A Classification Based Model to Assess Customer Behavior in Banking Sector

Abdur Rahman
Department of Computer Science
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology
Islamabad, Pakistan
arehman780@yahoo.com

Muhammad Naeem Ahmed Khan
Department of Computer Science
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology
Islamabad, Pakistan
mnak2010@gmail.com

***Abstract*—A customer relationship management system is used to manage company relationships with current and possible customers. Following a thorough review of contemporary literature, different data mining techniques employed in different types of business, corporate sectors and organizations are analyzed. A model that would be helpful to identify customers' behavior in the banking sector is then proposed. Three classifiers, k-NN, decision tree and artificial neural networks are used to predict customer behavior and are assessed in order to determine which classifier performs better for predicting customer behavior in the banking sector.**

Keywords-customer; relationship; management; profitability; behavior; data mining; prediction

I. INTRODUCTION

Understanding customer nature and characteristics is very important for successful business growth. If we know the customer needs or customer buying patterns then we can design a market strategy to improve business and services. Company correspondence with customers is an element which affects customer loyalty. Authors in [1] proposed an efficient CRM-data mining model for customer behavior prediction which manages associations between organizations and customers. The model improves decision making process for retaining customers. The study compares two classification methods, Naïve Bayes and Neural Networks, and results show that Neural Networks performance is better. Authors in [2] evaluated classifiers performance in the banking sector. Data mining techniques are used in different domains for instance banking, CRM, medical, business strategy, weather forecasting and bioinformatics. A classification technique is used for prediction and based on accuracy of correctly classified instances, the study proposed a framework to evaluate performance of different classifiers such as logistic regression, Naïve Bayes and nearest neighbors. Authors in [3] designed an expert system for promotional marketing campaigns using fuzzy logic. Fuzzy logic is based on fuzzy sets. In fuzzy sets boundaries are not precisely defined. A fuzzy model is proposed for the selection of customers who should be targeted for deposit subscription schemes. The study selects customers based on age, annual income and loan attributes. Direct bank marketing dataset obtained from UCI repository is used for fuzzy logic model evaluation. Authors in [4] analyzed bank direct marketing using data mining techniques. The study investigates analyzed different classifiers such as multilayer perceptron neural network, nominal or logistic regression, Bayesian networks and decision tree model and applied on direct bank marketing dataset obtained from UCI repository. The classifiers performance was measured by sensitivity, specificity and accuracy. After experiments it is concluded that decision tree classifier had the higher accuracy.

Author in [5] presents a case study of data mining modeling techniques for direct marketing. CRISP-DM is a standard data mining project lifecycle and consists of six stages. The study focuses only on data preparation, modeling and evaluation stages of CRISP-DM and identifies gaps like the role of variable selection and data saturation, selection of model hyper-parameters and controlling the problem of over-fitting and under-fitting. Authors in [6] extracted actionable knowledge for direct marketing using decision tree and Naïve Bayes classifiers. The study used decision tree and Naïve Bayes classifiers to predict whether a client will subscribe a term deposit. Authors in [7] applied semi-supervised learning technique to improve CRM processes and its efficiency. It is generally believed that retaining existing customers is more profitable than attracting new ones. Customer data is rapidly growing, therefore, companies face problems in analyzing customer for retention. The study investigates the proposed algorithm on bank and insurance datasets. Authors in [8] develop a classification model to support cross selling in a mobile telecom market using fusion data mining techniques. Telecom sector is using state-of-the-art value added services (VAS) which generates more average revenue per user (ARPU). VAS is a digital service and adds extra features in the mobile phone network, for example, online games, image download, ringtones download, email, voucher and electronic transactions etc. Authors in [9] utilized a customer clustering technique to investigate customer behavior. The clustering is generally performed in services, revenue, usage and user categories attributes. Then inter-cluster analysis is performed on the generated clusters and evaluated the scattering of customers among the dissimilar group of attributes.

