iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: XI Month of publication: November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75208

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

Prediction of Closure of Accounts and Non-Performing Assets (NPA) - Regression and Attribute Oriented Induction

J. Stanly Thomas[1], Dr N Rajkumar[2]
[1]Research Scholar, Dravidian University, India,
[2]Director (Research), KGISL Institute of Technology,India

Abstract: This research work is fully focused on prediction technique which helps banks to predict and to avoid the specific account closure and non-performing assets. The significant of this research work has arisen owing to current shape of the banking industry further deteriorating would directly affect the economy of this country. The attribute oriented induction is used to filter the relevant attributes from noise text database. The prediction technique of multiple regression is used to highlight the factors which shall lead to the closure of accounts and non-performing assets on every stage. Further, this research work is added-up with recommended best practices for each scenario from the real time to achieve zero account closure and zero non-performing assets.
Keywords: Relevance Analaysis, Linear Regression, Multiple Regression, Attribute Oriented Induction, Prediction methodologies

I. INTRODUCTION

This entire research work is classified into prediction of account closure in first cycle and prediction of non-performing assets in next Cycle. As far as first cycle concerned, attribute oriented induction on relevant analysis undertakes a significant role in filtering of relevant details towards the prediction from the huge data warehouse of banking industry.

II. RELEVANCE ANALYSIS

As a banking industry is dealing with massive volume of data, filtering of relevant details for the purpose of analysing are be preceded by classification or prediction of attributes which do not have any relevancy on prediction process. Hence, in order to start the relevance analysis, the various reasons for account closures have been identified as below.

Sl.No	Reason for account closure	Datasets
1	Customer Dissatisfaction	Complaints database on all the modules including CASA, Advances, Deposits, Electronic transactions (ATM, point of sale, ECOM, Internet Banking, Mobile Banking), pension and para banking activities such as Insurance, Mutual Funds etc.
2	Un resolved Complaints	Complaints database
3	Electronic Transaction Decline (Technical Decline)	Electronic Transactions database
4	Shifting of residence/offices	Account Closure Request database
5	Lack of competent interest rates/attractive products	Marketing database
6	Fraudulent Transactions	Risk Management Database

Table 1 Reason For Closure Of Accounts And Related Database

III. MULTIPLE REGRESSION

In order to predict the multidimensional reasons that factors closure of account is called as multiple regression. This is nothing but prediction of group of reasons which shall cause the closure of accounts and it is an enhanced version of multiple parameters under linear regression. It enables the linear function characteristics under the response variable L under the multi-dimensional vectors. The regression with two parameter attributes of variables, X1 and X2, is

$$L = a + \beta X_1 + \beta X_2$$

where the 'L' called as response variable and Xs' are called as predictor variables. In case of constant variable of "L" and values of a and β are the coordinates and regression coefficients indicating the intercept of Y and slope of the line respectively.

Given 'n' samples or data points of the form $(x_1,y_1),(x_2,y_2)..(x_n,L_n)$ then, the regression coefficients can be estimated using this method of following equations

$$\beta = \frac{\sum_{i=1}^{n} (\overline{X}_i - x)(\overline{L}_i - L)}{\sum_{i=1}^{n} (\overline{X}_i - x)i^2}$$

$$\alpha = \overline{L} - \beta\, \overline{x}$$

where $\overline{x}$ is the average of $x_1, x_2, \ldots\ldots\ldots x_n$ and $\overline{L}$ is the average of $L_1, L_2, \ldots\ldots L_n$.

IV. RELEVANCE ANALYSIS USING ATTRIBUTE ORIENTED INDUCTION (FIRST CYCLE : PREDICTION OF CLOSURE OF ACCOUNTS)

The emphasis of reducing noise data directly improves the accuracy and speedy process of any algorithm. The basic concept of attribute oriented induction on relevance analysis is to arrive at the balanced generalization of data which is exactly relevant and informative (not too high/not too low) for prediction. Further, in order to extract the relevant data in speedy manner the distinct attributes should be lesser but, the coverage of database should be higher percentage. The following table 2.3 shows the distinct attributes which have been tested to arrive at the balanced generalisation of data with the unrestricted threshold value to measure the banking data warehouse.

