0.63	-0.42	-0.27	-0.14	-0.44	-0.46
poutcome	0.01	0.01	-0.01	0.00	0.05	-0.01	0.00	0.17	0.01	-0.02	0.02	0.06	-0.32	-0.63	1.00	0.32	0.27	0.20	0.32	0.29
emp.var.rate	0.03	0.01	-0.06	-0.01	0.17	-0.01	0.00	0.38	0.10	0.00	0.00	0.13	0.19	-0.42	0.32	1.00	0.80	0.32	0.98	0.92
cons.price.idx	0.01	-0.01	-0.05	-0.02	0.16	-0.01	0.02	0.62	-0.16	0.00	0.02	0.12	0.09	-0.27	0.27	0.80	1.00	0.15	0.73	0.59
cons.conf.idx	0.12	0.05	-0.04	0.03	0.04	0.00	0.01	0.29	0.22	0.00	0.00	-0.01	-0.03	-0.14	0.20	0.32	0.15	1.00	0.39	0.21
euribor3m	0.04	0.01	-0.07	-0.01	0.16	-0.01	0.01	0.38	0.21	0.00	0.01	0.12	0.20	-0.44	0.32	0.98	0.73	0.39	1.00	0.95
nr.employed	0.02	0.00	-0.07	-0.01	0.16	-0.01	0.00	0.23	0.21	0.00	0.00	0.13	0.24	-0.46	0.29	0.92	0.59	0.21	0.95	1.00

Figure 2: Linear negative correlation

5.4 Naive Bayes

Naive Bayes performs arithmetical prediction, such as predicts class participation possibility. A basic Naive Bayes classification model guarantees analogical performance with decision tree and selected neural network classification models [2,13].

Mathematical formulae for Bayes theorem is described as:

$$P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)}$$

here X and Y shows as events, P(X) and P(Y) indicates the probabilities of X and Y without regard to each other. P(X|Y) is a conditional probability of observing event X given that Y is true. P(Y|X) is the probability of observing event Y given that X is true.

In this experiment, the dataset is analyzed on WEKA data mining tool with NB model and its outcomes is illuminated in the (Figure 7, Table 7).

5.5 k-nearest Neighbour

k-nearest Neighbour classification model is an instance learning based approach that is influenced to the lazy learning method. Instance based, also familiar as memory-based learning method that matches new problem instances with previously picked instances at training, that is stored in the memory. In k-NN classification, the result is a class membership. The object classification is decided based on majority vote of its neighbour [13].

In this experiment, the dataset is analyzed on WEKA data mining tool with kNN model and its outcomes is illuminated in the (Figure 8, Table 8).

VI. CLASSIFICATION PERFORMANCE MEASUREMENTS

Confusion matrix, alternatively known as error matrix, is a table representation that is frequently applied to depict classification model performance based on the test dataset for which the true values are known with arrangements of actual and predicted sides. There are various quality parameters such as correctly classified, incorrectly classified, specificity, sensitivity, precision, error rate, ROC and F1 measure described (table 2) for the performance measurements on the confusion matrix.

VII. RESULTS & COMPARISION

The performance measurement of subscription of term deposit prediction by five data mining algorithms is evaluated based on 20 attributes as mentioned in materials & tool section. This dataset consists of 41188 records with 4640 records with term deposit subscribed (yes) and 36548 records without term deposit subscribed (no). Bank marketing dataset records were divided in tenfold, each fold was used in testing and rest folds were applied as training throughout cross validation.

Table 1: J48_Decision tree:

		Predicted Class		Total Actual
		Yes	No	
Actual Class	Yes	2498 (62.75%)	2142 (5.76%)	4640
	No	1483 (37.25%)	35065 (94.24%)	36548
Total Predicted		3981	37207	41188

Confusion matrix of prediction result is tabulated (Table 3, Figure 3) for J48 decision tree, and other classifications like, SMO, Artificial neural network, Naive bayes and k-nearest neighbour are shown in Figure 3.

Figure 3 depicts predictions of these machine learning models. It is declared from the results that

Naive Bayes predicts topmost number of true positives (number of records predicted as yes and it does have term deposit subscribed) and SMO predicts topmost number of true negatives (number

of records predicted as no and it doesn't have term deposit subscribed) (Figure 3).