Authors in [10] developed an online information system and used it to store large amount of transaction data of customers that is available on the internet. These systems automatically capture customer transaction records from web browsing histories. Authors developed a new algorithm intended for produce RFM sequential patterns based on RFM notions. Authors in [11] focused on the profitability element in the banking sector. Web data and commercially accessible data are studied for predictive performance. Different data mining techniques are applied for predictive performance like decision trees, logistic regression and bagged decision trees. Authors in [12] analyzed ski resort's impact on sales and propose promotional and advertising strategies using decision tree. Ski resort uses different communication methods like public relations, advertising and sales promotion to communicate with their key market segments. The technologies and services incorporate micro blogging services, resort websites and online coupon services. Ski resorts segment customers into two major categories i.e. millennial or generation Y (less than or equal to 35) and non-millennial (greater than 35) and use promoting and advertising for them. Authors in [13] analyzed data mining based framework to identify shopping patterns. Understanding the key reasons why buyers enter their preferred stores plays an important role in achieving competitive advantage and retaining their market shares. Today, business analytics are helpful to explore a huge amount of data in order to gain customers insights and improve customer relationships. Authors propose a data mining based framework which could be used to discover patterns in customers' visits to a supermarket and identify their shopping missions. Authors in [14] analyzed customer loyalty in CRM using clustering and classification techniques. The study used a K-means algorithm for clustering purpose. This study extends the traditional RFM model by adding a weight (W) parameter thus renaming it as WRFM. The study classifies customer product loyalty for B2B concept joining WRFM with K-means and applied K-optimum for better cluster selection. Authors in [15] calculated CLV for customer segmentation and applied it in a health and beauty company. Generally, companies are concentrating on customer profitability and loyalty to improve their market share. Marketing analysis method based on RFM is used for customer segmentation. Secondly, a count item parameter is added in the RFM model. Authors in [16] proposed two methods: cluster analysis for data mining and apriori algorithm for association rules mining. Tourism plays a major role in regional and national economic development. Many companies in the world are designing product for tourism industry and this is a growing source of domestic and foreign earnings. Their proposed model is implemented in Phoenix Tours International company in Taiwan. Rules, knowledge and knowledge patterns are extracted from the dataset and they proposed solutions and suggestions for a tourism company for new product development and CRM. Authors in [17] analyzed lifestyle segmentation of customers using data mining technique. Authors in [18] developed a CRM approach using association rules and sequential patterns for a small size online shopping mall. Authors in [19] evaluated a data mining based CRM technique for enhancing profitability in terms of RFM and drew a critical review of significance of the proposed techniques. In corporate sector, particularly banking sector, the retention of customer is an important factor. We propose a classification based model to assess the customer behavior in banking sector with respect to RFM. Three classification models like decision tree, ANNs and k-NN are studied and evaluated.

In any case, certain concepts are to be further defined and elaborated:

A. Customer Life Time Value (CLV) or Customer Loyalty

Customer happiness and attraction are the core objectives of any company. Therefore, customer loyalty or customer lifetime value (CLV) is calculated in terms of recency, frequency and monetary variables (RFM).

B. RFM Analysis

Customer relationship management (CRM) system is used to manage company relations with the existing and prospect customers. RFM stands for recency, frequency and monetary. This technique is commonly used in database marketing and direct marketing to identify customer behavior or purchasing behavior. We represent customer behavior in three variables.

- Recency: How recently did the customer purchase? This is the interval between the time of most recent and current customer purchase.
- Frequency: How frequently do they purchase? This is the number of transactions in a specific period.
- Monetary (value): How much do they spend? This is the amount of money that customer spent during a specific period.

C. Decision Tress

Decision tree is a data mining and classification technique. This technique is mostly used for prediction and classification. A tree consists of paths, branches and nodes. A collection of branches is called path and represents the attribute value. Class value is represented by leaves. Each path in decision tree represents a rule which is used for classification or prediction. Decision tree divides data into subsets or nodes. Root node is the top node. Root node represents the complete dataset. Tree pruning is preformed when tree is completely built. Pruning is started from the leaf node.

D. Weka Tool

It is a machine learning toolkit extensively used for education, projects and research. It is an open source tool and developed by Waikato University, New Zealand. All the major machine learning algorithms are implemented in Weka. Data mining tasks like classification, clustering, association rules, data preparation or data preprocessing etc. can be performed using this tool.

E. ANN (Artificial Neural Network)

ANN is a data mining classification technique and it is used as a classifier. This structure consists of layered approach. Collection of neurons is called a layer. We can define multiple

or hidden layers in the ANN structure. Computation is performed in the hidden layer.

F. k-Nearest Neighbors (k-NN)

k-NN is a very simple data mining technique and is used for classification purposes. Here the symbol "k" is a static value and mostly it takes an odd value like 3, 5 and 7. Euclidean distance formula is used for measuring distances between the two entities and serves as a similarity index.

II. PROPOSED CLASSIFICATION MODEL

In the proposed model, three classifiers are employed. The purpose of using three classifiers is to determine which of them performs better for banking sector related datasets. The scheme of the proposed model is shown in Figure 1.

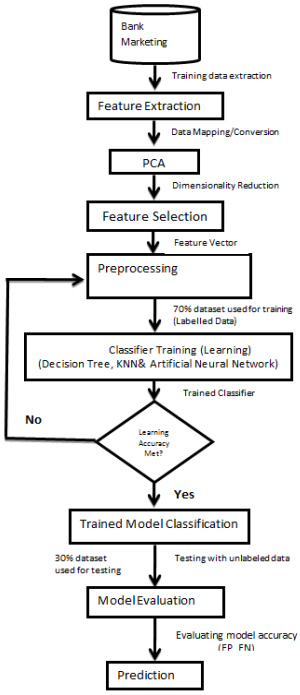

Fig. 1. Proposed classification model.

A. Dataset

Our model starts with acquiring the dataset which has been made available in the form of a CSV (comma separated values) file and is later on stored in an SQL table. The dataset is related to the direct marketing campaigns of a Portuguese banking institution between May 2008 and November 2010. The dataset details are: Name: bank-full.csv, number of instances: 45,211, Number of attributes per instance: 16 (plus one attribute indicating class label).

B. Feature Extraction

We used the bank marketing dataset available at UCI machine learning repository which consists of 45,211 instances and 17 attributes. There are 16 input variables and one output variable. The dataset attributes are: age, job, marital status, education, default (for credit information), balance, loan (housing), loan (personal), contact, day, month, duration, campaign, p-days, previous, p-outcome and class label. We analyzed data and selected 8 input attributes that provide sufficient information for classification. For preprocessing and classifier evaluation, we created mapping code based on heuristics and details are provided in Table I.

TABLE I. SELECTED ATTRIBUTES FOR CLASSIFICATION

Attribute	**Mapping Code**	
Job	**PCA, kNN & DT**	**ANN**
blue-collar	1	0100
management	2	0101
technician	3	0110
admin.	4	0111
services	5	1000
retired	6	1001
self-employed	7	1010
entrepreneur	8	1011
unemployed	9	1100
housemaid	10	1101
student	11	1110
unknown	12	1111
Attribute	**Mapping Code**	
Marital Status	**PCA, kNN & DT**	**ANN**
married	1	01
single	2	10
divorced	3	11
Attribute	**Mapping Code**	
Education	**PCA, kNN & DT**	**ANN**
secondary	1	00
tertiary	2	01
primary	3	10
unknown	4	11
Attributes	**Values**	**Mapping code PCA, kNN, DT & ANN**
Credit	Yes/No	1/0
Housing Loan	Yes/No	1/0
Personal Loan	Yes/No	1/0
Term Deposit	Yes/No	1/0

C. PCA (Based on Dataset Conversion)

The selected attributes are then passed to PCA (Principal Component Analysis) for dimensionality reduction as we have a total of 16 attributes which are surely high in dimensions to classify accurately. We employed PCA, a feature available in the WEKA tool, to obtain the pertinent set of attributes which are most suitable for classification purposes. We achieved better accuracy (99.61%) than the contemporary studies as shown in Table II. Also, we found from the experimental results that k-NN outperformed the other two classifiers DT and ANN used in our study on the basis of the accuracy measure. A comparison of true positive rates and false positive rates is provided in Table III.

D. Feature Selection

The set of attributes returned by the PCA in the form of eigenvalue constitute our “Feature Vector” to be used for training the classifier.

