PREDICTING AND CLASSIFYING BANK CUSTOMERS BY APPLYING CLASSIFICATION METHODS OF DATA MINING

B.PRIYANKA

(M.Phil (Research scholar), Tamil University, Department of Computer Science, Thanjavur, Tamil Nadu, India)

Abstract:

Data mining is a significant area for various commercial organizations comprising banking sector. It is a procedure of analysing the data from numerous viewpoint and précising it into valued information. Classification method is one of the most important techniques in data mining. In classification method is instinctively learns the possessions of classes. In this paper, banking customer dataset is used to predict and classify the accuracy of the customer's using a transaction process, using of internet banking and the customer's subscribed a term deposits. Classification techniques such as Bayes net, Naive Bayes, Logistic Regression, K-nearest neighbours and J48 is applied for the banking customer dataset. Finally, the research shows that the performance measures of customer's transaction process: J48 provides the best accuracy rate is 86.75%, internet banking of the customer's:J48 provides the best accuracy rate is 92.25% and the customer's subscribed a term deposits: Bayes net provides the best accuracy rate is 90.25%.

Keywords — **K-nearest neighbours, Logistic Regression, Naive Bayes, J48.**

I. INTRODUCTION

Data mining is a method of finding a model and extract or mining information from large amount of data. Data mining is generally mentioned to as Knowledge discovery from data (KDD) process. Data mining is described as a procedure that uses Machine Learning, Mathematical, and Artificial Intelligence and arithmetical method to take out and be familiar. The purpose of applying data mining method in banking organization is to use the obtainable data to retain its best client and to identify opportunities. Data mining can be used in banking sector for decision making and predicting. In bank organization knob the enormous quantity of difficulty to resolve. Data mining methods can be mostly used to analysis the future client's behaviour, relationship between clients' and depository, forecast of deception uncovering in acclaim cards, ATM or deduction cards, put down, loan, internet bank and dealings. Bargain hunter is a most important meaning of the reservoir manufacturing; punter segmentation is not likely to be familiar by income of which customers hold persuasive regulars and which trade include better credit. Client rapport managing (CRM)

Is a well-liked association technique focus on the civilizing of friendship flanked by company in addition to customers. Client connection managing can be described as the decision making exertions utilized to manage the commerce associates with customers by amalgamation commerce events and technology to follow to understand the commerce regulars. Customer Relationship Management and Data Mining methods are recently raising widespread consideration in banking sector. Customer Relationship Management and Data Mining techniques aids business to identify valid client's and forecast in data mining technique classification is one of the most common learning method. In classification various techniques can be used to analysis the dataset as an input and forecast attribute as a target value. Classification is the procedure of discovering a model that explain and differentiate data classes and concepts. Categorization is forms of data investigation so as to can be second-hand to passage model clarify important information lessons or to calculate future statistics trend. Classification procedure is a oversee sophistication procedure; it is skilful bottom on independence and cluster of student permit.

Supervised classification method covers a numerous data mining techniques constructed on training data. In this research data mining classification methods such as Bayes net, Naive Bayes, Logistic Regression, K-nearest neighbour and J48 algorithms are used to predict and classify the accuracy of the customer's using a transaction process, using of internet banking and the customer's subscribed a term deposits. Finally the research shows that the performance measures of customer's transaction process: J48 provides the best accuracy rate is 86.75%, internet banking of the customer's:J48 provides the best accuracy rate is 92.25% plus the customer's subscribe a word deposit: Bayes mesh provides the most excellent correctness rate is 90.25%.

II. LITERATURE SURVEY

[1] Haya Abdullah Alhakbani and Mohammad Majid al-Rifaie proposed to solve the imbalanced data using a Hybrid of Data-level and Algorithmic level solutions and oversampling the majority class, optimizing the cost parameter, gamma and kernel type of Support Vector Machines(SVM) using a grid search.

[2] Femina Bahari T proposed to predict the behavior of customers to enhance the decision making process for retaining valued customers,two classification methods like Naive Bayes and Neural Networks are used to compare the accuracy.

[3] Doha Hassan et al gossip on the subject of how information removal technique know how to be utilize to achieve end product so as to is additional precise, categorization method like Neural scheme, clutch up Vector gear, alternative plants, K-nearest neighbour, Naive Bayes and Logistic degeneration be gifted to be hand-me-down for amass telemarketing dataset from devoid of delay advertising group to determine the accurateness pace.

[4] Xujuan Zhou et al projected a fresh client acquaintance supervision(CKM) skeleton base on information drawing out, manage rapport sandwiched between bank organization plus their clientele, Neural network plus connection law are practical to forecast the behaviour of clientele plus to add to the choice creation procedure for recall appreciated clientele in bank modern.