J48 Decision tree confusion matrix shows that it has second highest true positives and also it predicts second highest true negatives (Figure 3).

ANN confusion matrix shows that it has third highest true positives and third highest true negatives (Figure 3).

kNN confusion matrix shows that it has fourth highest true positives and third highest true negatives (Figure 3).

SMO confusion matrix shows worst performer in the sense of lowest true positives and Naive Bayes predicts worst performance in true negatives (Figure 3).

Table 4 to table 8 explains various classification chronicle measurements and Figure 4 to Figure 8 illuminates fourteen classification performance parameters especially correctly classified, incorrectly classified, mean absolute error, root mean squared error, relative absolute error, specificity, sensitivity and precision, FPR, NPV, RMC, ROC, PRC, and F1 measure.

Figure 4 declared that J48 decision tree outperformed over all other machine learning methods with topmost classification accuracy of 91% while second highest classification accuracy is achieved by SMO of 90% (figure 5, table 5).

On the other hand, Naive Bayes has found highest sensitivity of 61.7% (figure 7, table 7) and J48 decision tree has got second highest sensitivity of 53.8% (figure 4, table 4).

Whereas SMO acquires topmost specificity of 98% (figure 5, table 5) and J48 decision tree has got second highest sensitivity of 95.9% (figure 4, table 4).

SMO has found highest precision of 65.1% (figure 5, table 5) and J48 decision tree has got second highest precision of 62.7% (figure 4, table 4).

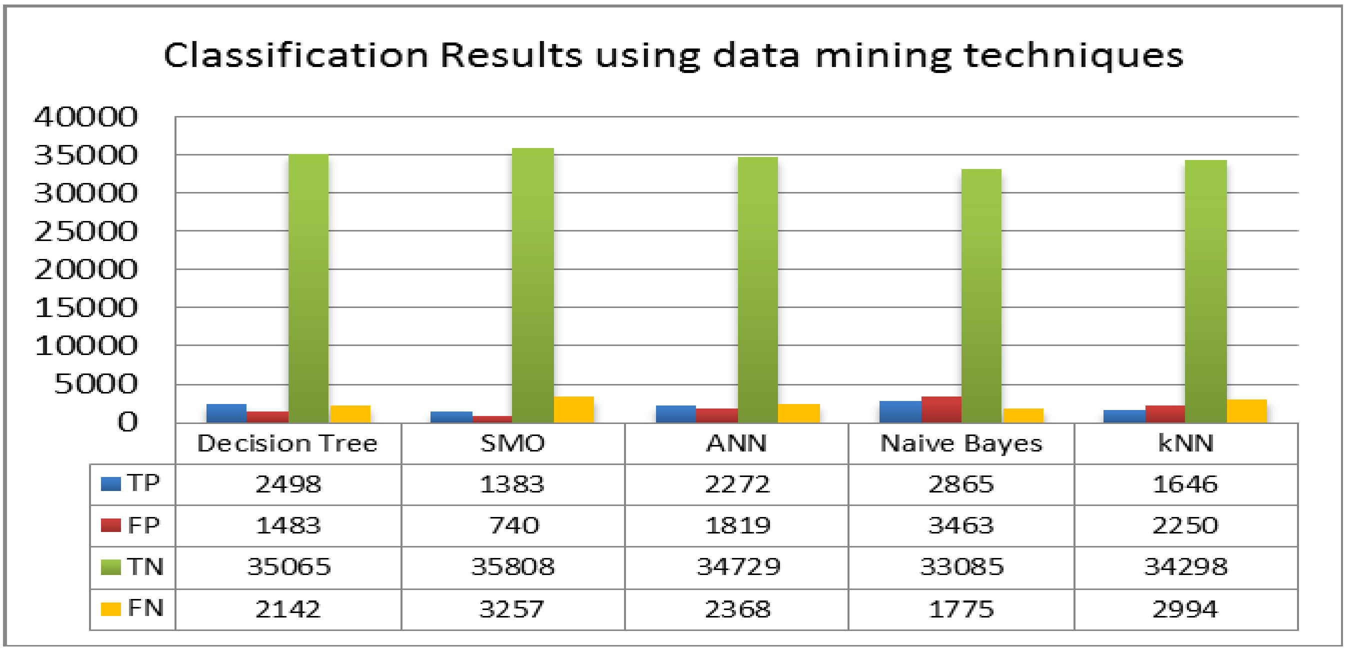
J48 has found lowest RMC of 8.8% (figure 4, table 4) and SMO has got second lowest RMC of 10.2% (figure 5, table 5).

