Pattern Analysis On Banking Dataset

Amritpal Singh, Amrita Kaur, Jasmeet Kaur, Ramandeep Singh, Shipra Raheja

Abstract: Everyday refinement and development of technology has led to an increase in the competition between the Tech companies and their going out of way to crack the system andbreak down. Thus, providing Data mining a strategically and security-wise important area for many business organizations including banking sector. It allows the analyzes of important information in the data warehouse and assists the banks to look for obscure patterns in a group and discover unknown relationship in the data.Banking systems needs to process ample amount of data on daily basis, related to customer information, their credit card details, limit and collateral details, transaction details, risk profiles, Anti Money Laundering related information, trade finance data. Thousands of decisions,based on the related data, are taken in a bank daily. This paper analyzes the banking dataset in the weka environment for the detection of interesting patterns based on its applications ofcustomer acquisition, customer retention management, and marketing and management of risk, fraudulence detections.

Keyword: Data Mining, Business, Customer acquisition, retention management, Marketing, Fraud and Risk Management

1 INTRODUCTION

Technology has open up thevarious opportunities for banking sector for the efficient deliver to the customers. A large amount of data is generated with the banking systems on everyday basis,relating to customer information, their credit card details, limit and collateral details, transaction details, risk profiles, Anti Money Laundering (AML) related information, telex messages, and trade finance data. Multiple decisions are taken by bank in a day for its beneficial impact on the customers. These decisions are taken with the considerations of credit frauds, relationship beginnings, investment decisions, Illegal financing related and AML. There is a dependency on reports that are generated and drill down tools provided by the banking systems to arrive at these critical decisions that is a manual process, involving errors and protracting procedure due to large volume of historical and transactional data involved in the banking database. Intriguing patterns and knowledge can be mined from this huge volume of data that is used for the decision making. It furnishes an outline of data mining techniques and procedures. It equips an internal view for the routine approaches used in banking areas to make the decision making process easier and productive.

2 DATA MINING

Data mining is a technique used to extract vital information from existing huge amount of data and enable better decision-making for the banking and retail arena. Data Warehouse allows combination of data using data warehousing from databases into an acceptable format so that the data can be mined to extract facts, which is then analyzed and the captured information is used throughout the organization to support decision-making and managerial intention.

- *Amritpal Singh, Amrita Kaur, Jasmeet Kaur, RamandeepSingh are currently pursuing bachelor's degree program in computer science engineering in Guru Gobind Singh Indraprastha University, India. E-mail: 08jasmeet.kaur@gmail.com*
- *Shipra Raheja is currently pursuing phDin computer science engineering from SVU, India. E-mail: shipraraheja@gmail.com*

"Thederivation of implied, formerly unknown, andperhaps useful information from data" [9]is well suited procedure of Data Mining, which isfurther described as "the science of extracting useful information from large databases or data warehouse" [6]. Knowledge discovery possess in its primary tasks ofData mining [7].

3 KDD PROCESS

Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a data warehouse. The steps included in KDD process begins with the tasks of data preparation and selection, data cleansing, incorporating precedent knowledge on data sets and interpreting accurate solutions from the observed results, which is must for KDD process."KDD is a combinational field of involving multiple technologies for data managerial activities like database management and data warehousing, statistic machine learning for intelligent retrievals, and goals accomplishment for decision making, and others such as visualization and parallel computing."[4]

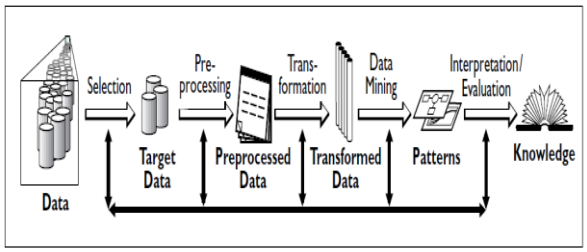

Fig. 1: *Steps constituting the KDD process*

The KDD process leads to its final conclusion of knowledge by involving following steps:

1. Data Selection involves the decision of relevant data to be analyzed and retrieved from the various data locations.
2. Data Preprocessing is the process of data cleaning and data integration. Irrelevant data are removed from the collected data in the Data Cleaning stage. Data Integration is a stage where multiple data sources are combined in a common source.
3. Data Transformation involves the transformation of the selected data into forms appropriate for the mining procedure.
4. Data Mining is the crucial step in which clever techniques are applied to extract potentially utile

patterns. The data mining technique to be used are then decided.

5. Interpretation and Evaluation identifies interesting patterns representing knowledge based on given proportions. The discovered knowledge is visually bestowed to the user.

4 DATA MINING TECHNIQUES

Data mining techniques provides the better business decision accomplishments, for detecting customer behavior patterns useful in formulating marketing strategies, sales techniques, profitable tracks and customer support schemes.

4.1ASSOCIATION

Association is also described as the relation technique as it leads to pattern discovery based on a relationship between items in the same dealings. Association technique is for finding patterns where one event is connected to another [14]. These verdicts help businesses to make vital decisions related to pricing, selling and to design the marketing methodologies. [8]Dataset generates a number of possible Association Rules but a high proportion of the rules are not valuable.**Apriori**algorithm defines frequent item set mining andassociation rule learning over transferable databases. It begins by identifying the frequent individual items in the database and item sets enhancement. [12] Association rules are determined using Apriori on the frequent item sets which highlight general trends present. [16]

4.2CLASSIFICATION

Classification is a classic data mining technique used to classify each item in a set of data into one of predefined set of classes or groups, based on machine learning. The methods of classificationfall back on mathematical approaches including statistics arena, linear programming methodology, neural networks, and decision trees. Fraudulent cases detection and credit risk appliance are well suited to this analysis. Learning and Classificationare the main steps included in data classification. Learning analyzes dataset using classification algorithms whereas Classification on the test data are used to estimate the accuracy of the classification rules [12, 15], that allows implementation of rules to the new data tuples. Fraudulent applications are traced by detecting complete records of frauds and invalid activities.**J48 classifier**allowsimplementing of the classification features for accounting for decision trees pruning, absent values, continuous attribute value ranges, rules derivation. In the WEKA data mining tool, J48 is an open source Java implementation. [2]

4.3 CLUSTERING

Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar essence. Clustering technique implementsclasses and puts objects in its respective class, whereas classification techniques assign objects into predefined classes.Preprocessing approach for attribute subset selection and classification is what we call Clustering of. [5] The **K-means algorithm**is a popular approach to find clusters due to its simplicity of implementation and rapid executional speed. It is included in data mining tools and machine learning literature. The employed simplicity makes the algorithm behavior complex. [1]

4.4SEQUENTIAL PATTERNS

Sequential patterns analysis is one of data mining technique that seeks to discover or identify analogous patterns, progressionsand general trends in transaction data over a business period. Analysis leads to discovery of various hidden trends, which is highly predictive of future events. Time series analysis methods are used that relate events in time based on a series of preceding events.[2]

5 WEKA

Weka is open source software for data mining under the GNU General public license, developed at the University of Waikato in New Zealand. The system is written using object oriented language java. “Weka” stands for the Waikato Environment for knowledge analysis. Weka provides implementation of state-of-the-art data mining and machine learning algorithmsincluding association techniques, classification rules, visualization, clustering methods, regression, filtering techniques by using weka tool. The data provides any meaningful information that can be used to know anything about any object. Information is than converted to knowledge to use with KDD. [4]

6 APPLICATION IN THE BANKING SECTOR

Extensive use of Internet Technology has immensely raised the application field of banking sector to involve a large customer force for the better decision making process.

6.1Customer Relationship Management

Customer relationship management (CRM) is a strategy that can help them to build long-lasting relationships with their customers and increase their revenues and profits. The CRM focus is shifting from customer acquisition to customer retention with the appropriate allocations of funds, time and resources. [3] Selling a product to the customer is outdated and demoded concept, with the objective to reach the heart of the customer and hence to develop a sense of belongingness for the organization. The Organizational data bases are storing billions of bits of information about the customers. Data mining isfunctional in all the three phases of a customer relationship cycle: Customer Acquisition, boosting customer value and Customer retention [5]. Customer profiling to group the like-minded customers in to one group and hence their dealing is the task of data mining techniques [12].

6.2Marketing

Interpreting the previous trends to determine the present demands and predicting the customer behavior of various products and assistance in order to grab more business opportunities and anticipate behavior patterns. Data mining has a number of techniques helpful for identifying profitable customers from non-profitable ones. Cross selling is another major area of development in banking where banks make an attractive offer to its customer by asking them to buy additional products. [13]

6.3Fraud Detection

While dealing with banks, the customers and the banks have the chances of falling an easy prey to fraudulent

activities and altercations. So both the parties need to establish the security parameters and wish not to be hampered while dealing.Detecting and preventing fraudulent activities is a consideration of data mining techniques, which will assist the organization to focus on the ways and means of analyzing the customer data in order to identify the patterns that can lead to scams [8, 12].

6.4Risk Management

Data mining technique helps to distinguish borrowers who repay loans promptly from those who don't. Customer behavior is determined as to identify that providing loan to him will result in bad loan decisions. Bank executives are using Data mining technique to analyze the behavior and reliability of the customers using credit cards, [10] andanalyzing them for making prompt or delay payment if the credit cards are provided to them.

7 IMPLEMENTATIONAND ANALYSIS

The Banking dataset is made to analyze forpattern detection for the business intelligence and decision based effectiveness for the banking systemthat are analyzed in weka environment and results are shown as follows.

7.1Data Visualization

The multiple transactions are carried out daily in a banking environment based on deposits and withdrawals that is analyzed through the patterns for the customer involvement. Banks engrosses customer that are single males with the low salary employers who don't prefer taking loans. The involvement of middle aged group is flat with the banking procedures but most of them are privately engaged. Customers populating after 2011 goes for taking the loans. Customer'sinterest is in depositing the money at their preferred location-Delhi, but they have low current balance. Transactionscarried through cheque is in the ultimate number. The result is evaluated as such.

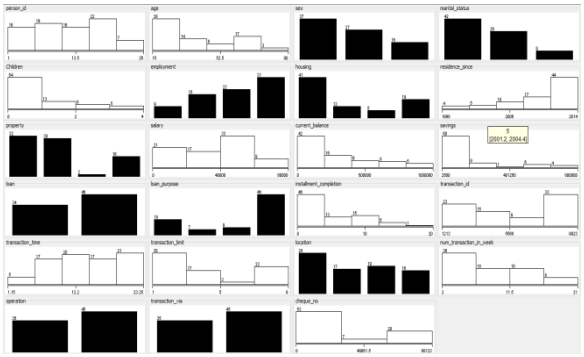

Fig 2: *Visualization patterns for attributes of customers involved with the banks*

7.2 Analysis using Clustering

Cluster formation implements simple K-Means to allow multiple iterations to follow to form the clusters based on the mean value of the data involved.

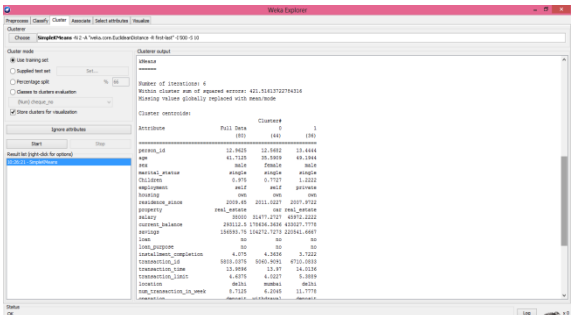

Fig. 3: *Implementation of Simple K-Means Algorithm*

It attempts to determine patterns for loan taken by the customer by their employment status and age. The privately employed customers have maximum involvement for loans, specifically car loans are preferred among all age groups whereas government employers hesitates in opting for loan facilities. The interest rates are preferentially less for in demand loans, whilst the education loan which is not preferred has comparatively higher interest rate.

Fig. 4: *Clustering analysis for loan with employment*

Salary plays a significant role in applying for loans as the customers having salary in the range of 25k-30k opt for loans of vehicles(mostly car), and more than 30k salariedprefer taking home loans whereas education loans are preferred by the 40k salaried customers. Many high salaried individuals don't take loans, which requires to be considered for the betterment of the bank.

Fig. 5: *Clustering Implementation of salary with loan*

Current balance in the account of a customer is depicted that describes the presence of high amount with the customers having salary greater than 50,000. Specifically the person with id 11 is showing fluctuating current balance with large number of transactions in a day.

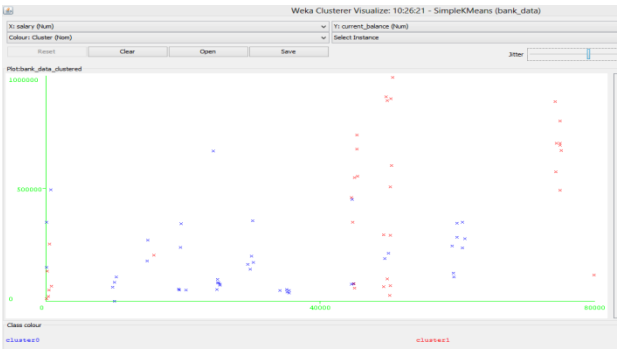

Fig. 6: *Clustering applicability on salary with balance*

7.3Analysis using Classification

Classification used to classify item set of data into predefined set of groups which is done by J48 tree pruning classifier.

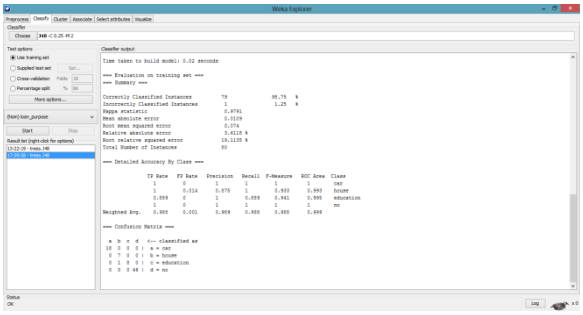

Fig. 7: *J48 Classifier Implementation*

Those customers takingloans have a current balance of more than 3L and they opt for education loans whereas the others owes house prefer taking car loans and who lives in free housing don't opt for loans and owes real estates .

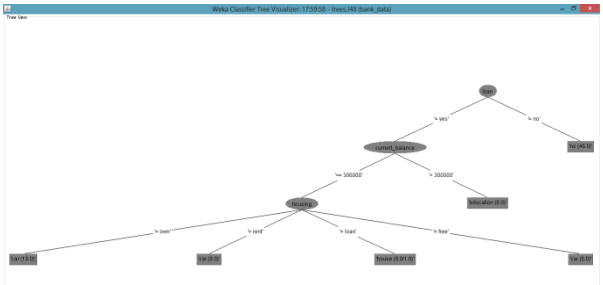

Fig. 8: *Determination of decision tree based on loan*

Customers having age limit less than 25 and those having less than 6 number of transactions in a week goes for car loans with the maximal installments of ten.

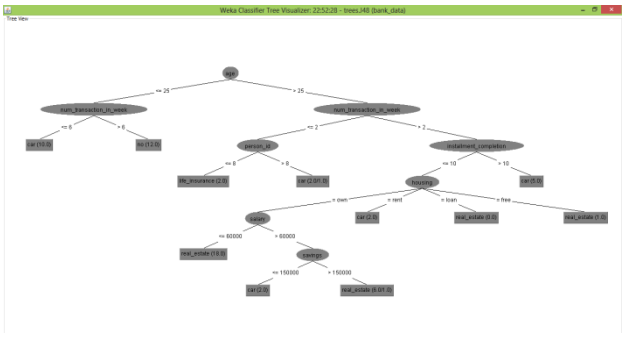

Fig. 9: *Classifier tree for Insurance*

Most of the car and property owners are interested in withdrawing money whereas customer's not owing property will deposit in their savings.

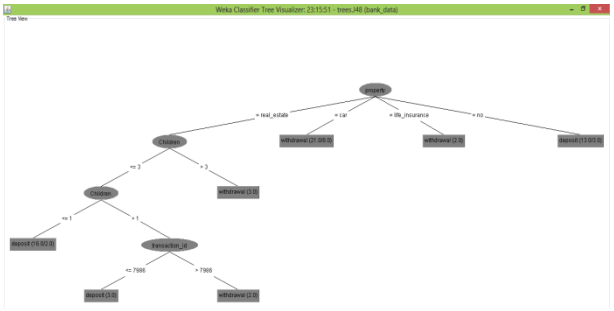

Fig. 10: *Classification based on transactional operations*

8 CONCLUSION

The following patterns are analyzed:

1. CRM depicts the pattern for customer relationships with the low involvement of females and young adults in the banking procedureswhereas middle aged are more involved and prefer branch having better services and in proximity.
2. On analyzing CRM trends, we concluded that salary is a prime factor for loan consideration: as education and personal loans are opted by low salaried customers followed by car loans and property loans.
3. Multiple transactions that occur near about the same time from different locations depicts the fraudulent analogies.Concurrency in withdrawals and deposits patterns of high amounts are done in a short period.
4. Marketing involved formulating schemes based on the customer's income. Insurance policies are preferentially taken by middle class and upper middle class whereas the lower class people prefers the loan schemes.
5. Risk is a considerable factor for the banking sector that involved analyzing fraudulent transactions from cheques along with exceeding or skipping the loan installments, dates many a time.
6. The trend ofhigh savings in a customer account and locker account is relevant to their marital status as the married couples prefer banking facilities with a consideration for their benefits.

REFERENCES

[1] Ian Davidson. “Understanding K-Means Non-hierarchical Clustering”, Univ. of California publications,2014.http://academic.research.microsoft.com/Keyword/21448/k-means-algorithm

[2] A. Floares., A. Birlutiu. “Decision Tree Models for Developing Molecular Classifiers for Cancer Diagnosis”, WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia.

[3] K. Chitra, B.Subashini, “Customer Retention in Banking Sector using Predictive Data Mining Technique”, International Conference on Information Technology, Alzaytoonah University, Amman, Jordan, 2011.www.zuj.edu.jo/conferences/icit11/paperlist/Papers/

[4] Xindong Wu, Xingquan Zhu , Gong-Qing Wu. ,”Knowledge and Data Engineering” , IEEE Transactions on (Volume:26 , Issue: 1) ,2011. http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?reload=true&punumber=69

[5] S.P. Deshpande, Dr. V.M. Thakare,”System and Applications: A Review Data Mining “, International Journal of Distributed and Parallel system, September2010.

[6] D. Muraleedharan, “Modern Banking: Theory and Practice”, PHI Learning private Limited, 2009.

[7] Rajanish Dass, "Data Mining in Banking and Finance: A Note for Bankers", Indian Institute of Management Ahmadabad, 2008.

[8] Dr.Madan Lal Bhasin, “Data Mining: A Competitive Tool in the Banking and Retail Industries”, the Chartered Accountant October, 2006.

[9] Hillol Kargupta, Anupam Joshi, Krishnamoorthy Siva Kumar, Yelena Yesha, "Data Mining: Next Generation Challenges and Future Directions", Publishers: Prentice-Hall of India, Private Limited, 2005.

[10] I.H. and Frank, “Data Mining: Practical machine learning tools and techniques. “. 2nd edition Morgan Kaufmann, San Francisco, 2005.

[11] M. De Martino, A. Bertone, R. Albertoni, H. Hauska, U. Demsar, M. Dunkars. :”Technical Report of Data Mining”, INVISIP IST-2000-29640, Information Visualisation for Site Planning, WP No2: Technology Analysis, D2.2, 2002.

[12] S. S. Kaptan, N S Chobey, "Indian Banking in Electronic Era", Sarup and Sons, Edition 2002.

[13] S.S.Kaptan, “New Concepts in Banking”, Sarup and Sons, Edition, 2002

[14] M. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New Algorithms for Fast Discovery of Association Rules", Proc. 3rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'97, Newport Beach, CA), 283-296 AAAI Press, Menlo Park, CA, USA 1997.

[15] Drummond, C. and Holte, “An alternative to ROC representation.” Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Publishers, San Mateo, CA.

[16] Aggarwal, R., T. Imielin´ ski, and A. Swami, "Mining association rules between sets of items in large databases". In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD ’93, New York, NY, USA, pp. 207–216. ACM.