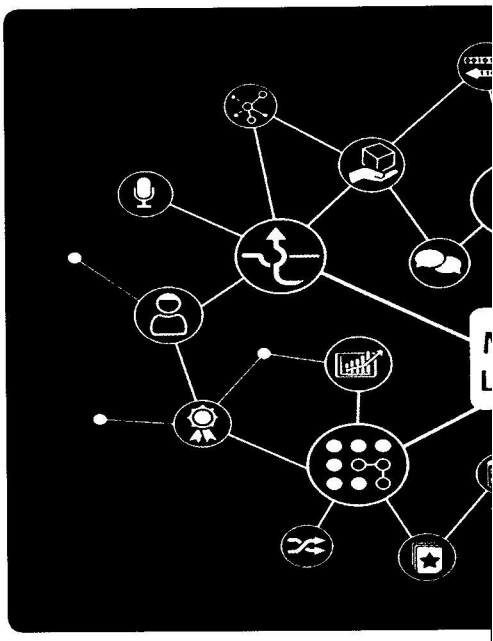
PHÂN TÍCH DỮ LIỆU NGÂN HÀNG VỚI MACHINE LEARNING

ThS. HOÀNG THỊ DUNG
Học viện Báo chí và Tuyên truyền



Ngày nay, dữ liệu lớn (Big Data) được tạo ra từ nhiều lĩnh vực trong đó có ngành Ngân hàng. Dữ liệu này chứa thông tin có giá trị cần được lưu trữ, xử lý, quản lý và phân tích... vì nó giúp tăng lợi nhuận kinh doanh. Ngành Ngân hàng đóng vai trò rất quan trọng trong nền kinh tế đất nước và khách hàng là tài sản chính của ngân hàng. Do đó, bài viết tập trung về vấn đề mà các ngân hàng đang phải đối mặt, đó là việc giữ chân khách hàng và phát hiện gian lận.

1. Đặt vấn đề

Ngành Ngân hàng tạo ra một khối lượng lớn dữ liệu mỗi ngày, bao gồm tài khoản khách hàng, lịch sử giao dịch, biến động tài chính,... Phân tích dữ liệu có thể được sử dụng để trích xuất thông tin giúp khám phá kiến thức từ Big Data. Các ngân hàng hiện nay đang phải đối mặt với nhiều thách thức như giữ chân khách hàng, phát hiện gian lận, quản lý rủi ro và các phân khúc khách hàng. Tăng lợi nhuận kinh doanh, tăng lượng khách hàng là phương pháp hiệu quả để phát triển ngân hàng. Vì vậy, điều quan trọng là xác định hành vi của người sử dụng dịch vụ, khách hàng đang hoạt động hay không hoạt động và giữ chân họ. Học máy (Machine Learning) giúp xử lý Big Data một cách nhanh chóng, tiện lợi, thông minh bằng cách phát triển các thuật toán để tạo ra thông tin chi tiết và mạng Neural nhân tạo được sử dụng để giám sát quá trình thực hiện, phân loại dữ liệu.

2. Một số nghiên cứu liên quan

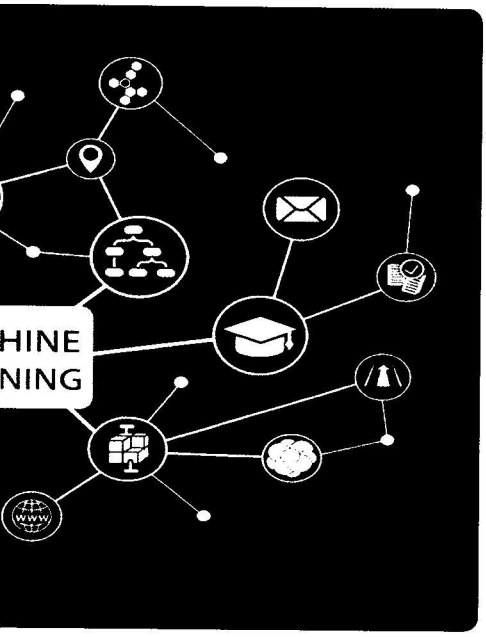
Đánh giá kết quả về phân tích dữ liệu ngân hàng, tài chính đã được nhiều nhà nghiên cứu phát triển, thực hiện bằng các phương pháp, kỹ thuật khác nhau, có thể kể đến:

- Customer Churn (tỷ lệ khách hàng rời đi) là phần trăm khách hàng ngừng sử dụng một dịch vụ hay kết thúc hợp đồng (không muốn gia hạn) với một công ty nào đó. Customer Churn thường được sử dụng trong các công ty kinh doanh với mô hình thanh toán theo kỳ (subscriber-based service model). Yong Shic và cộng sự đã thảo luận về dự đoán Customer Churn trong các ngân hàng thương mại. Họ đã sử dụng thuật toán hỗ trợ Vector Machine cho mục đích phân loại.

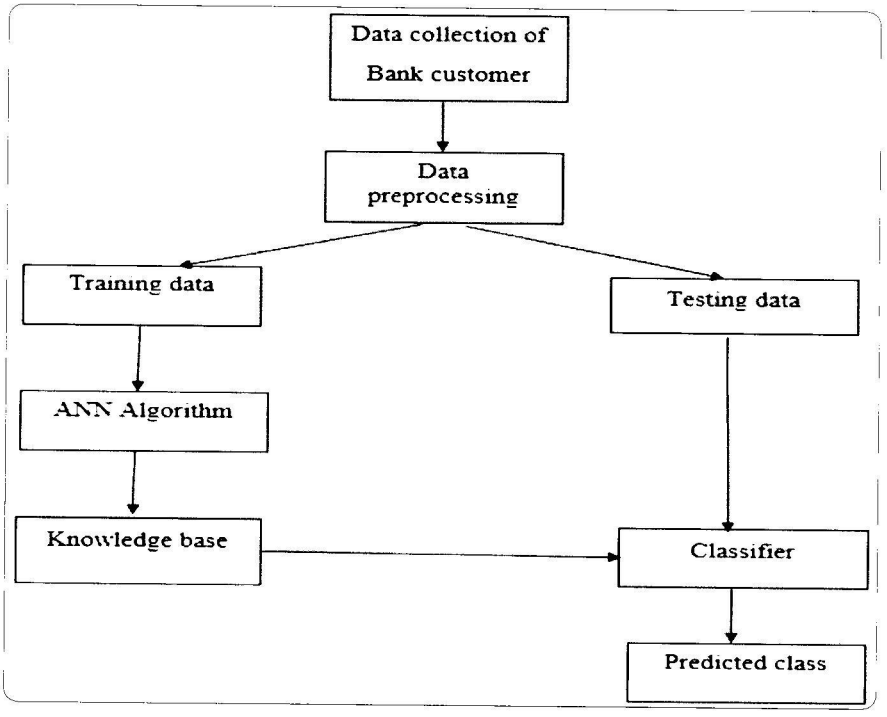
- Phương pháp lấy mẫu và mô hình hồi quy Logistic được dùng để cải thiện hiệu suất của vector hỗ trợ (Support Vector Machine - SVM). SVM là một khái niệm trong thống kê và khoa học máy tính cho một tập hợp các phương pháp học có giám sát liên quan đến nhau để phân loại và phân tích hồi quy.

Iain Brown và cộng sự so sánh các kỹ thuật khác nhau như mạng Neural, hồi quy Logistic, tăng độ dốc (gradient boosting là một thuật toán tối ưu hóa lặp bậc nhất để tìm một cực trị của một hàm khả vi), là thuật toán giám sát rừng ngẫu nhiên (random forests) và SVM để phân tích bộ dữ liệu chấm điểm tín dụng. Họ đã kiểm tra hiệu suất của kỹ thuật với sự mất cân bằng ngày càng tăng trong bộ dữ liệu theo phương pháp lấy mẫu và đưa ra kết luận rằng, tăng độ dốc hoạt động tốt hơn so với các kỹ thuật khác trong trường hợp mất cân bằng lớn.