The chart undoubtedly shows that J48 decision tree defeats all other methods on F1 score with 58% (figure 4, table 4). Whereas ANN achieves highest ROC of 89.1% (figure 6, table 6), and J48 decision tree is the second runner in ROC, that achieve 88.4% (figure 4, table 4).

Performance measures based on Confusion Matrix:

True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)
Number of client subscribed term deposit was positive and predicted right	Number of client subscribed term deposit was negative and predicted right	Number of client subscribed term deposit was negative but predicted wrong (Type I error)	Number of client subscribed term deposit was positive but predicted wrong (Type II error)

Measures	Formulae	Description
Correctly classified	$CA = (TP + TN)/(total\ sample)$	Shows percentage of correctly classified records.
Incorrectly classified	$IC = (FP + FN)/(total\ sample)$	Shows percentage of incorrectly classified records.
Mean absolute error	$\sum(\lvert f(x_i) - y_i\rvert)/N$	Calculates the mean magnitude of the errors without pondering its leaning.
Root mean squared error	$\sqrt{\sum(f(x_i) - y_i)2/N}$	Calculates the mean magnitude of the error with quadratic calculation principle
Relative absolute error	$\sum(\lvert f(x_i) - y_i\rvert)/\sum(\lvert(\bar{y}_i - y_i\rvert)$	computes the entire absolute error and normalized it through dividing by the whole absolute error
Sensitivity/Recall/TPR	$TPR = TP/(TP + FN)$	Ratio of predicted positive record to the total actually positive records.
Specificity/TNR	$TNR = TN/(TN + FP)$	Ratio of predicted negative record to the total actually negative records.
Precision/PPV	$PPV = TP/(TP + FP)$	Ratio of predicted positive record to the total predicted positive records.
FPR	$FPR = FP/(FP + TN)$	Ratio of predicted positive record to the total actually negative records.
NPV	$NPV = TN/(TN + FN)$	Ratio of predicted negative record to the total predicted negative records.
RMC	$RMC = \frac{Type\ I\ Error + Type\ II\ Err}{total\ sample}$	Ratio of overall incorrectly records to the total number of records (also known as **Error Rate**).
ROC	X-axis & Y-axis plot FPR & TPR respectively	ROC curve is a strategy for imagining, sorting out and choosing classifiers in view of their performance.
PRC	X-axis & Y-axis plot recall & precision respectively	Used as a additional of ROC to obtain the full graphic when assessing and compare to the tests.
F1 Measure	$F1 = 2TP/(2TP + FP + FN)$	It is a weighted average of the recall and precision.



	Decision Tree	SMO	ANN	Naive Bayes	kNN
TP	2498	1383	2272	2865	1646
FP	1483	740	1819	3463	2250
TN	35065	35808	34729	33085	34298
FN	2142	3257	2368	1775	2994

Figure 3: Classification output of data mining techniques.

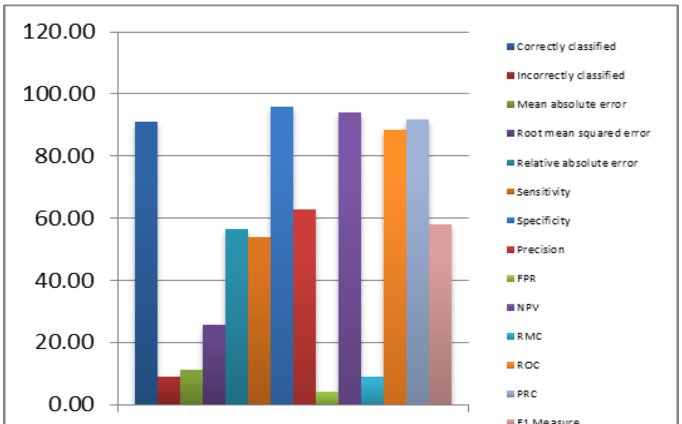


Figure 4: J48_Decision Tree

Measure Parameters	Outcome (%)
Correctly classified	91.20
Incorrectly classified	8.80
Mean absolute error	11.32
Root mean squared error	25.85
Relative absolute error	56.62
Sensitivity	53.8
Specificity	95.9
Precision	62.7
FPR	4.1
NPV	94.2
RMC	8.8
ROC	88.4
PRC	91.8
F1 Measure	58