Sl.No	Database	Distinct Field	No. of Distinct Value	Overall Tuple
1	Customer Complaint (along with number of days pending) Reason: Customer dis-satisfaction & un resolved customer complaints	Account Number	5602375	5613129
		Customer ID	5124861	
		Module ID	10	

2	Electronic Transaction Reason: Electronic Transaction Technical Decline	Account Number	4917845	5152713
		Debit/Credit Card Number	4865034	
		Module ID	2	
3	Account Closure Database Reason: Shifting of Residence/Office	Account Number	12867	12992
		Customer ID	12549	
		Module ID	4	
4	Marketing Database Reason: Lack of competent interest rates/attractive products	Scheme code	58	58
		Module ID	7	
5	Risk Management Database Reason: Fraudulent Transactions	Account Number	127	127
		Customer ID	127	
		Module ID	2	

Table 2 Distinct Value Analysis On Data Attributes

In order to predict the closure of accounts accurately and fast after considering 100% coverage to the data tuples, maintaining of less distinct value data attribute as prime generalization is significant. The above table is clearly shows the generalization of query based on the distinct attributes "Module ID" is the less distinct value which shall be used for prime generalization.

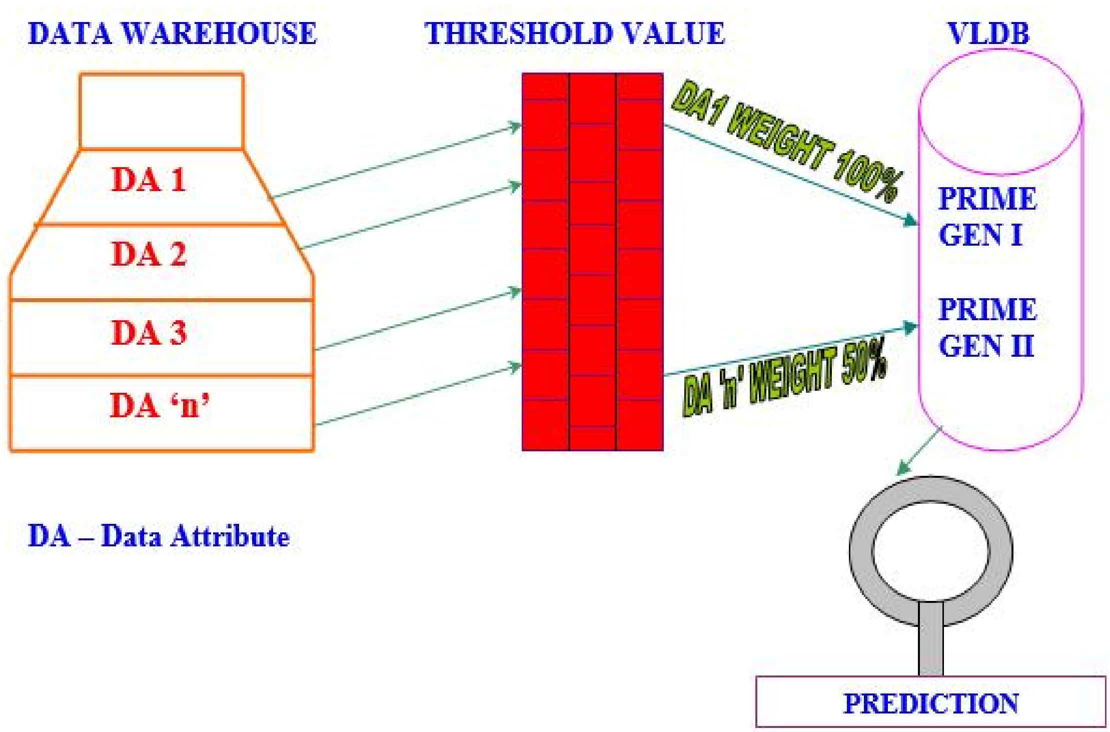


FIG 1 Attribute Oriented Induction

The coverage of entire database by the data attribute is called as weight over the target class. This shall be derived as say 't_a' denoted as the target class. The transformation percentage of tuples of the unshaped data or initial working relation to shaped target class is called weight of 't_a' denoted as 't_n'. This shall be derived as follows

$$\text{Weight}=\text{count}(t_a) / \sum_{i=1}^{n} \text{count}(t_a)$$

Under the generalized relation where *n* denotes the number of tuples for the target class. The tag ***ta*** exists under the range of t_1 to t_n. Therefore, the range for the weight is [0.0, 1.0] or [0%, 100%].