A.B. Adeyemo và cộng sự dự đoán chu kỳ Customer Churn bằng cách sử dụng kỹ thuật khai thác dữ liệu. Họ đã sử dụng dữ liệu khách hàng từ hồ sơ do Ngân hàng Nigeria cung cấp, rồi phân tích bằng công cụ WEKA (Waikato Environment for Knowledge Analysis là một bộ phần mềm học máy được Đại học Waikato, New Zealand phát triển bằng Java, theo Giấy phép công cộng GNU); K-Means (phương pháp lượng tử hóa vector dùng để phân các điểm dữ liệu cho trước vào các cụm khác nhau) và JRip (cách thức lặp để giảm lỗi) được sử dụng làm thuật toán điều



Hình 1. Sơ đồ kiến trúc hệ thống



khiển. Kết quả thu được cho thấy các phương pháp được triển khai có thể xác định các mẫu hành vi của khách hàng.

U. Devi Prasad và cộng sự nghiên cứu mô hình người sử dụng dịch vụ của các ngân hàng ở Ấn Độ. Họ đã sử dụng các kỹ thuật khai thác CART (Classification And Regression Tree) để chuyển đổi dữ liệu khách hàng thô thành dữ liệu hữu ích.

Cheng-Tao Chu và cộng sự đã triển khai mạng Neural Multilayer Perceptron (MLP) với tính năng đào tạo lan truyền ngược để dự đoán Customer Churn của một công ty viễn thông Jordan. Để xây dựng mô hình phân loại, họ đã sử dụng các cấu trúc liên kết khác nhau của MLP, sau đó đánh giá và so sánh hai phương pháp thay đổi sai số điển hình và phương pháp dựa trên trọng số ANN (Artificial Neural Network).

Như vậy, có nhiều thuật toán Machine Learning khác nhau được sử dụng để phân loại dữ liệu, có thể tóm tắt qua bảng sau: (Bảng 1)

3. Đề xuất phương án

3.1. Kiến trúc hệ thống

Nguyên lý hoạt động của hệ thống bao gồm nhiều giai đoạn như thu thập dữ liệu, tiền xử lý, xây dựng tệp dữ liệu đào tạo và thử nghiệm để kiểm tra hiệu suất của mô hình, triển khai giải thuật toán dựa trên trọng số ANN và phân tích kết quả. Sơ đồ kiến trúc hệ thống được thể hiện trong Hình 1.

3.2. Hoạt động của ANN

Cấu trúc của ANN bao gồm ba lớp

Bảng 1. Một số thuật toán học máy được sử dụng để phân lớp và cụm dữ liệu

Thuật toán Machine Learning	Mô tả
Decision Tree (DT)	Tạo ra các quy tắc hoặc mẫu ở dạng biểu thức điều kiện if-then-else.
K-means	Chia dữ liệu thành các cụm trên cơ sở trọng tâm. Các phần tử trong cùng một cụm nằm gần tâm của cụm đó.
Naive Bayes (NB)	Sử dụng định lý Bayes để đưa ra dự đoán, suy ra xác suất của dự đoán theo các sự kiện có trong dữ liệu.
Support Vector Machine (SVM)	Sử dụng các hàm Kernel tuyến tính và phi tuyến để xử lý các dữ liệu khác nhau.
Neural Network	Các hệ thống xử lý phân tán có khả năng tìm hiểu các mẫu phức tạp xuất hiện trong dữ liệu, với tỷ lệ chính xác cao.

cơ bản: Lớp đầu vào, lớp ẩn và lớp đầu ra. (Hình 2)

Các thành phần của mạng Neural:

- *Trọng số:* Trong ANN, Neuron nhận nhiều đầu vào đồng thời. Để xử lý, hàm tính tổng của phần tử được gán cho mọi đầu vào. Trọng số là các hệ số phù hợp quyết định cường độ đầu vào cung cấp cho mạng Neural.

- *Hàm tổng hợp:* Các đầu vào và trọng số tương ứng có thể được biểu diễn dưới dạng (I1, I2,... In) và (W1, W2,... Wn). Tổng đầu vào là tích của hai vector này.

- *Chức năng kích hoạt/chuyển đổi:* Chức năng này được áp dụng trên kết quả của hàm tổng đầu vào và đầu ra.

- *Hàm bước Step:* Đầu ra của hàm này là nhị phân. Giá trị nhận được phụ thuộc vào ngưỡng chỉ định Θ.

$$f(n) = \begin{cases} 0, n < \Theta \\ 1, n \geq \Theta \end{cases} \quad \ldots\ldots\ldots (1)$$

- *Hàm Log Sigmoid:* Hàm này nhận dữ liệu đầu vào và cho ra kết quả trong phạm vi từ 0 đến 1, theo biểu thức:

$$f(n) = \frac{1}{1+e^{-n}} \quad \ldots\ldots\ldots (2)$$

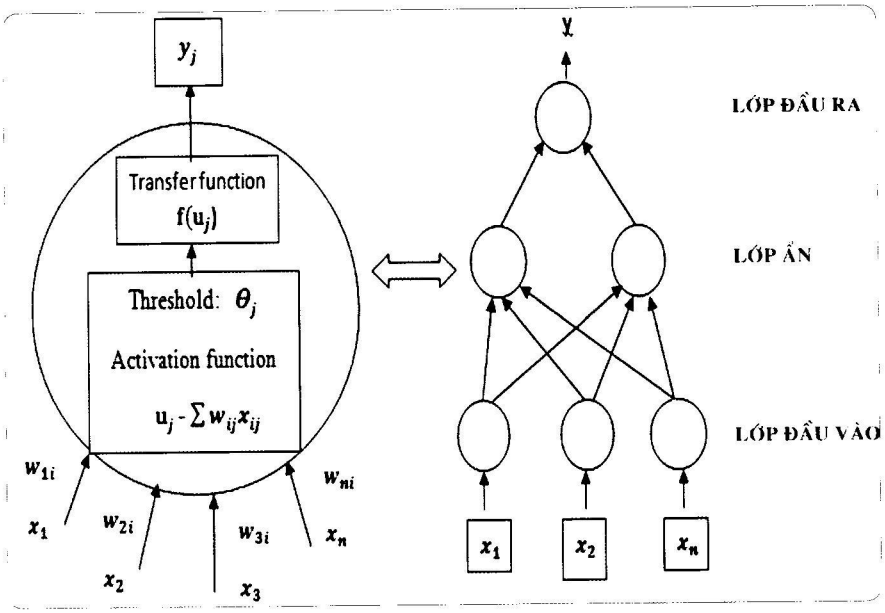
- *Tiếp tuyến Hyperbolic Sigmoid:* Hàm truyền Tan Sigmoid tạo ra đầu ra nằm trong khoảng từ -1 đến 1, theo phương trình:

$$f(n) = \frac{e^{n}-e^{-n}}{e^{n}+e^{-n}} \quad \ldots\ldots\ldots (3)$$

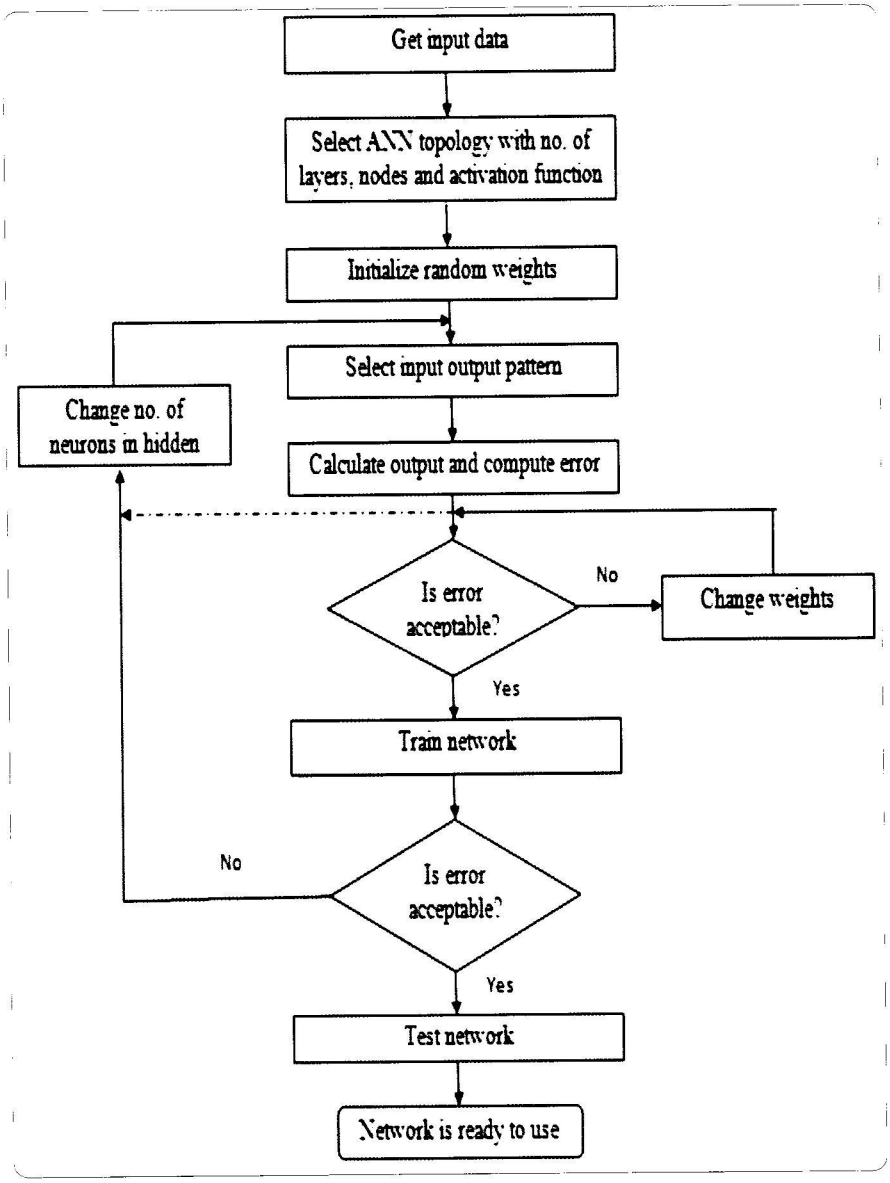
- *Chia tỷ lệ và giới hạn:* Nếu hàm chuyển đổi hoàn thành, kết quả có thể chuyển qua các quy trình bổ sung như quy mô và giới hạn.

- *Chức năng đầu ra:* Mỗi phần tử đầu vào có một đầu ra được liên kết với nó. Thông thường, đầu ra tương đương với kết quả của hàm chuyển đổi.

Hình 2: Một mạng Neural và Neuron nhân tạo



Hình 3. Thuật toán lan truyền ngược (BP)



- *Hàm sai số Error:* Sự khác biệt giữa đầu ra hiện tại và đầu ra thực tế được đánh giá là sai số và nó được chuyển đổi bởi hàm Error. Sai lệch này được lan truyền quay lại lớp trước để cập nhật trọng số.

- *Hàm đào tạo:* Được sử dụng để thay đổi trọng số trên các đầu vào của mỗi phần tử xử lý. Tỷ lệ đào tạo quyết định các điều chỉnh cần thiết để mạng hoạt động được tốt hơn.

3.3. Thuật toán lan truyền ngược (Back Propagation - BP)

Đối với mạng chuyển tiếp nguồn cấp dữ liệu, ANN sử dụng phổ biến nhất là thuật toán BP. Giải thuật toán này gồm hai giai đoạn là truyền đi và cập nhật trọng số. Trong giai đoạn đầu, tín hiệu chuyển tiếp từ lớp đầu vào đến lớp đầu ra cùng với việc thêm trọng số vào các Neuron và tính toán. Sau đó, sai số giữa giá trị thực tế và giá trị dự đoán được lan truyền ngược lại để sửa đổi trọng số. Trong giai đoạn thứ hai, trọng số ANN được điều chỉnh để giảm thiểu sai lệch. (Hình 3)

Các bước xử lý chính của thuật toán BP:

- *Bước 1:* Phân chia tín hiệu thành dữ liệu đào tạo và thử nghiệm.

- *Bước 2:* Khởi tạo tất cả các trọng số.

- *Bước 3:* Khi điều kiện dừng là Sai, thực hiện bước 4.

- *Bước 4:* Đối với mỗi đầu vào đào tạo:

+ Truyền đầu vào cho các nút trong lớp ẩn.

+ Mỗi đơn vị ẩn: Tính tổng các đầu vào có trọng số.

+ Mỗi đơn vị đầu ra: Tính tổng các đầu vào có trọng số.

+ Áp dụng tính toán đầu ra.

+ Mỗi đơn vị đầu ra nhận được một kết quả tương ứng với mô hình đào tạo đầu vào, tính toán sai số giữa thực tế và mục tiêu.

+ Truyền sai lệch về phía sau để điều chỉnh trọng số.

+ Mỗi đơn vị đầu ra cập nhật giá trị và trọng số.

+ Kiểm tra điều kiện dừng.

- *Bước 5:* Áp dụng mô hình trên dữ liệu thử nghiệm.

4. Kết quả thực nghiệm

Tệp dữ liệu 1: Được sử dụng để phát hiện gian lận. Dữ liệu này có sẵn tại kho lưu trữ học máy UCI, chứa thông tin của chủ tín dụng, tổng cộng 24 đầu vào và một đầu ra.

Tệp dữ liệu 2: Được dùng để giữ chân người sử dụng dịch vụ. Dữ liệu được chuẩn bị theo hướng dẫn của ngân hàng, chứa thông tin của khách hàng dưới dạng ID, tuổi, giới tính, tài khoản dư, thu nhập, trạng thái thẻ tín dụng, tình trạng hôn nhân, loại khoản vay, loại tài khoản, số lượng giao dịch, trình độ học vấn và công việc... của 1.000 khách hàng, chứa 12 đầu vào. Dữ liệu này có hai loại khách hàng: Hoạt động và không hoạt động. (Bảng 2)

Bộ dữ liệu đặt theo tỷ lệ 7:3, nghĩa

Bảng 2. Đặc điểm của 2 tệp dữ liệu

Dataset	Số đầu vào	Số tệp dữ liệu	Số tệp đào tạo	Số tệp thử nghiệm
Dataset1 (D1)	24	1000	700	300
Dataset2 (D2)	12	1000	700	300

Hình 4. Thử nghiệm các mẫu cho dataset1

	Actual	Predicted
897	1	1
899	0	1
903	1	1
904	1	1
910	1	0
914	0	0
921	1	1
922	0	1
940	1	1
949	0	1
954	1	0
957	1	1
966	0	1
976	1	1
979	0	0
980	0	1
988	1	0
992	1	1
993	1	0
999	1	1

Hình 5. Thử nghiệm các mẫu cho dataset2

	Actual	Predicted
106	1	1
110	1	1
116	0	0
119	1	1
126	1	0
127	0	0
173	1	1
573	1	1
576	1	1
582	0	0
583	0	1
584	1	1
587	0	0
588	0	0
848	1	1
995	1	1

là trên 1.000 hồ sơ thì 700 hồ sơ được sử dụng cho đào tạo và 300 hồ sơ còn lại cho mục tiêu và kết quả thử nghiệm. Ở đây, chức năng kích hoạt Sigmoid Logistic được sử dụng ở lớp ẩn. Ngưỡng 0,5 được sử dụng để chia dữ liệu thành các lớp, phân biệt khách hàng và nhóm tín dụng cho hai tệp dữ liệu. Hình 4 và Hình 5 cho kết quả thực tế và dự đoán của một số mẫu dữ liệu thử nghiệm từ dataset1 và dataset2.

Kết quả phân tích được thể hiện trong Bảng 3: Mô tả sai lệch trung bình (Root Mean Square Error - RMSE), độ chính xác về tệp dữ liệu đào tạo và thử nghiệm.

Bảng 3. Kết quả về dữ liệu đào tạo và kiểm tra

Dataset	Hidden nodes	RMSE		Accuracy %	
		Train	Test	Train	Test
D1	10	0.000003	0.0444	99	72
D2	7	0.000188	0.1093	99	98

Sai số cho giai đoạn đào tạo là ít cho cả hai tệp dữ liệu, chỉ có 0,28% và 0,014%, còn độ chính xác là 72%, 98% thu được từ dataset1 và dataset2. Do đó, khi sử dụng mô hình này, có thể xác định trạng thái hoạt động của khách hàng và tín dụng rất hiệu quả.■

TÀI LIỆU THAM KHẢO:

1. Utkarsh Srivastava, Santosh Gopalkrishnan, "Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks", ELSEVIER 2015.

2. Amir E. Khandani, Adlar J. Kim, Andrew W. Lo, Consumer credit-risk models via machine-learning algorithms, ELSEVIER 2010.

3. Francisca Nonyelum Ogwueleka, Department of Computer Science, Federal University of Technology, Minna "Neural Network and Classification Approach in Identifying Customer Behaviour in the Banking Sector: A Case Study of an International Bank", 2011.

4. Iain Brown, Christophe Mues," An experimental comparison of classification algorithms for imbalanced credit scoring data sets", Expert Systems with Applications 39 (2012) 3446–3453, ELSEVIER.

5. BenlanHea, Yong Shic, Qian Wan, Xi Zhao, "Prediction of customer attrition of commercial banks based on SVM Model", ELSEVIER 2014.

6. HosseinHakimpoor, Islamic Azad University, Birjand Branch, Iran "Artificial Neural Networks' Applications in Management", World Applied Sciences Journal 14 (7), 2011.

7. XinhuiTian, Rui Han, Lei Wang, Gang Lu, Jianfeng Zhan, "Latency critical big data computing in finance", ScienceDirect 2015.

8. K. Chitra, B.Subashini, "Customer Retention in Banking Sector using Predictive Data Mining Technique", ICIT 2011.

9. O. Oyeniyi A.B. Adeyemo, "Customer Churn Analysis in Banking Sector Using Data Mining Techniques", IEEE 2015.

10. Dr. K. Chitra, B. Subashini, "Data Mining Techniques and its Applications in Banking Sector", IJETAE 2013.

11. Dr. U. Devi Prasad Associate Professor Hyderabad Business School, GITAM University, Hyderabad "Prediction of Churn Behavior Of Bank Customers Using Data Mining Tools" 2012.

12. Kuchipudi Sravanthi, Tatireddy Subba Reddy, "Applications of Big data in Various Fields", IJCSIT 2015.

13. N. Sun; J. G. Morris, J. Xu, X. Zhu, "iCARE: A framework for big data-based bankingcustomer analytics", IEEE 2014.

14. XindongWu, Vipin Kumar, J. Ross Quinlan, JoydeepGhosh, Qiang Yang, "Top 10algorithms in data mining", Springer 2008.

15. Mohammad Ridwan Ismail, Besut, Terengganu, Malaysia,"A Multi-Layer Perceptron Approach for Customer Churn Prediction", International Journal of Multimedia and Ubiquitous Engineering Vol.10, No.7 (2015).

16. Cheng-Lung Huang, Mu-Chen Chen, Chieh-Jen Wang, "Credit scoring with a data mining approach based on support vector machines", ELSEVIER Expert Systems with Applications 33 (2007) 847-856.

17. Priyanka S. Patil, Nagaraj V. Dharwadkar. "Analysis of Banking Data Using Machine Learning", 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).