Table 4: J48_Decision Tree

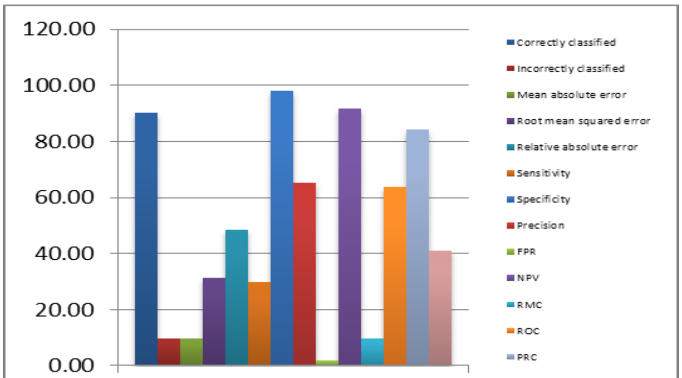


Figure 5: SMO

Measure Parameters	Outcome (%)
Correctly classified	90.30
Incorrectly classified	9.70
Mean absolute error	9.70
Root mean squared error	31.15
Relative absolute error	48.54
Sensitivity	29.8
Specificity	98
Precision	65.1
FPR	2
NPV	91.7
RMC	9.7
ROC	63.9
PRC	84.4
F1 Measure	40.9

Table 5: SMO

OPEN ACCESS
International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

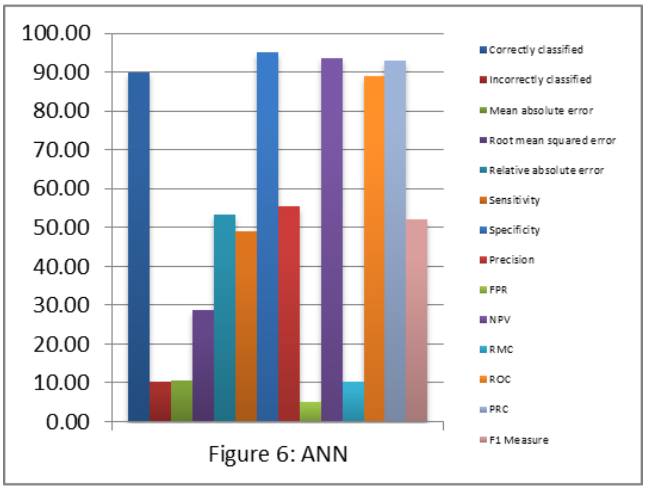


Figure 6: ANN

Measure Parameters	Outcome (%)
Correctly classified	89.83
Incorrectly classified	10.17
Mean absolute error	10.66
Root mean squared error	28.81
Relative absolute error	53.31
Sensitivity	49
Specificity	95
Precision	55.5
FPR	5
NPV	93.6
RMC	10.2
ROC	89.1
PRC	93.1
F1 Measure	52