A. Algorithm – Attribute Oriented Induction

Function attrelevance()
Begin
S -> Select Task Relevant data // Selection of various related databases as above
Prepare for generalization (S)
Begin
    Read S and gather all values of elements
// if knowledge on data in the database is presence then collect with desired attribute
//For each attribute, determine whether the generalization is to be removed or not //based on the threshold value as follows
if (no.of distinct value > threshold value) then
Begin
    Remove generalization
Else
    Accept generalization
End
// Compare each attributes by no. of distinct values and fix prime attribute with the // less distinct value and index based on the same
// so that accurate and speedy prediction shall be possible
End
End function

V. MULTIPLE REGRESSION ON PREDICTION

In this research work, multiple regression has learned with the background process based on the historical data which leads to closure of accounts. Hence, the algorithm is able to predict the similar situation in real time environment with the sensitivity called "Low" and "High". Basically, this research work is tuned to handle individual reasons which lead to closure of accounts and it shall be identified based on the average value fixed by the user for each category of reason. However, historical data analysis reveals that all other categories are finally migrated to "Customer dissatisfaction" which are immediately result in closure of accounts. Therefore, the sensitivity flag is fixed as "High" for this category.

This algorithm handles combination of categories which leads to closure of accounts based on the average value of each category. For instance, the same account is filtered under continuous technical decline of transactions as well unresolved complaints. The technical decline is the terminology used for the scenario where a customer account in order but owing to technical issues system could not facilitate the desired function. In this scenario, average of both the category shall be summed up and the alert value will be in "Higher" degree of value. The below pictorial representation illustrate the role of multiple regression under account closure prediction.