TABLE II. RESULTS COMPARISON

Ref	Classifier	Total	Sensitivity	Specificity	Accuracy
[2]	ANN	4521	40.88	94.85	88.63
[7]	ANN	31589	65.67	93.28	90.92
[18]	k-NN	18084	93.87	37.87	87.33
[7]	DT	31589	76.75	94.92	93.23
[18]	DT	4521	64.68	97.78	93.96
Our Model	k-NN	31648	100.00	93.24	99.61
	k-NN (k=3)	31648	99.60	13.68	94.74
	k-NN (k=5)	31648	99.90	2.79	94.40
	k-NN (k=7)	31648	99.99	0.56	94.36
	DT	31648	0.00	100.00	94.34
	ANN	31648	0.00	100.00	94.34

Sensitivity, Specificity and Accuracy are % percentages

TABLE III. TRUE POSITIVE AND FALSE POSITIVE RATES

Ref	Classifier	True Positive Rate	False Positive Rate	ROC Area
[2]	ANN	0.410	0.052	0.847
Our Model	K-NN	0.996	0.064	1
	K-NN - (k=3)	0.947	0.815	0.952
	K-NN - (k=5)	0.944	0.917	0.915
	K-NN - (k=7)	0.944	0.938	0.885
	DT	0.943	0.943	0.5
	ANN	0.943	0.943	0.567

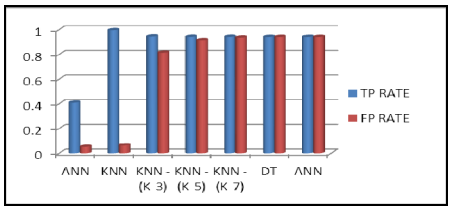

Fig. 2. True positive and false positive rates.

E. Preprocessing

The feature vector obtained in the previous step is then preprocessed to check if any noise (missing value, duplication values, and typos) are checked in the dataset. In addition, another important activity performed in the preprocessing is the encoding of the dataset in order to make it suitable for feeding the classifier for training.

F. Classifier Training (Learning)

We used three data mining classifiers: DT, k-NN and ANN. We applied the Weka tool for preprocessing and data classification. Firstly, we applied unsupervised preprocessing filter PCA in Weka on the dataset for attributes selection. For k-NN classifier, (lazy. IBk) algorithm is used to trained our model using different values of k such as k=1, k=3, k=5 and k=7. For the ANN classifier, we applied the multilayer Perceptron algorithm. For the DT classifier, we trained the model by using the J48 algorithm. The dataset is divided between 70:30 ratios as training data and test data. The Weka tool is used for training and testing (learning) model. Learning model accuracy is checked through MSE (mean squared error). If desired data accuracy is met then the trained model will be saved, otherwise the preprocessing step will be performed again. Test data is passed to the learned model to evaluate model accuracy. Classifiers' accuracy is measured through false positive and false negative values and is helpful in the model prediction.

III. EXPERIMENTATION

We used three classifiers k-NN, DT and ANN. We applied all classifiers on a bank marketing dataset that has been obtained from UCI website and Weka tool was used for model accuracy and prediction. Results are compared based on correctly and incorrectly classified instances. Sensitivity, specificity and accuracy have been calculated for model evaluation.

- Sensitivity or true positive rate (TPR) is calculated on the basis of the following formula:

$$TPR = TP/P = TP/TP+FN \quad (1)$$

- Specificity (SPC) or true negative rate (TNR) is calculated on the basis of the following formula:

$$SPC = TN/N = TN/FP+TN \quad (2)$$

- Accuracy (ACC) is calculated on the basis of the following formula:

$$ACC = TP + TN/P+N \quad (3)$$

- Precision or positive predictive value (PPV) is calculated on the basis of the following formula:

$$PPV = TP/TP+FP \quad (4)$$

A. k-NN

We performed multiple experiments using different values of "*k*" like 1,3,5,7 and their results are shown in Table IV.

B. Decision Tree (DT)

We run decision tree on the training dataset and predicted the test dataset result. The comparison of results shows that k-NN performs better than DT.

C. ANN

We performed ANN experiments on the same dataset (training dataset and test dataset) and found that kNN performs better than ANN.

TABLE IV. K-NN RESULTS

Value of K	TP Rate	FP Rate	Precision	ROC Area
1	0.996	0.064	0.996	1.000
3	0.947	0.815	0.935	0.952
5	0.944	0.917	0.927	0.915
7	0.944	0.938	0.931	0.885

TABLE V. DT RESULTS

TP Rate	FP Rate	Precision	ROC Area
0.943	0.943	0.890	0.5

TABLE VI. ANN RESULTS

TP Rate	FP Rate	Precision	ROC Area
0.943	0.943	0.890	0.567

IV. CONCLUSION AND FUTURE WORK

A model used to identify customer behavior in the banking sector using k-NN, DT and ANN is proposed in this paper. Multiple experiments are performed on bank marketing datasets in order to evaluate the model. Model specificity, sensitivity and accuracy are recorded. It is concluded that ANN outperforms DT and k-NN.

REFERENCES

[1] T. F. Bahari, M. S. Elayidom, "An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour", Procedia Computer Science, Vol. 46, pp. 725-731, 2015

[2] A. Shrivastava, B. Kumari, "Implementation of classifiers and their performance evaluation", International Journal of Engineering Research Online, Vol. 3, No. 2, pp. 71-78, 2015

[3] N. Khan, F. Khan, "Fuzzy based decision making for promotional marketing campaigns", International Journal of Fuzzy Logic Systems, Vol. 3, No. 1, pp. 64-77, 2013

[4] H. A. Elsalamony, "Bank Direct Marketing Analysis of Data Mining Techniques", International Journal of Computer Applications, Vol. 85, No. 7, pp. 12-22, 2014

[5] A. Nachev, "Application of Data Mining Techniques for Direct Marketing", in: Computational Models for Business and Engineering Domains, pp.86-95, ITHEA, Rzeszow – Sofia, 2014

[6] M. Karim, R. M. Rahman, "Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing", Journal of Software Engineering and Applications, Vol. 6, pp. 196-206, 2013

[7] S. Emtiyaz, M. Keyvanpour, "Customers behavior modeling by semi-supervised learning in customer relationship management", arXiv preprint arXiv:1201.1670, 2012

[8] H. Ahn, J. J. Ahn, K. J. Oh, D. H. Kim, "Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques", Expert Systems with Applications, Vol. 38, No. 5, pp. 5005-5012, 2011

[9] I. Bose, X. Chen, "Exploring business opportunities from mobile services data of customers: An inter-cluster analysis approach", Electronic Commerce Research and Applications, Vol. 9, No. 3, pp. 197-208, 2010

[10] Y. L. Chen, M. H. Kuo, S. Y. Wu, K. Tang, "Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data", Electronic Commerce Research and Applications, Vol. 8, No. 5, pp. 241-251, 2009

[11] J. D'Haen, D. Van den Poel, D. Thorleuchter, "Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique", Expert Systems with Applications, Vol. 40, No. 6, pp. 2007-2012, 2013

[12] P. Duchessi, E. J. Lauria, "Decision tree models for profiling ski resorts' promotional and advertising strategies and the impact on sales", Expert Systems with Applications, Vol. 40, No. 15, pp. 5822-5829, 2013

[13] A. Griva, C. Bardaki, S. Panagiotis, D. Papakiriakopoulos, "A Data Mining Based Framework to Identify Shopping Missions", Mediterranean Conference on Information Systems, August 4, 2014

[14] S. M. S. Hosseini, A. Maleki, M. R. Gholamian, "Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty", Expert Systems with Applications, Vol. 37, No. 7, pp. 5259-5264, 2010

[15] M. Khajvand, K. Zolfaghar, S. Ashoori, S. Alizadeh, "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study", Procedia Computer Science, Vol. 3, pp. 57-63, 2011

[16] S. H. Liao, Y. J. Chen, M. Y. Deng, "Mining customer knowledge for tourism new product development and customer relationship management", Expert Systems with Applications, Vol. 37, No. 6, pp. 4212-4223, 2010

[17] V. L. Migueis, A. S. Camanho, J. F. Cunha, "Customer data mining for lifestyle segmentation", Expert Systems with Applications, Vol. 39, No. 10, pp. 9359-9366, 2012

[18] B. Shim, K. Choi, Y. Suh, "CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns", Expert Systems with Applications, Vol. 39, No. 9, pp. 7736-7742, 2012

[19] A. Rahman, M. N. A. Khan,"An Assessment of Data Mining Based CRM Techniques for Enhancing Profitability", International Journal of Education and Management Engineering, Vol. 7, No. 2, pp.30-40, 2017