Table 6: ANN

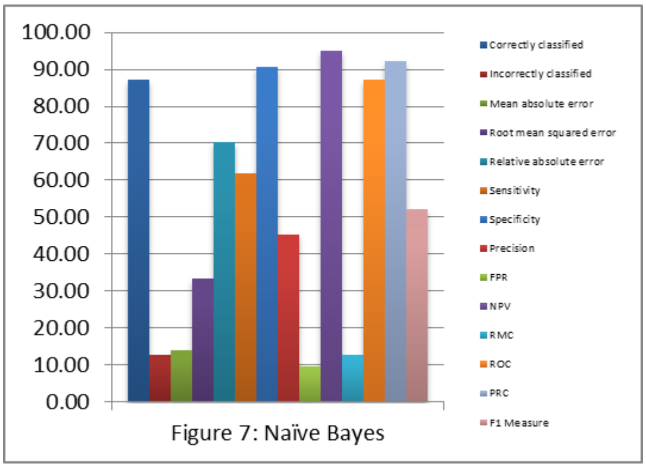


Figure 7: Naïve Bayes

Measure Parameters	Outcome (%)
Correctly classified	87.28
Incorrectly classified	12.72
Mean absolute error	14.06
Root mean squared error	33.25
Relative absolute error	70.30
Sensitivity	61.7
Specificity	90.5
Precision	45.3
FPR	9.5
NPV	94.9
RMC	12.7
ROC	87.1
PRC	92.2
F1 Measure	52.2

Table 7: Naive Bayes

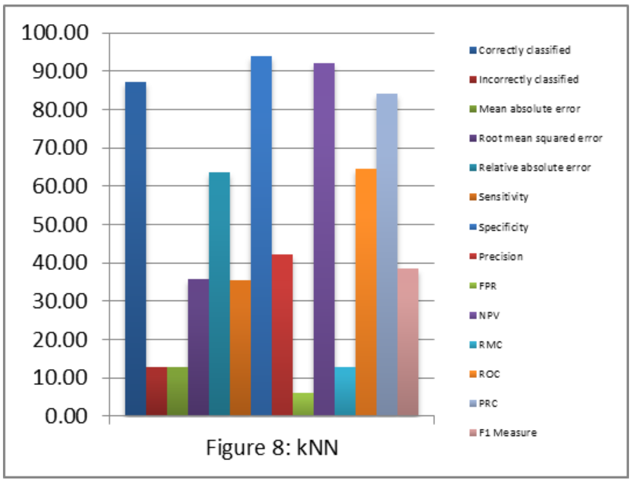


Figure 8: kNN

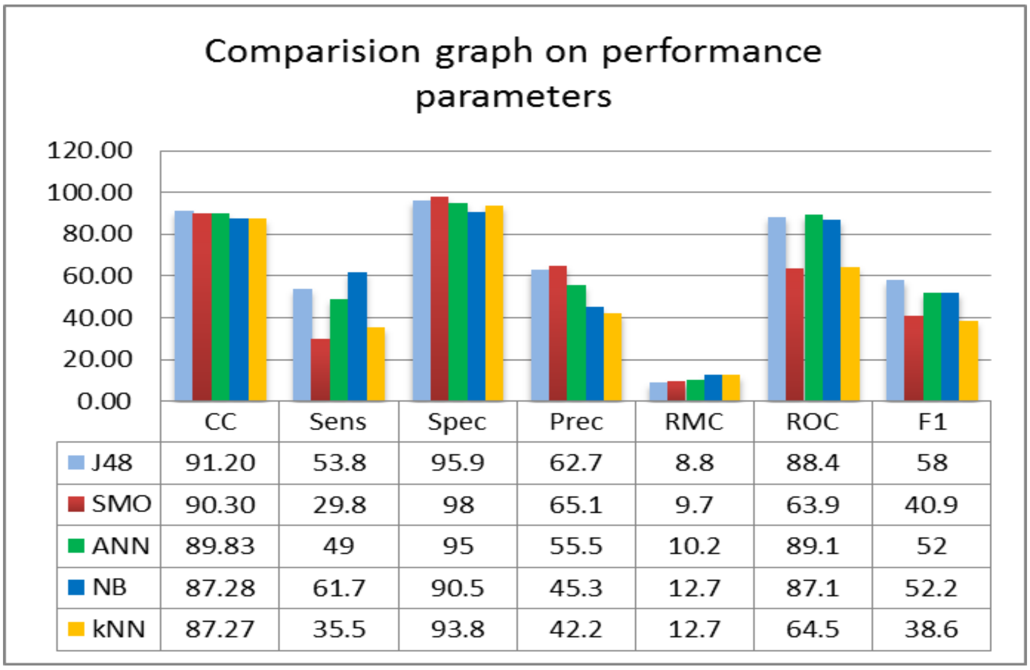
Measure Parameters	Outcome (%)
Correctly classified	87.27
Incorrectly classified	12.73
Mean absolute error	12.73
Root mean squared error	35.68
Relative absolute error	63.69
Sensitivity	35.5
Specificity	93.8
Precision	42.2
FPR	6.2
NPV	92
RMC	12.7
ROC	64.5
PRC	84
F1 Measure	38.6