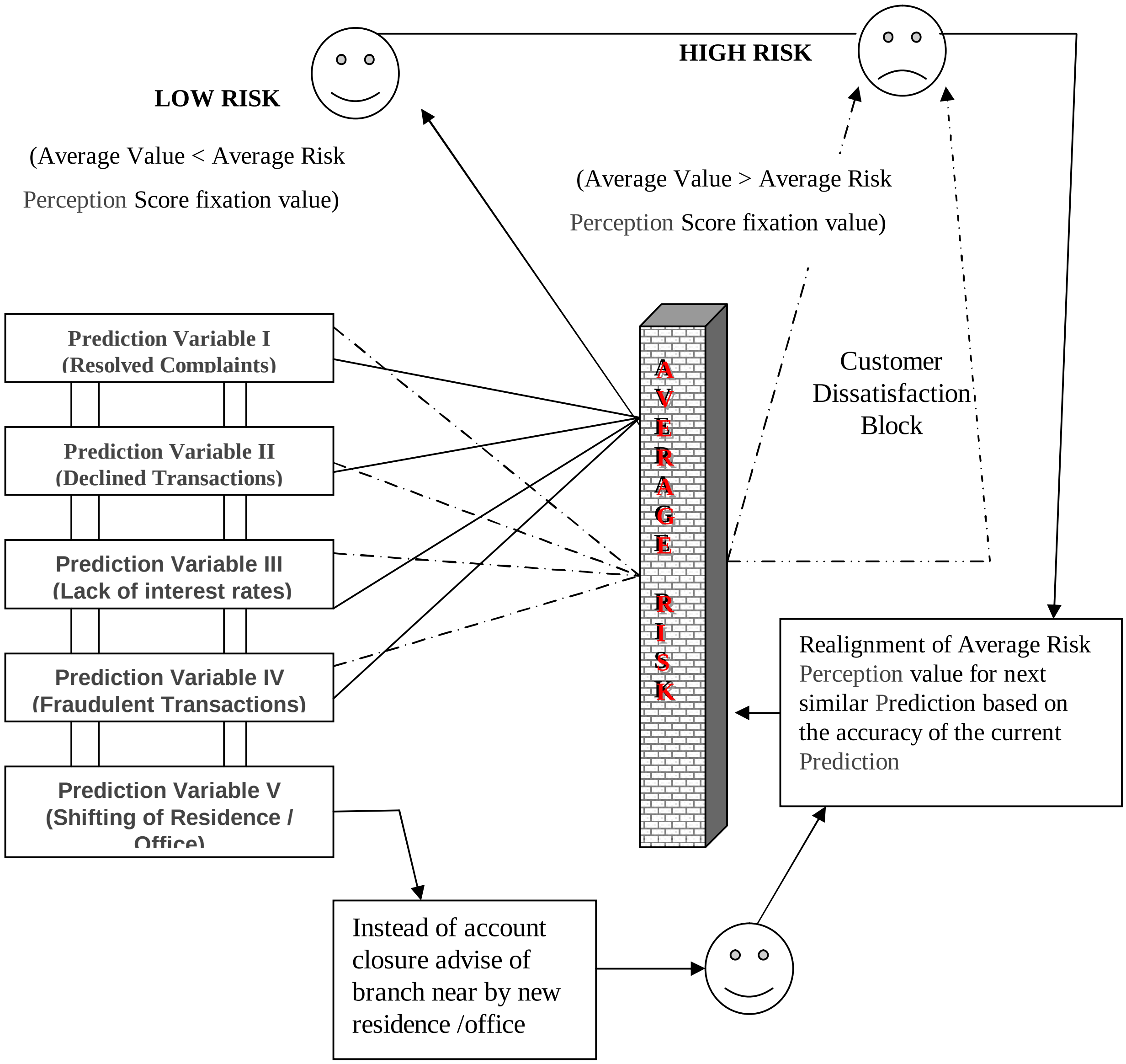


FIG 2 Multiple Regression On Prediction

A. Algorithm for Multiple Regression on Prediction (Closure of account) :

1) Fixation of risk perception average score value for each category other than "Customer dissatisfaction" since it is normally "High" alertness is advised to the user.
2) If the predicator variable is single and linear regression value is more than the risk perception average score value it shall be alerted as "HIGH" or otherwise "LOW" to the user.
3) If the predicator variable is more than one and multiple regression value is directly alerted as "HIGH" instead of checking the risk perception average score value since, it falls under more than one scenario.
4) Fixing of occurrences for the migration to "Customer Dissatisfaction".
5) Either occurrences is more then the fixation limit or the predicator variable is more than one then the category shall be moved to "Customer Dissatisfaction" with "HIGH" alert and immediate attention required on the same.

If the advised prediction is accurate and user shall input the accuracy level so that average risk perception shall be increased for the next similar analysis.

VI. SECOND CYCLE: PREDICTION OF NON PERFORMING ASSETS (NPA)

An asset quality is depending upon the credit worthiness of the borrower. The monitoring of asset quality on different stages is mandatory to predict and avoid the non-performing asset as this reduces the profitability of any bank. Hence, the research on prediction of non-performing assets is very significant for the banking industry to undertake the profitable business. Basically, the assets are classified as follows and our research work is playing vital when the asset/account is in standard. The main motive of this research is to predict the account/assets which may migrate to sub-standard, so that bank shall initiate necessary action to avoid the above migration. Since, the asset quality depends upon the repayment behavior of scheduled Earnest Money Interest (EMI), we shall predict the NPA based on the same.

Asset Quality	Expression of terminology
Standard Accounts/Assets	Active accounts intact in all the aspects
Sub-standard Accounts/Assets	Accounts are overdue for the period not exceeding 12 months and classified as non-performing assets- NPA under the books of accounting.
Doubtful Accounts/Assets	Accounts are overdue for the period exceeding 12 months and classified as non-performing assets- NPA under the books of accounting.
Loss Accounts/Assets	Accounts are remains overdue and unrealizable.

TABLE 2 Asset Quality Terminologies

As discussed earlier, the NPA is directly depends upon the EMI payment the number of regression also fixed on the same. Therefore, the predictor variable is strictly focusing on EMI schedule, payment behavior and the deviation on fixed average risk perception shall escalate the deviation and course of action to different layers of people.

A. Linear Regression on Prediction

The predictor variable is fixed on EMI payment and average risk perception value is fully depends on user configuration for the same. The escalation and follow up matrix of the deviation is also depends on the user configuration. The algorithm is fed with the background knowledge of EMI schedule and the amount. It triggers SMS alert and email alert to borrower before 10 days of the EMI schedule about the due payment. This is configured with the algorithm which completely overcomes the possibility of non-payment of EMI due to the borrower unawareness about the schedule and the amount. This algorithm is maintaining the risk value inherited from EMI payment. It adjusts the risk value based on the promptness in EMI payment (low for regular payers and high for others). Based on this risk value bank shall report the borrower credit quality to the CIBIL. Similarly based on this risk value, the SMS alert and email alert triggered to different people of the bank to take follow up action on the same to avoid NPA.

