Customer Churn Prediction using Logistic Regression with Regularization and Optimization Technique

B. Arivazhagan, R.S. Sankara Subramanian

Abstract: ***Customer Relationship Management (CRM) is a challenging issue in marketing to better understand the customers and maintaining long-term relationships with them to increase the profitability. It plays a vital role in customer centered marketing domain which provides a better service and satisfies the customer requirements based on their characteristics in consuming patterns and smoothes the relationship where various representatives communicate and collaborate. Customer Churn prediction is one of the area in CRM that explores the transaction and communication process and analyze the customer loyalty. Data mining ease this process with classification techniques to explore pattern from large datasets. It provides a good technical support to analyze large amounts of complex customer data. This research paper applies data mining classification technique to predict churn customers in three variant sectors Banking, E-commerce and Telecom. For Classification, enhanced logistic regression with regularization and optimization technique is applied. The work is implemented in Rapid miner tool and the performance of the prediction algorithm is assessed for three variant sectors with suitable evaluation metrics.***

Keywords: ***CRM, Logistic regression, Regularization technique, Optimization algorithm.***

I. INTRODUCTION

The process of data mining involves [6] multiple steps that start with the selection of data that consists of observed values with certain attributes that may be generally historical data. The selected data are then cleaned and preprocessed. Cleaning is made in order to remove the inconsistencies and preprocessing is applied in order to get relevant information to the mining process and reduce the problem complexity. The data set is then analyzed to identify patterns and the final step is validation and visualization. These steps are performed iteratively until meaningful business knowledge is extracted. CRM consists of four dimensions namely Customer Identification, Customer Attraction, Customer Retention and Customer Development which share common goal of creating a deeper understanding with customers to maximize customer value. In this case, Data mining techniques ease the process by extracting hidden customer characteristics and behaviors from large databases. The main advantages of CRM implementation includes retaining the customer from moving to the next platform and achieving a better image of the organization in front of clients.

Revised Manuscript Received on June 30, 2020.

B. Arivazhagan, Research Scholar, Department of Computer Science, Erode Arts and Science College, Erode, E-mail: arivazhaganphd 2019 @ gmail.com

DR. R.S. Sankara Subramanian, Associate Professor, Department of Computer Science, Erode Arts and Science College, Erode, E-mail: rsankarprofessor@gmail.com

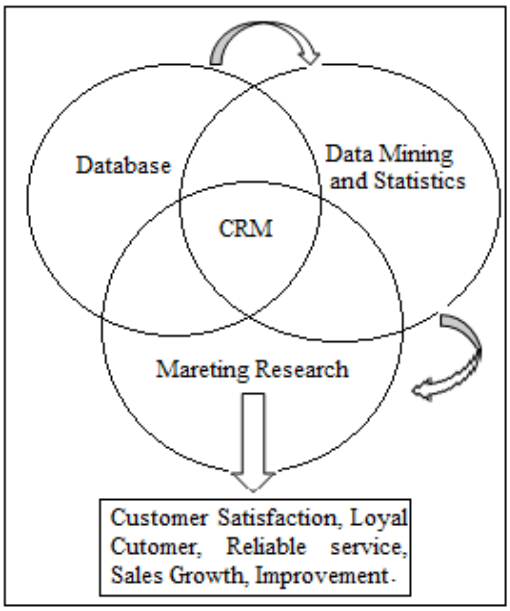

Fig 1. CRM Process

Fig. 1. Shows the CRM with the objective of the analysis using data mining techniques.

A. Types of CRM

1) Strategic - Focus on the development [11] of a customer-centric business culture.

2) Operational - Gives an overall view of the functions such as single customer view, a single page for each customer for a company. It includes client information, past sales, previous marketing efforts and summarizes all of the relationships between the customer and the company. The three main components are sales force automation, marketing automation and service automation.

3) Analytical - Analyze customer data [13] which is collected from multiple sources and to ease the business managers with more informed decisions. It uses techniques such as data mining, correlation, and pattern recognition. This analysis helps to improve customer service. For example, through the analysis of a customer buying behavior, a company might know which customer buy lot of products.

4) Collaborative - Incorporate suppliers, vendors & distributors and share customer information across the organizations.

5) Customer data platform – The marketing [14] departments that assembles data about individual people from various sources into one database, with software systems.

Data Mining Techniques for CRM

1) Classification - Collect various attributes [8] together into various categories, which are used to draw further conclusions. It contains a class label that describes the instances with the attributes. For example, the credit card holder can be classified based on the category 'Low risk, Medium risk, High risk'.

2) Prediction - Recognize and understand the historical trends to chart accurate prediction of what will happen in future. For example, the review of consumers' credit histories and past purchases will help to predict about the buying behavior of consumers in future.

3) Clustering - Groups chunks of data together based on their similarities by using distance functions. For example, the E-commerce products can be grouped based on the frequency of its moving as "Highly preferable, Moderately preferable and Least preferable".

4) Association – It specifies events or attributes that have high correlation with another event or attribute. For example, if there is a link while a customer buys a specific item (Bread); there is a chance to buy a second item (Jam).

The rest of the paper is organized as follows: Section 2 explains the variant sector used in this research work for customer churn prediction. Section 3 discusses about the related works in churn prediction, Section 4 elaborates the existing and proposed methodology, Section 5 list out the results and Section 6 concludes the analysis.

II. APPLICATION OF CRM IN VARIANT SECTORS USED IN THIS WORK

1) E-Commerce – In whole sale Electronic Commerce numerous customers [10]are registered with unique customer ID. Hence this online environment has repeated number of transactions. The challenging issue in this platform is to find out churn customer in prior they left out. Generally, the customer will move on to next platform if they are not satisfied with the products and services. The enterpriser has to retain the customer to improve their profit. So, the transaction details for each customer should be analyzed to find the customer with least transaction (permanently) and they are termed as churners.

The goals of customer churn prediction in E-Commerce are:

- Planning and organizing target customer.
- Retaining the current loyal customers.
- Improve sales marketing.
- Improve business processes and productivity.

2) Banking Sector - Banking provide multi-channel communication with customers in a consistent and efficient manner to maintain its growth. For this, banking has efficient data collection, unified view of clients, enhanced decision making. This allows banks to know their client's details and make use of customer interactions across multiple channels. This makes the banking sector to increase revenue and responds to new customers. Some customer may take variant loans and will not respond to the bank in time and some may has very small amount in savings for a long time. These customers are termed as churners.

The goals of customer churn prediction in Banking are:

- Identify the customers' patterns with non responding activities to reduce fraudulent issues.
- Able to know trust worthy customer.
- Improve the Banking service and revenue.

3) Telecom Sector - This telecom service provider offers many services in online. The online environment should have features of high streaming capacity, speed in downloading, good technical support, device protection etc. If their service is not up to the mark, the customer will move on to other service provider. To avoid this, the service provider should know the value of the customer with the customers' interested and satisfied services. The customer satisfaction with the provided service is must to retain them. So history of the customer in getting the service should be analyzed to expand the customer base.

The goals of customer churn prediction in Telecom sector are:

- Identify the loyal customer to provide discounts n services to retain them.
- Able to predict the customer who is liable to move other service providers.
- Able to know the customer satisfaction.

III. LITERATURE REVIEW

Alexiei Dingli *et al* [1] designed a customer retention model with the values recency, monetary and frequency in E-Commerce sector. Churn is defined by analyzing the customer transactions. Two types of datasets are created in which the first set, four months data are loaded and in second set, eleven month data are loaded. The customer is termed as churners if they stops the transaction in a specified period. For churn analysis, the dataset is divided into two frames with predictive and active windows. In first frame the active customers are found out by analyzing the frequency of transaction and the customer with least transaction are removed. In the second frame, if the activity customers still remain active then they are termed as non churners and the remaining are churners. The model is implemented with two classification algorithms namely Random forest and Logistic regression. Out of this, Random forest provides high accuracy and the logistic regression classifier needs some enhancement to give best accuracy. Anjali Nair *et al* [2] presents a new feature selection method proposed to resolve CRM dataset with efficient data mining technique to improve data quality and it enhances the performance of classification. This paper implements a hotel management system which includes customer registration, checking the availability of rooms, ordering food. The system check availability of rooms from database, maintain database of customers, rooms, menu, etc.

Data Mining would be applied to the large dataset and the customers is classified based on the criteria selected. This eases the organization to help improve their business. The key feature of this implemented system is to improve customer relationship by providing discount or additional privileges to the loyal customers. Data mining technique KNN is used for categorizing the customer.

Eun Whan Lee *et al* [3] discover patients' loyalty with the hospital by analyzing their medical service usage patterns. This paper proposes a data mining application in customer relationship management for hospital inpatients. It models the patterns of the loyal customers' medical services usage via a decision tree. The patients were divided into two groups according to the variables of the model. The patient who has high frequency of medical use and expenses was defined as loyal clients. The predictable factors of the loyal clients are duration of stay, selectable treatment, surgery and accompanying treatments, range of patient room, and ward from which they were discharged. Through the hospital CRM that was introduced in this work along with data-mining method clustering and decision tree classification technique, the hospitals improve their management activities.

Ibrahim M.M.Mitkees *et al* [4] proposed a model to predict customer behavior with the attributes defined in telecom sector. Sample dataset is taken from the IBM Watson analytics and missing values are preprocessed. This research predicts the customer whether they are churn or not with data mining prediction algorithms. Also the customers are grouped with clustering algorithms based on their offered services and also association rule mining is applied to get the top most patterns with confidence level one. The customer who stops transaction in the previous month is labeled as churn. Data mining classification algorithm logistic regression, MLP are implemented and in that MLP gives high accuracy, in clustering Dbscan gives high accuracy. The association rule mining is applied with one metric confidence and it produces ten rules. But single evaluation metric is discussed for classification and no evaluation metric is discussed for clustering.

Ketutgde Manik Karvana *et al* [5] compare data mining techniques for churn prediction in banking industries. The churn is termed as who stops transaction in bank and close his accounts recently. Data collected in the form of market share data in the form of interviews, documentary studies and observations. K-fold cross validation is used for training set for learning. This method works by dividing the full set into number of k and repeat the iterations to have good training sets. Other than this cross validation, three types of sampling mathods are used namely stratified sampling and two types of percentage split. These methods are evaluated with data mining classification algorithms individually. On the whole, stratified sampling gives high accuracy for all the classification algorithms.

Mohammed Hassouna *et al* [7] compare two classification techniques for customer churn prediction in telecom sector. Logistic regression in data mining technique used to predict occurrence probability and decision tree with variables' relationships in the dataset is used for analysis with two variant sets of data. The data sets are obtained from UK mobile telecommunication warehouse with 17 variables with target variable. The initial process includes replacing the missing values, discretisation of numerical values and feature selection. Then the dataset is evaluated with the classification algorithms and found decision tree outperforms the logistic regression model as it is failed in exploring the significance of relationships between the variables.

Qui Yanfang *et al* [9] put forth a method to predict churn customers in E-commerce environment by analyzing the factors such as users online duration, number of logins, attentions on the site. This research paper applies logistic regression to calculate the retention rate with sigmoid function, likelihood method. The original features in the dataset are changed to logical induction by the sigmoid function to classify the customers as churn and non-churn. The data is extracted from the real e-commerce platform and four evaluation measures are taken for analysis. The factors considered to predict churners are user's interest rate, share rate churn time and daily usage. The accuracy seems to be high but the work use already labeled set.

IV. METHODOLOGY

A. Existing method – Churn Prediction using Logistic regression (LR)

Logistic regression is a mathematically-oriented approach to analyze the effect of variables on the other variables. Prediction is made by forming a set of equations connecting independent values with the dependent field. In binary logistic model, the dependent variable has two values. The regression algorithm is evaluated to classify the customer as churn and non-churners.

Logistic regression is used to estimate the probability of certain class for binary values with sigmoid function. Logistic regression model is used for binary classification in this work for Churn & Non-churn.

$$f(y) = p(y = 1/0|x_1, \dots, x_n) \quad (1)$$

In equation (1), 'y' is the dependent variable and 'x' is the set of n independent variables. The value 'y=1 or true' implies the churn customer and 'y=0 or false'' implies the non-churn customer. Equation (2) is a sigmoid function. It is a real function having the characteristic of 'S'-shaped curve.

$$y = 1/1 + e^{-x} \quad (2)$$

Equation (2) is the sigmoid function used in logistic regression.

$$y = \beta 0 + \beta 1\, x_1 + \beta 2\, x_2 + \cdots + \beta n\, x_n \quad (3)$$

In equation (3), '$\beta 0$' is the constant value and '$\beta 1$, $\beta 2$.., β_n' are weights assigned to each independent variable from which the category of the dependent variable is to be found.

Procedure LR
Input: Datasets in three sectors
Output: Classified output as churner and non-churner

Step 1: Start with three variant sector dataset.
Step 2: Churn prediction with rule based method.
Step 3: Find logistic regression hypothesis using sigmoid function in equation (2).
Step 4: Calculated regression decision boundary using equation (3).
Step 5: Computing the regression parameter with cost function.
Step 6: Fix the target value for each instances.

Advantages

- Logistic regression performs well with linearly separable data.
- The algorithm is easy to implement, interpret and very efficient to train.

Disadvantages

- The outcome is fully depends on the set of independent variable.
- The model is affected from high bias and prone to error while having more attributes.

B. Proposed Method – Churn Prediction using Logistic Regression with Optimization technique and Regularization (LR-OR)

Logistic regression is combined with regularization and optimization technique to improve the accuracy of the learner. Initially the proposed method creates a new attribute 'Churn' filled with values true for churners and false with non-churners by using aggregate and grouping operators.

L1 Regularization technique – Regularization adds information [12] in order to solve over fitting. It improves the generalizability of the base learner and minimizes the expected error over the possible dependent and independent variables in predicting the outcome. The objective function minimize the high regression coefficients and it is calculated as,

$$\hat{\beta} = \min_{\beta} -LL\,(\beta; y, x) \qquad (4)$$

Where, LL is the log likelihood function, β is the coefficients, y is the dependent variable and x is the independent variable.

To penalize the high coefficients, regularization term (λ) is added to the objective function as below,

$$\hat{\beta} = \min_{\beta} -LL\,(\beta; y, x) + \lambda R(\beta) \qquad (5)$$

Optimization learning function- L-BFLS

It expands as "Limited memory [15] – Broyden – Fletcher-Goldfarb and Shanno" algorithm which is an iterative method for solving optimization problems. The target of the algorithm is to minimize the function f(x) with the real vector 'x'. The algorithm uses an inverse Hessian matrix. The algorithm starts with initial optimal value x_0 and iteratively estimates x_1, x_2, .. x_n. These estimates form the derivatives as,

$$g_k = \nabla f(x_k) \qquad (6)$$

Equation (6) is used to form the Hessian matrix. The matrix has the second order partial derivatives with n x n matrix.

The determinant is called Hessian determinant which is defined as

$$H(f(x)) = J(\nabla f(x)) \qquad \text{--- (7)}$$

Where, 'J' is the Jacobin matrix.

Procedure LR-OR
Input: Datasets in three sectors
Output: Classified output as churner and non-churner

Step 1: Start with three variant sector dataset.
Step 2: Churn prediction using aggregate and grouping operator.
Step 3: Find logistic regression hypothesis using sigmoid function in equation (2).
Step 4: Calculated regression decision boundary using equation (3).
Step 5: Train the model with regularization, optimization technique L-BFLS learning function using Equation (5), (6), (7).
Step 6: Fix the target value for each instances.

Advantages

- Improves the accuracy with regularization and optimization technique and reduce over fitting.
- Regularization penalizes the high coefficients and train the model to perform better by reducing the generalization error.

V. RESULT AND DISSCUSSION

A. Dataset

The three variant datasets are taken from Kaggle repository. Churn prediction is found out by using aggregate and grouping operator in all the three sectors.
1) E-Commerce dataset has 5035 instances with eight attributes. After applying aggregate and grouping operator to group the customer based on the unique transaction the dataset has 1345 instances with three attributes 'Customer id, Quantity and count (customer id)' and one attribute 'Churn' with labeled as true, false. The customers with least transaction are termed as churners.

Churn	Quantity	Cusid	count(Cusid)
false	1	1311	13
true	1	1373	1
true	1	12738	1

Fig. 2. E-Commerce dataset with churn analysis

2) Bank dataset has 3500 instances with eighteen attributes and one attribute 'Churn' with labeled as true, false. In this dataset, account balance, credit score, loan details, campaign and outcome of the campaign are considered to predict the churn customer.

Customers who have very low balance with low credit score and have low number of contacts during campaign are predicted to be churn and labeled as churners in a separate attribute.

Churn	Customer Id	marital	education	balance
false	10001	married	secondary	4500
false	10002	married	secondary	1270
false	10003	married	secondary	2476

Fig. 3.**Bank dataset with churn analysis**

3) Telecom dataset has 3910 instances with eighteen attributes and one attribute 'Churn' with labeled as true, false. The telecom service provider offers six services. The customers having less than three services with less than one year contract and pays limited total charges are predicted to be churn and labeled as churners in a separate attribute.

Churn	Customer Id	Partner	Dependents	tenure
false	12653	No	No	11
true	12654	No	No	1
false	12655	No	No	10

Fig. 4. **Telecom dataset with churn analysis**

Fig 2, 3, 4 shows three variant sectors E-Commerce, Bank and Telecom with churn prediction. This attribute is created using aggregate and grouping operator. This attribute is taken as a target or label and it is classified using proposed logistic regression LR-OR.

B. Evaluation measure

1. Accuracy – TP + TN / TP+FP+TN+FN
2. Error rate – FP + FN / TP+FP+TN+FN
3. F-Measure – 2(precision*recall) / (precision + recall)
4. Kappa Statistics– $(p_o - p_e) / (1-p_e)$
Where, p_o = observed accuracy; p_e=expected accuracy;
5. Youden (Delta P) – (Sensitivity + Specificity) -1
6. Processing Time – Total execution time to classify the data.

C. Performance Analysis for classified data

Table –I: Performance Analysis

Metrics	E-Commerce		Bank		Telecom	
	LR	LR-OR	LR	LR-OR	LR	LR-OR
Accuracy	92.51	96.72	93.02	96.12	92.35	96.92
Error rate	7.49	3.28	6.98	3.38	7.65	3.08
F-Measure	91.32	95.57	91.97	95.02	91.01	95.63
Kappa	0.829	0.912	0.845	0.902	0.821	0.927
Youden	0.832	0.915	0.851	0.905	0.825	0.931
Time	3.01	5.23	3.99	6.21	3.89	6.18

Table-I. shows the performance analysis of existing LR method and proposed LR-OR method with six evaluation measures. The Accuracy, Error rate and F-Measure are in percentage while Kappa statistics and Youden values are in rate (number), Processing time is noted in seconds. Though the proposed method takes high processing time, it gives high values for accuracy, F-measure, kappa statistics and youden with less error rate.

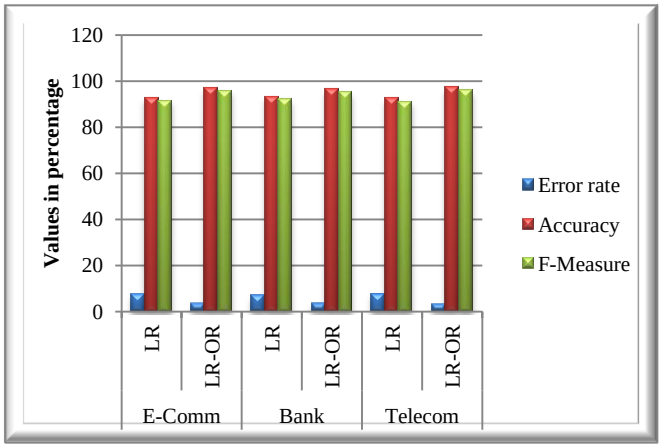

Fig. 5. Performance analysis for Error rate, accuracy, F- measure

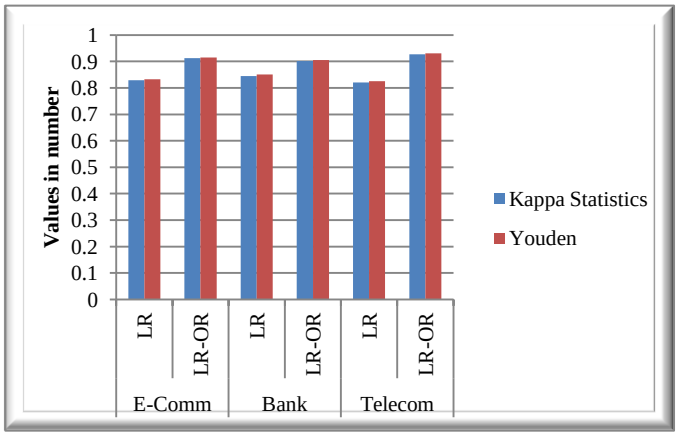

Fig. 6. Performance analysis for Kappa Statistics and Youden

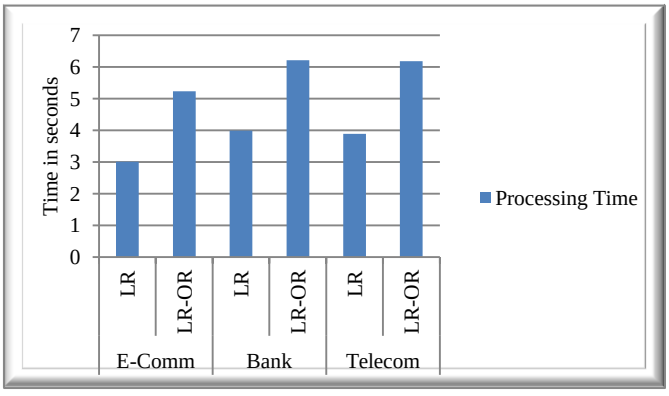

Fig. 7. Processing time analysis

Fig 5, 6 and 7 shows the performance analysis in charts. The proposed method takes high processing time but it gives high accuracy with less error rate.

VI. CONCLUSION AND FUTURE WORK

Customer Churn Prediction is an important field in Customer Relationship Management (CRM) as it analyzes the customer historical data and predicts the loyal and churn customers. This prediction should be found out to improve the profitability by increasing the customer base. This research work analyze three variant customer oriented fields and predict , classifies the churn customer with enhanced Logistic regression model with Regularization and optimization technique LR-OR. The performance of the existing logistic regression LR and proposed LR-OR method is assessed with six evaluation metrics in Rapid miner tool and the results shows the proposed method outperforms the existing method by increasing Accuracy, F-Measure, Kappa Statistics, Youden and minimize the error rate. But the LR-OR method takes high processing time for the enhancement.

In future, the work may be extended to further improve the accuracy with less processing time.

REFERENCES

1. Alexiei Dingli, Vincent Marmara, "Nicole Sant Fournier, *Enhancing Customer Retention through Data Mining Technique*, Machine Learning and Applications:" An International Journal (MLAIJ) Vol.4, No.1/2/3, September 2017.
2. Anjali Nair Rohan Khasgiwala Snigdha Mishra, " Improving Customer Relationship Management Using Data Mining", International Journal of Scientific & Engineering Research, Volume 7, Issue 2, February-2016.
3. Eun Whan Lee, " Data Mining Application in Customer Relationship Management for Hospital Inpatients", Healthcare Informatics Research, 2012.
4. Ibrahim M.M.Mitkees, Sherif M. Badr, Ahmed Ibrahim Bahgat ElSeddawy, *Customer Churn Prediction Model using Data Mining techniques*, IEEE, 2017.
5. Ketut Gde Manik Karvana, Setiadi Yazid, Amril Syalim, and Petrus Mursant, *Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry*, IEEE, 2019.
6. Michael, J.A.B. and Gordon, S.L. Data Mining Techniques: For Marketing and Sales, and Customer Relationship Management. 3rd Ed., Wiley Publishing Inc., Canada
7. Mohammed Hassouna , Ali Tarhini , Tariq Elyas & Mohammad Saeed AbouTrab, *Customer Churn in Mobile Markets: A Comparison of Techniques*, International Business Research; Vol. 8, No. 6; 2015.
8. Pavels Goncarovs, " Data Analytics in CRM Processes: A Literature Review", Information Technology and Management Science, vol. 20, December 2017.
9. Qiu Yanfang, Li Chen, *Research on E-commerce User Churn Prediction Based on Logistic Regression*, IEEE, 2017.
10. Ummugulthum Natchiar, Dr.S.Baulkani, "Customer Relationship Management Classification Using Data Mining Techniques", International Conference on Science, Engineering and Management Research (*ICSEMR)*, IEEE, 2014.
11. Zaki, Mohammed J., and Wagner Meira Jr. Data Mining and Analysis: Fundamental Concepts and Algorithms. New York: Cambridge UP, 2014.
12. www. Knime. Com /blog/ regularization for logistic logisticregression-l1-l2-gauss-or-laplace
13. www.rolustech.com/blog/data-mining-crm.
14. managementstudyguide.com/customerrelationship-management.htm.
15. en.wikipedia.org/wiki/Limited-memory_BFGS

AUTHOR PROFILE

B. Arivazhagan, is a research Scholar at Erode Arts and Science College, Erode, TamilNadu. He had two years of teaching experience in the same instituition. He has completed B.C.A and M.Sc Computer Science in PSG College of Arts and Science, Coimbatore, TamilNadu. He was awarded M.Phil (Networking) in Computer Science from Bharathiar University. His research areas are Networking and Data Mining. He presented one paper in International conference and published one paper in International journal.

Dr. R. Sankara Subramanian obtained his Master Degree in Science. He was awarded M.Phil in Computer Science from Bharathiar University. He received one Ph.D (Software testing) in Computer Science from Dravidian University and another Ph.D (Network Security) in Computer Science from Bharathiar University. He has 27 years of teaching experience in Erode Arts and Science College, Erode, TamilNadu and guided more than 56 M.Phil research scholars and 6 Ph.D Scholars. He has published 30 research papers in national, international conferences and journals. His current research area includes Software Testing, Network Security, Digital Image Processing and Data Mining.