Table 8: kNN

VIII. CONCLUSION

The primary goal of this research study is towards prediction of term deposit subscription using WEKA data mining tool. WEKA has mainly five sections, Explorer, Experimenter, KnowledgeFlow, Workbench and Simple GUI. Out of the five section, Explorer and KnowlegeFlow are used for classification experiment and ROC curve generation for positive and negative class (Figure 10, Figure 11). In this study, five techniques of classification were used i.e. J48 Decision tree, SMO, ANN, NB and kNN. These methods were applied using WEKA data mining software to assess the classification accuracy that was obtained after experiment of these methods. After experiment through these methods, the results were compared on the base of correctly classified and other performance measures that are described in table 2.

With the use of Explorer procedure it has concluded that J48 decision tree is the overall leading performance classifier method of data mining in this experiment. J48 decision tree has achieved an accuracy of 91.2%, takes less time to execute and depicts ROC curve 88.4%, and had minimum error rate of 8.8%.



	CC	Sens	Spec	Prec	RMC	ROC	F1
J48	91.20	53.8	95.9	62.7	8.8	88.4	58
SMO	90.30	29.8	98	65.1	9.7	63.9	40.9
ANN	89.83	49	95	55.5	10.2	89.1	52
NB	87.28	61.7	90.5	45.3	12.7	87.1	52.2
kNN	87.27	35.5	93.8	42.2	12.7	64.5	38.6

Figure 9: Comparison graph on specific performance parameters for classifier

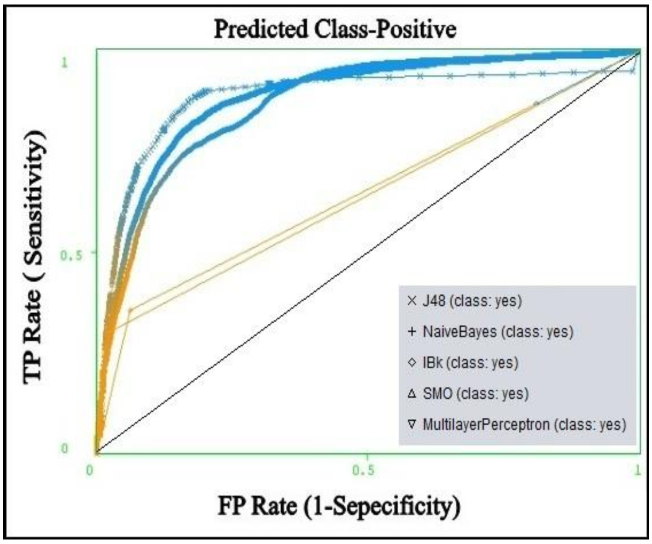


Figure 10: ROC curve for positive class

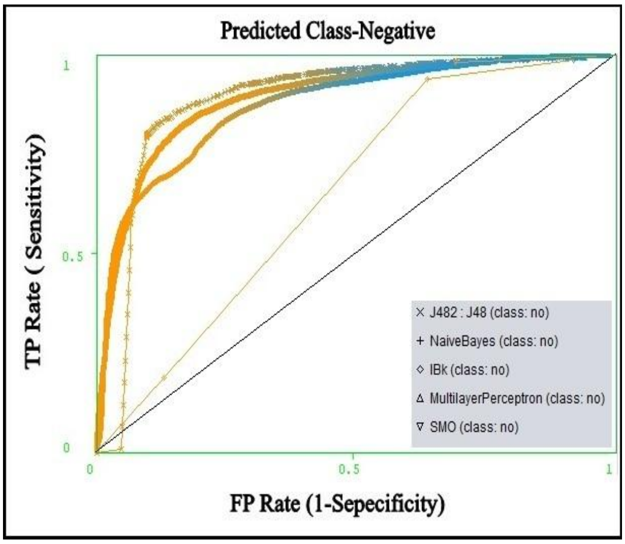


Figure 11: ROC curve for negative class

SCOPE FOR FURTHER RESEARCH

Apart from marketing data mining can be further used in to extract information form a huge set of data warehouse and enable or leads to a better decision making in the field of banking sector. It can be further use in customer acquisition and retention of customer, the most valuable customers, providing segment base product. It can be also used in fraud detection and risk management.

List of Abbreviation:

WEKA : Waikato Environment for Knowledge Analysis
ARFF : Attribute Relation File Format
DT : Decision Tree
SVM : Support Vector Machine
SMO : Sequential Minimal Optimization
ANN : Artificial Neural Network
MLP : Multilayer Percepton
NB : Naive Bayes
kNN : k- Nearest Neighbour
CC : Correctly Classified
IC : Incorrectly Classified
FPR : False Positive Rate
NPV : Negative Predictive Value
RMC : Rate of Misclassification
PRC : Precision Recall Curve
ROC : Receiver Operative Characteristics

REFERENCES

1. Bache, K., & Lichman, M. (2013). UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. University of California, School of Information and Computer Science. *Irvine, CA*.

2. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, *54*(6), 627-635.
3. Cortez, P. (2010, July). Data mining with neural networks and support vector machines using the R/rminer tool. In *Industrial Conference on Data Mining* (pp. 572-583). Springer Berlin Heidelberg.
4. Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A., & Joshi, A. (2008, March). Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th international conference on Extending database technology: Advances in database technology* (pp. 668-677). ACM.
5. Hormozi, A. M., & Giles, S. (2004). Data mining: A competitive weapon for banking and retail industries. *Information systems management*, *21*(2), 62-71.
6. Kotler, P., Keller, K. L., Ancarani, F., & Costabile, M. (2014). *Marketing management 14/e*. Pearson.
7. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124-136.
8. Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22-31.
9. Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology. In *Proceedings of European Simulation and Modelling Conference-ESM'2011* (pp. 117-121). Eurosis.
10. Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, *52*(2), 464-473.
11. Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.
12. Raza, K. (2010). Application of Data Mining in Bioinformatics. Indian Journal of Computer Science & Engineering (IJCSE), 1(2): 114-118.
13. Raza, K., & Hasan, A. N. (2015). A comprehensive evaluation of machine learning techniques for cancer class prediction based on microarray data. *International journal of bioinformatics research and applications*, *11*(5), 397-416.

AUTHORS PROFILE

Dr. Md. Rashid Farooqi is currently working as an Assistant Professor at Department of Management, School of Commerce & Business Management, Maulana Azad National Urdu University (Central University), Hyderabad, Telangana. He also shared his knowledge at reputed LN Mishra College of Business Management, Muzaffarpur (Bihar).
He has completed his B.Sc. (Hons) in Statistics from Aligarh Muslim University (AMU), Aligarh. He did Master of Business Administration (MBA) from Faculty of Management Studies and Research (FMSR), L.N.M.I. Patna with specialization in Marketing & also got Master of Business Administration (MBA) degree with specialization in Agriculture Business. He has qualified UGC-NET in Management. He has completed his Ph.D. in Management from Jamia Millia Islamia (JMI), New Delhi. He did a short term certificate course in Soft Skill from IIT, Roorkee.
His area of research is CRM, Rural Marketing, Research Methodology, Sales & Advertising Management.

Naiyar Iqbal is currently doing Doctor of Philosophy° Ph.D. (Computer Science) and he has also received his Master of Technology° M.Tech. (Computer Science) degree from Department of Computer Science and Information Technology, Maulana Azad National Urdu University (Central University), Hyderabad, Telangana. He has got his Diploma in Computer Applications° (DCA) from National Council for Promotion of Urdu Language, New Delhi. He has also completed his Bachelor of Computer Applications° (BCA) and Master of Computer Applications° (MCA) degrees from Indira Gandhi National Open University (Central University), New Delhi.
He has four years experience in teaching at Under Graduate (UG) & Post Graduate (PG) levels. His area of research is Data Mining, Machine Learning, Soft Computing, Artificial Intelligence, Bioinformatics.