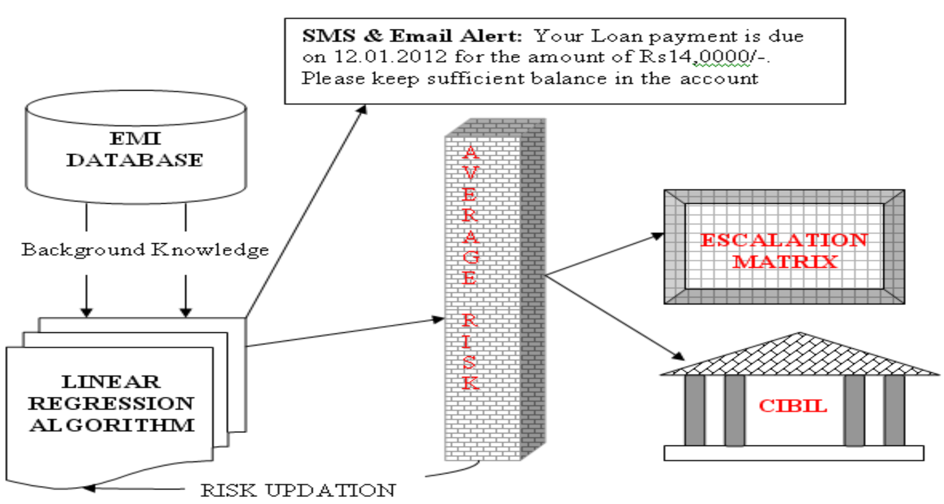


FIG 3 Linear Regression On Npa Prediction

B. Algorithm of Linear Regression on prediction

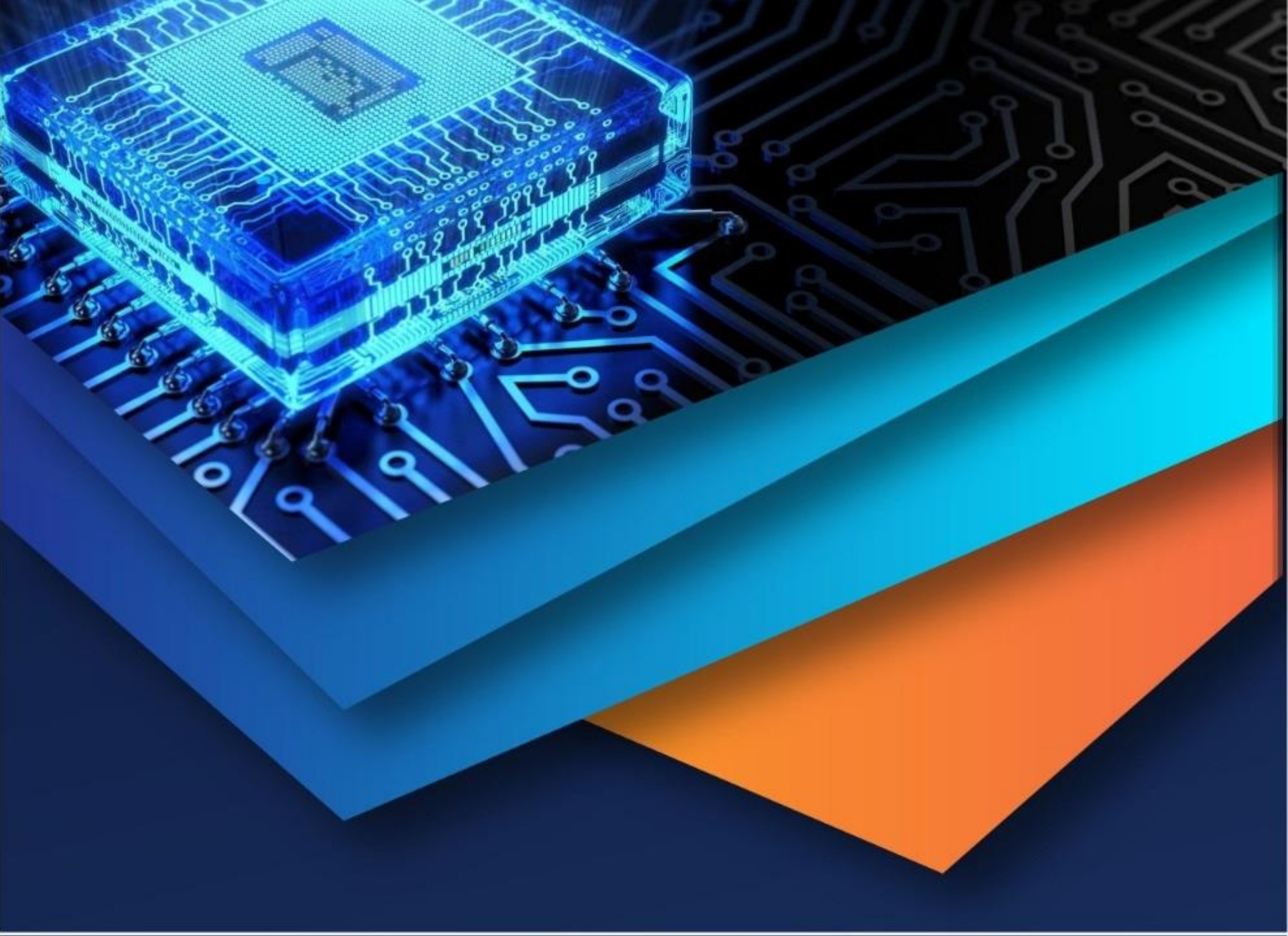
1) Literate the system with EMI payment schedule details
2) Assign the predictor variable as EMI payment
3) Configure SMS and email alert about the EMI due in advance
4) Analyse EMI payment attendance
5) Load risk value based on EMI payment
6) Configure Average Risk perception value to withhold the risk and avoid the NPA
7) Assign the linear regression behavioral pattern based on the average risk perception value
8) Configure escalation matrix based on the behavioral pattern and advise