[5] Olatunji Apampa projected to pull through the performance of cataloging algorithms hand-me-down in the stockpile payer support retort educated guess of an unnamed Portuguese compilation by yield of the not intentional wood have a meeting.

[6] Hritik Mittal et al proposed the imbalanced bank marketing dataset for direct marketing campaign and different classification algorithms used to compare the accuracy and cfs Subset Eval method is used for the dimensionality reduction b evaluating the compartment of feature base on the prognostic aptitude in addition to position out the doubling-up bordered by the ideal skin steadiness.

[7] Venkatesh Yadav et al proposed various Machine Learning Algorithms like Random Forests, Random Tree, Rep Tree, Naive Bayes,J48 and Decision Trees before and after refinement of data and highly developed arithmetical technique be practical for efficient psychoanalysis of store telemarketing in sequence in regulate to augment figure of subscribe regulars.

[8] Sergio Moro et al proposed a Data Mining techniques approach to predict the success of telemarketing calls for selling bank long-term deposits.

[9] Alaa Abu-Srhan et al planned several dream method are practical to a bank straight advertising dataset, a number of oversampling method are second-hand to improve the correctness of the forecast of a client's payment to a word deposit

[10] Shamala Palaniappan et al future to add to the accuracy of the bank shopper profile from side to side classification as healthy as identify a collection of clientele.

III.   DATASET DESCRIPTIONS AND METHODOLOGY

Dataset Description

In this investigate, the dataset hold 400 example by means of 19 quality of store clientele dataset is worn to forecast and categorize the exactness of the customer's by a treaty itinerary, through of internet bank plus the regulars. Group of students tag deal, internet bank and aim put down principles (Yes or No) so as to determine whether a client using a deal, internet bank and the client subscribe a expression put down. The characteristic person's name, portrayal, kind plus their principles are particular in the stand 1

Table - 1

No	Attribute Name	Attribute Description	Attribute Type	Attribute Values
1	Age	Age of the customer	Numeric	20 to 76
2	Job	Job type of the customer	Categorical	Admin, Homemaker, Employed, Unemployed, Student, Technician, Self-employed, Retired, Management, Blue collar.
3	Marital	Marital status of the customer	Categorical	Single, Married, Divorced
4	Education	Education level of the customer	Categorical	Primary, Secondary, Graduate, Postgraduate, Unknown, No education
5	Default	Customer has a praise in non-payment?	Binary	Yes, No, Unknown
6	Balance	Average yearly balance	Numeric	4000 to 100000
7	Loan	Does the customer have a personal or any other loan?	Binary	Yes, No
8	Contact	Contact communication type	Categorical	Mobile, Telephone
9	Day	Last contact day	Categorical	Mon, Tue, Wed, Thu, Fri
10	Month	previous get in touch with month	Categorical	Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec
11	Duration	Last contactduration in seconds	Numeric	30 to 986
12	Campaign	Number of customers communication during this campaign	Numeric	1 to 50
13	PreviousDays	Number of days passed previously the next contact day during a campaign	Numeric	1 to 871
14	Previous	Number of contacts created to the customer prior to the campaign	Numeric	0 to 25
15	Previous Outcome	Result of the previous campaign	Categorical	Nonexistent, Success, Failure
16	ATM	Does the customerhave an ATM card?	Binary	Yes, No
17	Transaction	Does the customer using atransaction process ?	Binary	Yes, No
18	Internet Banking	Does the customerusing an internet banking?	Binary	Yes, No
19	Target Deposit	Customer subscribed a termdeposit?	Binary	Yes, No

Methodology:

In this research, bank customer's dataset is used in data mining classification techniques. The proposed system has been implemented using five classification techniques like Bayes Net, Naive Bayes, Logistic Regression, K-Nearest neighbor and J48 is used to predict and classify the accuracy of the customers using of transaction process, using of internet banking and the customers subscribed a term deposits. Then the research shows that the performance measures of customer's transaction process: J48 provides the best accuracy rate is 86.75%, internet banking of the customer's:J48 provides the best accuracy rate is 92.25% and the customer's subscribed a term deposits: Bayes net provides the best accuracy rate is 90.25%.

IV. CLASSIFICATION TECHNIQUES

In this research, five categorization practices similar to Bayes net, Naive Bayes, Logistic deterioration, K-nearest neighbours plus J48 is second-hand to forecast and categorize the correctness velocity.

1. Bayes net: Bayes net is recognized as Bayesian network. Bayesian categorization is based on Bayes theorem. Bayesian classifiers are statistical classifiers. They can predict class association probability, such as the likelihood that a known tuple belong to an exacting group of students.

2. Naive Bayes: Naive Bayes is a probabilistic classifier. Naive Bayes is based on probability models that include strong individuality assumptions. Naive Bayes handles both constant and separate data, highly climbable with number of predictors and data points. For huge dataset Naive Bayes model is used as simple and easy to build.

3. Logistic Regression: Logistic Regression is a statistical analysis technique expended to forecast a data value based on previous observations of a dataset.

4. K-Nearest neighbour: K-Nearest neighbour is a data mining techniques and it is used for classification process. K-Nearest neighbour is a supervised learning algorithm; it classifies a particular instance based on the simple majority of its closest K instances. It is used to prognostic power and low computation time and it generally creates highly competitive results.

5. J48: J48 algorithm is used to classify dissimilar applications and execute accurate results of the classification.J48 is a conservatory of ID3. J48 algorithm is a straightforward C4.5 pronouncement Tree for categorization. It generate a twofold tree.J48 is the best apparatus erudition algorithm to scrutinize the data unquestionably and continuously

V. EVALUATION METRICS

In this research, the confusion matrix will be mentioned to as an initial source for evaluation. The bewilderment surrounding substance can be illustrated as exposed in the Table 2.

Table 2: Confusion Matrix

Retrieved?		
	Yes	No
Relevant? Yes	TP	FN
No	FP	TN

Where,

TP-True Positives is the positive tuples that were correctly labelled by the classifier.

TN-True Negatives is the unenthusiastic tuples that be appropriately labelled by the classifier.

FP-False Positives is the negative tuples that were incorrectly labelled by the classifier.

FN-False Negative is the encouraging tuples that be mistakenly labelled by the classifier.

Depending on the confusion matrix, a number of metrics have been defined as Accuracy, Precision, Recall and F-measure.

Accuracy: Accuracy is measured by the percentage of test set tuples that are properly classified by the classifier.

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN)$$

Precision: Precision is the ratio of all true positive in contradiction of the sum of all outcomes, both negative and positive.

$$\text{Accuracy} = (TP) / (TP+FP)$$

Recall: remember is the relation of optimistic tuples that are correctly recognized or factual positives, in disagreement of the figure of all factual positives and fake negatives.

$$\text{Recall} = (TP) / (TP+FN)$$

F-Measures: F-Measures are the harmonic mean between precision and recall.

$$\text{F-Measures} = 2(\text{Precision}*\text{Recall})/(\text{Precision} + \text{Recall})$$

VI. PERFORMANCE ANALYSIS

In this research, performance evaluation and analysis of the data mining techniques are presented

and the result of the experiments is discussed in this section.

EXPERIMENTS AND RESULTS:

In this research, Bank customer dataset is performed using weka tool. The Waikato surroundings for information examination (weka) is a mechanism knowledge toolkit worn extensively for investigate. Weka is set up by Waikato institution of higher education, New Zealand plus is unlock basis software on paper in Java (GNU public license).It hold of collection of mechanism knowledge algorithms and gear for information removal everyday jobs such as information Pre-processing, organization, cluster, connection Rule, and Regression, forecast and mental picture techniques [13].

In this research, Data mining classification techniques such as Bayes net, Naive Bayes, Logistic Regression, K-nearest neighbours and J48 algorithms are used to predict the bank customer's dataset and classify the accuracy of the customer's using a transaction process, using of internet banking and the customer's subscribed a term deposits. The result of Bayes net, Naive Bayes, Logistic Regression, K-nearest neighbours and J48 algorithms are used to predict and classify the transaction process of the customer and their performance measures like accuracy, precision, recall and F-measures are shown in the Table 3

Table 3: Performance events of customer's contract procedure

Algorithms	Percentage Accuracy (%)	Precision	Recall	F-measures
Bayes net	83.50%	0.283	0.283	0.283
Naive Bayes	82.50%	0.338	0.543	0.417
Logistic Regression	85.50%	0.342	0.283	0.310
K-nearest neighbors	81.50%	0.196	0.196	0.196
J48	86.75%	0.000	0.000	0.000

The effect of Bayes net, inexperienced Bayes, Logistic deterioration, K-nearest neighbours plus J48 algorithms are worn to envisage and organize the internet bank of the purchaser plus their presentation events like correctness, accuracy, remember plus F-measures are exposed in the bench.

Table 4: Performance measures of internet banking of the customer's

Algorithms	Percentage Accuracy (%)	Precision	Recall	F-measures
Bayes net	88.75%	0.294	0.323	0.308
Naive Bayes	84.50%	0.208	0.355	0.262
Logistic Regression	87.75%	0.179	0.161	0.169
K-nearest neighbors	89.75%	0.344	0.355	0.349
J48	92.25%	0.000	0.000	0.000

The result of Bayes net, Naive Bayes, Logistic Regression, K-nearest neighbours and J48 algorithms are used to predict and classify the customer's subscribed a term deposits and their performance measures like accuracy, precision, recall and F-measures are shown in the Table 5

Table 5: Performance measures of customer's subscribed a term deposits.

Algorithms	Accuracy	Precision	Recall	F-measures
Bayes net	90.25%	0.739	0.340	0.466
Naive Bayes	87.25%	0.486	0.360	0.414
Logistic Regression	86.50%	0.441	0.300	0.357
K-nearest neighbors	83.25%	0.293	0.240	0.264
J48	90%	0.750	0.300	0.429

in conclusion, the investigate demonstrate that the routine events of customer's deal procedure ,J48 provide the most excellent correctness speed is 86.75% .The presentation events of internet bank of the customer's,J48 provide the most excellent correctness pace is 92.25%.The presentation events of customer's subscribe a word deposit, Bayes net provide the most excellent correctness rate is 90.25%.

CONCLUSION & FUTURE WORK

In this investigate, the forecast and categorize the correctness speed of store client dataset like customer's by a deal process, by of internet bank plus the customer's subscribe a word deposits was future. Data mining techniques such as Bayes net, Naive Bayes, Logistic Regression, K-nearest neighbour and J48 algorithms are applied in the bank customer dataset. Best model that achieves high performance accuracy rate, the research shows that the performance measures of customer's transaction process, J48 provides the best accuracy rate is 86.75% .The presentation events of internet bank of the customer's, J48 provide the most excellent correctness speed is 92.25%.The performance measures of customer's subscribed a term deposits, Bayes net provides the best accuracy rate is 90.25%.

In landscape, this labour container is slow to utilize further employment of creative writing facsimile similar to cluster classifiers, euro covered with tresses classifier gain additional machine in order algorithms to get better the estimate and catalog aptitude. Too the additional great bona fide earth collection customer datasets be hypothetical to be second-hand for strenuous the opportunity new scaffold.

REFERENCES

1. Haya Addullah Alhakbani and Mohammad Majid al-Rifaie,"Handling class imbalance in direct marketing dataset using a hybrid data and algoritmic level solutions", SAI Computing Conference in 2016, IEEE.

2. Femina Bahari T and Sudheep Elayidom M,"An efficient CRM data mining framework for the prediction of customer behaviour", International Conference on Information and Communication Technologies (ICICT 2014) in 2015, ELSEVIER.

3. Doha Hassan,Ali Rodan,Maher Salem and Moayyad Mohammad,"Comparative study of using data mining techniques for bank telemarketing data",In 2019,IEEE.

4. Xujuan Zhou,Ghazal Bargshady,Moloud Abdar,Xiaohui Tao,Raj Gururajan and K C Chan,"A case study of predicting banking customers behaviour by using data mining",6th International Conference on Behavioral,Economic and Socio-Cultural Computing in 2019,IEEE.

5. Olatunji Apampa,"Evaluation of classification and Ensemble Algorithm for bank customer marketing response prediction",Journal of International Technology and Information Management in 2016,Vol-25,Issue-4,Artical

6. Hritik Mittal,Jagrit and Shubham,"Classifiction of imbalanced banking dataset using dimensionality reduction,"Proceeding of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019) in 2019,IEEE.

7. Venkatesh Yadav, M. Sreelatha, T.V. Rajinikanth, "Classification of telemarketing data using different classifier Akgorithms", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN:2278-3075,Vol-8,Issue-122.

8. Sergio Moro,Paulo Cortez and Paulo Rita,"A data-driven approach to predict the success of bank telemarketing",Decision Support System in 2014, doi:10.1016/j.dss.2014.03.001.

9. Alaa Abu-Srhan,Sanaa Al Zghoul,Bara'a Alhammad and Rizik Al- Sayyet,"Visualization and Analysis in bank direct marketing prediction",International Journal of Advanced Computer Science and Applications in 2019,Volume-10,No:7.

10. Shamala Palniappan,Aida Mustapha,Cik Feresa mohd Foozy and Rodziah Atan,"Customer profiling using classification approach for bank telemarketing",International Journal of Informatics Visualization in 2017,Volume- 1,No:4-2,e-ISSN:2549-9904,ISSN:2549-9610.

11. Daniel grzonka,Grazyna Suchacka and Barbara borowik,"Application of selected supervised classification methods to bank marketing campaign",Information System in Mangement in 2016,Volume-5(1) 36-48.