VII. CONCLUSION

The attribute oriented relevance analysis and prediction techniques on Closure of accounts and non-performing assets shown its real worthiness in the banking industry. The era of doing this research work is thrust on the prediction techniques especially in banking industry to avoid great damage on the business and to sustain the business. The attribute oriented induction on relevance analysis deploys the strong platform to filter the very large banking database. The multiple regression method has increased the accuracy level of this prediction due to the combination of various predictor attributes. This research work shall be enhanced further on other critical areas of banking industries like prediction of money laundering, credit worthiness of the loan applicant, letter of credit, bill of lading and bank guarantee.

REFERENCES

[1] Heckerman. Efficient approximations for the marginal likelihood of incomplete data given a Bayesian network. Machine Learning, 29:181–212, 2007.
[2] S. Chib. Marginal likelihood from the Gibbs output. JASA, 90:1313–1321, 1995.
[3] C. Carter, Markov chain Monte Carlo in conditionally Gaussian state space models. Biometrika, 83:589–601, 2006.
[4] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. IEEE Trans. on Info. Theory, 14:462–67, 1968.
[5] S. Liu. Mixture Kalman filters. J. Royal Stat. Soc. B, 2000.
[6] T. H. Cormen, and R. L. Rivest. An Introduction to Algorithms. MIT Press, 1990.
[7] G. F. Cooper. A simple algorithm for efficiently mining observational databases for causal relationships. Data Mining and Knowledge Discovery, 2007.
[8] A. Tuzhilin, “Abstract pattern discovery in databases” IEEE Trans. Knowledge and Data Engineering, 1993
[9] J. Li, “Efficient mining of emerging patterns – Discovering trends and differences”, Int. Conf. Knowledge Discovery and Data Mining,1999
[10] R. Duda and P.Hart, “Pattern Classification and Scene Analysis.”, New York, ohn Wiley & Sons, 1973.
[11] D.W. Cheung, and Y. Fu, “A Fast Distributed Algorithm ,” Proc. 2006 Int’l Conf. Parallel and Distributed Information Systems, PP. 2006 Int’l Conf. Data Enf., PP. 106-114, New Orleans, Feb. 2006.
[12] R. Agrawal and R. Srikant, “Mining Sequential Patterns,” Proc. 1995 Int’l Conf. Data Eng., pp.265-276,Tucson,Ariz., May 2007.
[13] J.H. Friedman, “A recursive partitioning decision rule for non parametric classifiers”, IEEE Trans. On Com,1977
[14] S. Chaudhuri and U. Dayal, “An Overview of Data Warehousing and PLAP Technology,” AGM SIGMOD Record, vol. 26, pp. 65-74, 2007.
[15] T. Fu and J. Han, ”Meta-Rule-Guided Mining of Association Rules in Relational Databases,” Object-Oriented Databases (KDOOD ’95), PP. 39-46, Singapore, Dec. 1995.
[16] M.S. Chen, and P.S. Yu, “Data Mining: An overview from a Database Perspective,” IEEE Trans. Knowledge and Data Engg., Vol.8, PP.866-883,2006.



10.22214/IJRASET

INDEX COPERNICUS 45.98

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY