Framework to predict NPA/Willful defaults in corporate loans: a big data approach

Girija Attigeri, Manohara Pai M M, Radhika M Pai
Department of Information and Communication Technology, Manipal Institute of Technology,
Manipal Academy of Higher Education, India

Article Info

Article history:

Received Mar 6, 2019
Revised Apr 9, 2019
Accepted Apr 18, 2019

Keywords:

Big Data
Corporate Loan
Financial Fraud
Machine Learning
Non Performing Assets
Parameterization
Willful default

ABSTRACT

Growth and development of the economy is dependent on the banking system. Bad loans which are Non-Performing Assets (NPA) are the measure for assessing the financial health of the bank. It is very important to control NPA as it affects the profitability, and deteriorates the quality of assets of the bank. It is observed that there is a significant rise in the number of willful defaulters. Hence systematic identification, awareness and assessment of parameters is essential for early prediction of willful default behavior. The main objective of the paper is to identify exhaustive list of parameters essential for predicting whether the loan will become NPA and thereby willful default. This process includes understanding of existing system to check NPAs and identifying the critical parameters. Also propose a framework for NPA/Willful default identification. The framework classifies the data comprising of structured and unstructured parameters as NPA/Willful default or not. In order to select the best classification model in the framework an experimentation is conducted on loan dataset on big data platform. Since the loan data is structured, unstructured component is incorporated by generating synthetic data. The results indicate that neural network model gives best accuracy and hence considered in the framework.

Copyright © 2019 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Manohara Pai M M,
Department of Information and Communication Technology,
Manipal Institute of Technology, Manipal Academy of Higher Education,
Manipal, 576104, India.
Email: mmm.pai@manipal.edu

1. INTRODUCTION

Banking system with integration of advent technology helps to foster the economic development. They perform mainly two important functions. One is mobilizing deposits by providing attractive interest rates to convert inert savings into active capital and second is distributing these deposits through loans to the corporates to grow further that directly helps in economic development. Availing loan has become an easy process in India with the credit and cheque settlements. Banking as well as Non-Banking Finance Companies offer different types of loans according to requirements of corporates [1, 2]. The requirements can be purchase of inventory, payment of long unpaid bills, building of infrastructure, purchase of equipment, loan repayments and so on [3]. Based on the requirements loans are broadly classified as Personal Loan, Credit Card Loan, Home Loan, Vehicle Loan, Education Loan, Loan against the Insurance Schemes/FD/Mutual funds, and Business Loan to Corporates. There are several business loans possible such as Working capital loan to use in day to day activities, Real Estate loan to buy a property for production, Venture loan to start up business, Line of credit loan for certain financial assistance periodically, Equipment loan to assist buying asset requirements, Term loan to acquire long term fixed assets, Loan against property for supporting business by providing security to the corporates, Cash Credit facility as overdraft against the security of the stock by pledging the current assets and Letter of creditwith which the bank guarantees that the seller will receive payment on certain conditions.

In this paper, focus is on the process of managing corporate loan, as the recovery of these loans is tedious task and it affects economy of the country heavily. Healthy banking system represents healthy economy of the nation. However there are hindrances to achieve the required set up for the same. All the corporate loans borrowed do not end up as assets for the bank. There are two outcomes of the loans: One is performing assets and another one is Non-Performing Assets (NPA). NPAs are the loans which generate the loss in capital of banks and are not easily recoverable by the banks. This is the most tedious challenge for banking sector as it impacts the performance by declining the profits. It has become major problem for all public sector and private sector banks. According to RBI report 2016, total gross NPA amount was 6 lakh crores. By 2017 there was increase of 1 lakh crore [7.31 Cr]. Losses are over four times more than the profits indicating NPA's power to trap the economy of the country in vicious debt cycle.

Banks provide big loans to corporates in order to achieve higher profits [4]. Companies start behaving as a defaulter by showing losses in company's finance. Some deliberately don't repay even if they have sufficient financial resources to pay. Companies propose to pay these loans by taking other loans from multiple banks. Unrecoverable loans put company to bankruptcy status. Due to the willful default behaviour, many genuine companies do not get economic support at the time of need and may end up in the failure. Hence, such companies may not be able to pay existing loans, and get defaulter tag. When an individual or business enterprise declines to fulfill payment commitments with financial institutions even when it has sufficient capacity for repayment, such a borrower unit is considered as willful default. More bad loans get generated because of existing bad loans. In order to prevent the failing economy the banks decide to lend money to save the companies going bankrupt. They take advantage of the situation leading to increased willful defaults, bringing the economy back to the same status.

It has been observed that there is a rising trend in NPAs, especially in public sector banks. There are several causes for this and there is a strong evidence for defining the relation between fraud and NPA [5]. RBI data obtained through RTI request indicate that 8670 loan fraud cases amounting Rs. 612.6 billion are recorded over last five financial years. These frauds are referring to cases where borrower deliberately tries to deceive the bank and does not repay the loan amount [6].

NPA is a major issue faced by all commercial banks as the banks are considering loan advances as revenue generating asset. Quality of this asset need to be taken care to improve the profitability of financial institutions, and ultimately financial climate of the economy as a whole. In this regard early detection of NPA is a big relief for such a major problem faced by all financial institutions. While analyzing the scenario, it is not just the poor economic conditions that resulted in NPA, but deliberate defaults have also resulted in huge piled up of NPA. Hence, more emphasis on identifying willful default is the need of the hour. The objective of this paper is to build a data model for early detection of the willful default. The process of classifying willful default involves considering various categories of parameters such as financial, personal, social etc. The data to capture these parameters could be structured, unstructured and needs continuous analysis. Also, the data need to be captured from various sources which are called as Heterogeneous sources. Hence, Big data technology is used to design a data model where the size, variety and complexity of data can be handled effectively [7]. Various classification algorithms are designed using big data approach to evaluate the classification data model for prediction of early stages of NPAs.

The rest of the paper is organized as follows. Section 2 discusses about the background of the work with respect to national and international scenario and literarture review. Section 3 describes parameterization process and data model for npa/willful defualt indentification. Section 4 explains the framework for NPA/willful default identification. The validation process of framework and evaluation of prediction model is explained in Section 5. Secion 6 drwas conclusion.

2. BACKGROUND

There are several policies, schemes which the law agencies have set up to deal with the falling economy of the country. The extensive literature survey has been carried out for studying the NPA scenario in India as well as in various other countries [8-10]. Also, the study has been done to identify various parameters which are useful for NPA identification process.

2.1. International scenario

China recently has 250 million dollars of bad debt. These are mainly the loans that are directly related to real estate, used to develop the infrastructure. The state of the NPA is due to political and social implications, legal impediments, bankruptcy laws, real estate. Italy also faced 207 billion dollar bad debt due to real estate. Bad debts along with national debts incurred huge financial crisis for the country. However as it is part of the Eurozone, it was saved by bailout funds provided to recapitalize banks.

Russia accounts up to 9.16% NPA ratio of the total loans. It is because the country is mostly dependent on oil and gas exports. When global oil prices crashed down, it marked collapse of Russia's oil and gas industries. In turn, banks approved loans to rescue the economy. However the sanctioned money never came back leading to the financial crisis. Adding to its economic sanctions imposed by America and other European countries caused slowdown in economic growth.

Spain faced debt/housing crisis of unpaid loans in 2008. But Government provided fix to the issue quickly with remedial measures. As a result, the bad debt decreased from 6.09% in 2016 to 5.7% in June 2017. Ireland is facing the NPA issue due to economic slowdown. It has set up National Asset Management Agency for insolvency services to support real estate and housing debtors. These remedial measures have dropped country's bad loan ratio from 27% in 2013 to 14.2 percent in 2016. This trend is contributing to success which can be attributed to successful debt restructuring programs.

2.2. Indian scenario

In Indian financial sector during 2017, the most discussed topics are GST, demonetization and NPAs. NPA has led to almost 10% of the loans impacting around 9 lakh crore, affecting Indian economy negatively. RBI is the Indian banking institute which coordinates and regulates the activities of the banks in Indian economy.

Major challenges faced by RBI are mentioned as follows [3]:

a. NPA: As explained earlier, NPA is the indicator to identify the status of the corporate loans which are not regularly settled and creates financial crisis for banks as well as for company. NPAs affect the financial growth of the bank and hence declines the economic condition of the country. In order to tackle this, Indian government has taken many initiatives.
   1) Temporary relief of several thousand crores
   2) Special courts to deal with companies having bad loans
   3) Reduced interest rates
   4) Merger of banks to reduce burden of bad loans

b. Bank Frauds and Cyber Threats: Bank frauds and cyber-attacks on financial transactions are illegal means of obtaining the money or assets especially from bank. One way to obtain money from a bank is to take out a loan, which bankers are more than willing to encourage if they have good reason to believe that the money will be repaid in full with interest. A fraudulent loan, however, is one in which the borrower is a business entity controlled by a dishonest bank officer or an accomplice. The "borrower" then declares bankruptcy or vanishes and the money is gone. The borrower may even be a non-existent entity and the loan merely an artifice to conceal a theft of a large sum of money from the bank. This can also be seen as a component within mortgage fraud.

   Today's robbers are doing robbery behind the internet using targeted and sophisticated cybercrime tactics. Some of the example attacks are phishing, Carbanak malware, SQL injection attacks, and attacks on bank database, credit cards and on online financial transactions. IT teams at banks have increased protection of customer data and limited credit card fraud, but the security of most banks' internal systems still need to be improved.

c. Increase in excess liquidity: The increase of penalty rate will increase the interest rates and excess reserve owned by banks. Therefore, the total liquidity in economy will increase rapidly without involving policy rate reduction mechanism (loose monetary policy), just when the liquidity should be restricted. The reason behind increase in excess liquidity in bank is the economic condition which is in liquidity trap. Liquidity trap is a condition where return from banking loan is too small to cover intermediation cost and banks get higher yield in reserves than giving loans. In this condition, expansive monetary policy will only cause increase in excess reserves. Due to increase in liquidity, financial crisis are increasing in the banks which leads to weakening the domestic currency with respect to international currencies.

2.3. Literature review

In Indian financial sector during 2017, the most discussed topics are GST, demonetization and NPAs. NPA has led to almost 10% of the loans impacting around 9 lakh crore, affecting Indian economy negatively. RBI is the Indian banking institute which coordinates and regulates the activities of the banks in Indian economy. Charan and Brar [11] have presented a study on stressed assets in India. They have mentioned about identification of number of factors that lead to this situation. They have identified broad categories of the reasons such as stress for global slow down, governance related issues, political factors as well as malintentions and misconduct. They also emphasize need for extensive research into the factors that cause deteriorating asset quality in public sector banks.

In the study on frauds in the Indian Banking Industry by Charan et. al., they used interview based approach to identify the reasons for frauds in banking sector [5]. They mention the main factors as lack of supervision from the management, lack of incentive mechanisms for employees, non-cooperative staff, corporate borrowers and third party agencies etc. One very important thing noted is absence of strong regulatory system and absence of tools and techniques to detect early warning signals.

Bardan and Mukhrjee [7] deal with willful default and its implications for profitability and decision-making process of the loans at banks. They examine the cases where the borrower defaults willfully by under reporting its cash flow. In the analysis they mention it is necessary for the regulator to choose lower loan capacity to avoid NPA levels at the bank due to willful default. However, it will exert sinking pressure on the profit level of the bank. Hence it will face a trade-off between greater incidence of willful default and higher profit of the bank. They also emphasize that the reason for increasing willful default is weak monitoring and supervision system, poor bankruptcy laws in developing countries like India. All these give opportunity for the borrower to willfully default the loan.

The research papers [12-16] show that the risks the banks face and default behavior were challenges even two decades before, however with the new technologies at hand challenges have become more difficult to address. As per the literature study and national and international scenarios, the aim of the work is to define the standard process to identify the parameters which can be used for early detection of fraud behavior and further helpful for early identification of NPAs/Willful default.

The objectives of the proposed approach to identify NPAs/Willful default are as follows

a. Understand the loan process
b. Identify data parameters for early identification of willful defaulters or NPAs
c. Identify suitable technology and develop model and algorithms for willful default identification

3. PARAMETERIZATION PROECESS FOR NPA/WILLFUL DEFUALT INDENTIFICATION

The process of loan sanctioning after the request for loan till the completion of it is shown in Figure 1. The main important block of the loan process is monitoring the financial health of the corporate/customer in order to understand the fraud or willful behavior. Monitoring financial health need various parameters which are closely associated with the purpose. Hence there is a requirement to define standard process to identify parameters which will be useful to define the data model for early prediction of frauds, willful defaults and further NPAs.

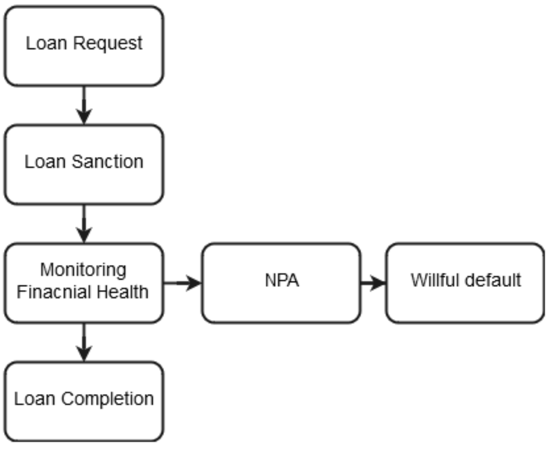

Figure 1. Loan sanction and recovery process

Parameterization process is the important process of identifying essential and critical parameters for carrying out a particular analytical task and coming out with valuable outcome. For willful default identification in the loan scenario the process is defined and is shown in Figure 2. The process starts with identifying sources that help to understand the various terminologies of the loan and causes of NPA and thereby willful default.

The defined parameterization process considers the parameters from different sources such as RBI document, literature, case analysis reports, brainstorming session, bank documents and so on. These parameters are huge and unstructured and hence need to be classified into broad categories to further capture specific parameters for each category along with the ranges. The process is dynamic in nature, covers parameters related to fraud and identifies the change in ranges as per the categories.

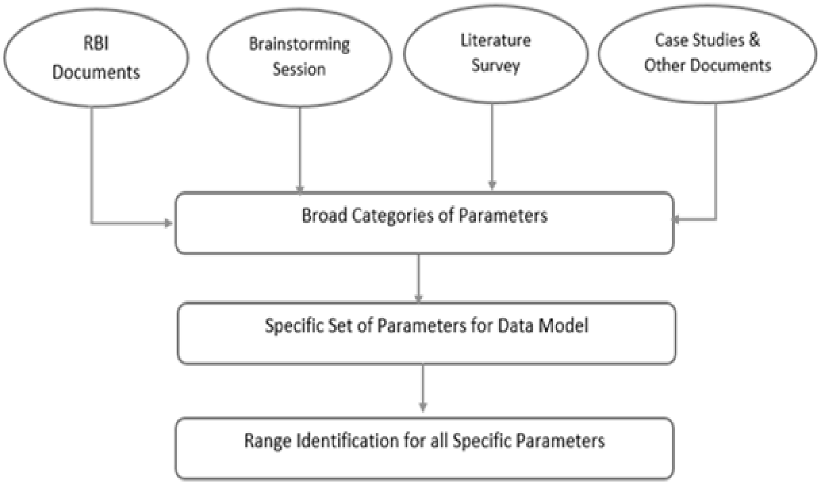
RBI Documents
Brainstorming Session
Literature Survey
Case Studies & Other Documents
Broad Categories of Parameters
Specific Set of Parameters for Data Model
Range Identification for all Specific Parameters

Figure 2. Parameterization process

According to RBI circular RBI/2015-16/100 [17] a willful default is considered to occur in any of the following four cases:

a. When there is a default in repayment obligations by the borrower unit to the financial institutions even when it has the capacity to honor the said obligations. There is deliberate intention of not repaying the loan.
b. The funds are not utilized for the specific purpose intended for which finance was availed but have been diverted for other purposes.
c. When the funds have been tapped off and not been utilized for the purpose for which it was availed. Further, no assets are available which justify the usage of funds.
d. Asset bought by the lenders' funds have been sold off without the knowledge of the lender.

Also in cases where a letter of comfort or guarantees are furnished by group companies of willfully defaulting units, these obligations are not honored when they are invoked by the lender, then such group companies are also considered to be willful defaulters.

RBI suggests in its document on data standardization [3, 6] that, there is a data requirement for proper supervision. The data is broadly divided into two groups 1) Data submitted by banks 2) Data generated or compiled by the supervisor. Furthermore, data can also have other characteristics which need to be considered. Table 1 shows these considerations suggested by RBI. Considering the data standardization requirement of RBI and increasing concern of loan frauds committed, the objective of the study emphasizes to define the set of parameters which help to detect willful default behavior and build a data model.

In order to understand the usefulness of the required parameters, brainstorming session was arranged with bank experts, company officials, loan supervisors and financial brokers. The discussion happened on scenarios with respect to banks, companies who are taking loans, and other financial scenarios. This session was extremely useful to obtain initial broad set of parameters to begin the process, These are shown in Figure 3.

After identifying initial level of parameters as per brainstorming session, further many parameters are identified by learning case studies, the literature survey and discussion with the domain expert. Based on all these inputs and studies, the identified parameters as early indicators are grouped into six groups as shown in Figure 4.

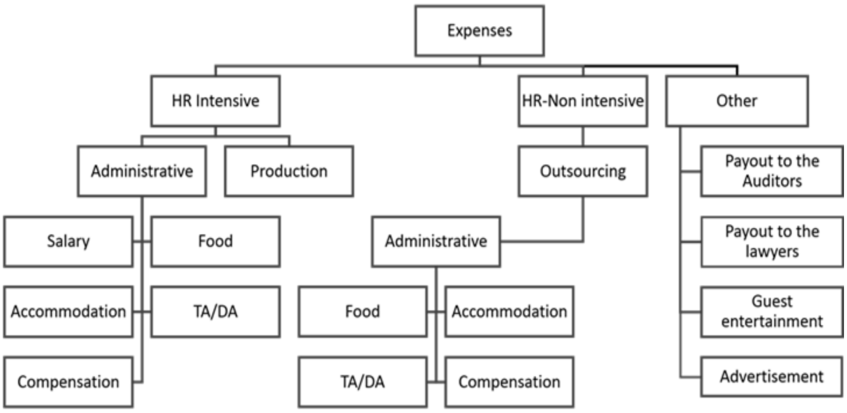

Figure 3. Initial Brainstorming parameters

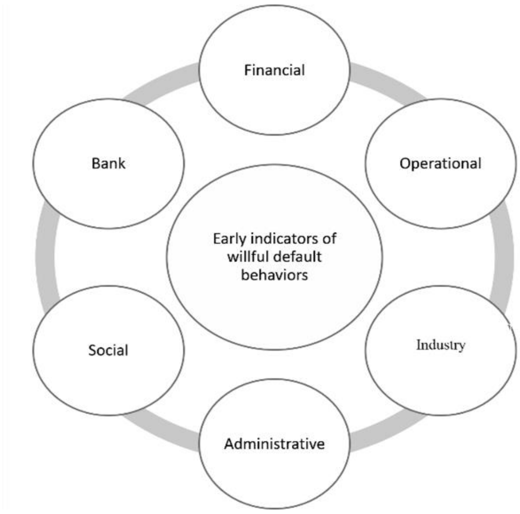

Figure 4. Broad groups of early indicators

Table 1. Guidelines suggested by RBI for supervision

Data Type		Submitted by Bank	Generated/Compiled by Supervisor
Structured	Numerical/ Financial	1) DSB Returns(XBRL)	7) Standard Annexes as part of onsite inspection
		2) Fraud Returns	8) Assessment of key financial/capital including validation/re-assessment of RBS riskk data furnished by bank
		3) FID Returns	9)Sores for aggregations of various risks as part of IRISc model
		4) RBS Risk Data (Data Collector Application)	10) Thematic/Sector/Industry/other bank-wide studies
		5) Financial Conglomerate Return (Excel)	
		6) Ad hoc data (Data Collector Application)	
	Textual	11) RBS Control gap information(Data Collector Application)	13) Comments/additional information on Control gap and Compliance by SSM
		12)RBS Compliance information (Data Collector Application)	14) Comments by Quality Assurance Division
Unstructured		15) Annual Reports	19) Working documents for supervisory assessment
		16) Policy Documents	20) Supervisory Reports
		17) Board Minutes	21) BFS Reports
		18) Reports of External Auditors	22) Communications to Banks

Source: Report to the Committee on Data and Information Management in the Reserve Bank of India (2014), RBI

The attributes under each of the group are defined as:

1. Financial

a. Financial leverage ratios :

i. Leverage ratios indicate fixed expenses obligations. Since the fixed expenses are period cost, it should be recovered from the period in which it is incurred. Worsening leverage ratio indicates that the company is not in the position to recover its fixed obligations.
   1) Asset Coverage Ratio: It indicates total backup of assets for each rupee of loan raised. If it is more than 1 then company can manage to repay its long term loans with existing assets.
   2) Debt Equity Ratio: This indicates outsiders' contribution to capital compared to owners' contribution. Ideal ratio is 1:1 but standard is fixed based on the gestation period and sector.
   3) Debt Service Coverage Ratio: This is calculated based on interest payment and interim period of 2 intervals. If the ratio is more, loan term should be less, if more years are given then it is a suspicion
   4) Debt/EBITDA Ratio : Debt is compared with Earnings Before Interest Tax and Depreciation Asset(ETBDA), which indicates the burden of the debt on profit
   5) Fixed Assets to Net Worth: This indicates to what extent fixed asset is financed by owners' contribution.
   6) Interest Coverage Ratio (ICR) : Interest is compared with Earnings Before Interest Tax and Depreciation, which indicates the burden of the interest on profit, how many times profit is sufficient to cover the interest
   7) Long Term Debt to Capitalization Ratio: It indicates that borrowed fund in the financial structure is less compared to owners' fund. It should be less than 1.
   8) Current assets current liability: This ratio indicates short term liquidity. It indicates the quality of working capital. For a manufacturing sector higher working capital is essential compared to service sectors
   9) Total Expense Ratio (TER) (Total expense/ Turn over): This indicates the proportion of the cost in revenue. Lower ratio is better indicator of profit margin.

ii. Interest to sales ratio: This indicates the proportion of debt cost to the revenue earned. Lower ratio is a better indication of profitability
   a. Credit rating agency: Credit ratings are given whenever there is new issue of securities apart from the company as a whole. Example CRISIL score etc.
   b. Write offs: It indicates poor collection policy and hamper the profitability.
   c. Current liability to fixed assets : Higher ratio indicates higher risk [Short term fund for long term projects]

2. Operational: These attributes indicate operational aspects of a company.
   a. Delay in payments to suppliers
   b. Delay in payments from the customers
   c. Losing customers
   d. Sudden changes in the suppliers, buyers
   e. Frequent changes in the business model
3. Administrative
   a. Diversion of funds: The loan amount is being used for purpose other than for which loan was sanctioned or the amount is diverted for personal gain.
   b. Lack of cooperation from the key personnel: If the key personnel is avoiding the discussion with financial institution or has negative outlook.
   c. Changes in administrators frequently: This indicates the problem with the company if there is frequent change in the administrative positions.
4. Industry
   a. Decline in the business growth
   b. Change in industry regulations affect the profitability of the company. It can be monitored by analysis of news articles
   c. Increase in cost of the raw materials
   d. Emerging markets, competitive company performance also affect the profitability of the company.
   e. Change in customer behavior with respect to segment: It can observed that if a customer segment changes then it has adverse effect on the company. It can be studied from broker analysis of Annual reports.
5. Social
   a. Social behavior and life style of the company officials
   b. Investment pattern of the company
   c. Expenses related to travel and other requirements

   d. Key personal outlook, director of the company is responsible for loan process which is taken for business/corporate requirements
   e. Social behavior of the directors of management people can be obtained through social media posts

6. Bank: There are several parameters which Banks maintain for each loan, some of them are listed below.
   a. Purpose of the loan
   b. Past loan status
   c. Annual Income
   d. Grade of the loan
   e. Credit score
   f. delinqencies
   g. delay in payments

Apart from these parameters, Companies Auditor’s Report Order (CARO) can be considered as the master document to analyse the parameters mentioned in. For instance if the CARO report says the assets are not validated, then it is a negative indicator. Then asset ratios need not be considered even if they look good. Companies Act, 2003 requires that the auditor’s report of specified companies should include a statement on the prescribed matters. These reporting requirements have been prescribed under the Companies (Auditor’s Report) Order, 2016. CARO report has information on Fixed Asset, Inventory, Loan given by Company, Loan to director and investment by the company, Deposits, Cost Records, Statutory Dues, Repayment of Loan, Utilization of IPO and further public offer, Reporting of Fraud, Approval of managerial remuneration, Nidhi Company, Related Party Transaction, Private Placement of Preferential Issues, Non Cash Transaction, Register under RBI Act 1934. This information also need to be considered as potential parameters for identification of NPAs based on pre-loan and post-loan performance analysis of the same. The pre-loan and post-loan performance analysis of the parameters mentioned above need to be done to understand the pattern of performance. If post-loan performance declined as compared to pre-loan performance then it is a negative indicator. However, continuous monitoring of the loan is required for early detection of the willful default behavior.

Considering all the above broad groups, the parameterization process is carried out to identify effective parameters for data model. For each parameter suitable data type and range or value indicating good loan are identified. The parameters, data type and values are shown in Table 2. The parameterization process followed to identify the parameters is unique and effective as all aspects and scenarios related to loan process have been taken into consideration while defining the final list of parameters. Hence, the process is highly feasible to implement with the help of Information and Communication Technologies (ICT). As the number of banks and companies are increasing and also number of loans increasing, the data capturing and analysis process for all these parameters is not able to be implemented using traditional ICT. Further, paper describes the big data based novel framework designed for loan process and data analysis.

4. FRAMEWORK FOR NPA/WILLFUL DEFAULT IDENTIFICATION

The parameterization process followed to identify the parameters is unique and effective as all aspects and scenarios related to loan process have been taken into consideration while defining the final list of parameters. Hence, the process is highly feasible to implement with the help of Information and Communication Technologies (ICT). As the number of banks and companies are increasing and also number of loans increasing, the data capturing and analysis process for all these parameters is not able to be implemented using traditional ICT. Further, paper describes the big data based novel framework designed for loan process and data analysis.

A novel framework for NPA/Willful default identification is designed and is represented in Figure 5. This framework mainly provides technical solution to handle the complete loan process staring from sanctioning to early identification of the willful default. For this process all the parameters required for early detection of NPA/willful default are identified through data parameterization process. These parameters need to be collected at the loan approval level and then continuous monitoring has to be done until loan is completed. During monitoring the pattern of loan payment, transactions carried out, behavioral and social traits are analyzed and if the pattern is not normal it is identified as outlier behavior and hence possible default case. This process is carried out longitudinally until the loan is fully paid or declared as NPA.

According to E\&Y survey [18] early warning signs to identify defaults must leverage technology and data analytical capabilities. Only technology can bring revolutionary shift in NPA management in India. Assistance of Automated solutions in data analysis can enable early indicators that will generate alerts before the situation becomes worse.

Table 2. Parameters identified for the data model

Sl no.	Feature Name	Values	Ideal Values
1	Asset Coverage Ratio	ratio	>1
2	Debt Equity Ratio	ratio	1:01
3	Debt Service Coverage Ratio	ratio	I.5
4	Debt/EBITDA Ratio	ratio	<1
5	Fixed Assets to Net Worth	ratio	<1
6	Interest Coverage Ratio (ICR)	ratio	<1
7	Long Term Debt to Total Asset Ratio	ratio	<1
8	Current assets current liability	ratio	>1
9	Total Expense Ratio (TER)	ratio	<1
10	Interest to sales ratio	ratio	<1
11	Credit rating agency	Ranking	Positive growth
12	Write offs	Frequncy	Frequncy
13	Current liability to fixed assets	ratio	<1
14	Creditors velocity	ratio	1
15	Stock velocity	ratio	1
16	Debtors velocity	ratio	1
17	Loss of sales	categorical	low
18	Supplier’s loyalty	categorical	high
19	Customers loyalty	categorical	high
20	businessModel	categorical	No
21	Diversion of funds	Boolean value(Yes/No)	No
22	Outlook of KMP	Boolean value(Good/Bab)	Good
23	Administrator Turnover	Boolean value(More/less)	Less
24	AR:Inventory Valuation	Boolean (Done/Doubtful)	Done
25	AR:Loan sactioned	Boolean(Good/Bad)	Good
26	AR:Statutory Dues	Boolean(Yes/No)	No
27	AR:Repayment of Loan	Boolean(Yes/No)	Yes
28	AR:Managerial remuneration	Boolean(Yes/No)	No
29	AR:Private Placement of Preferential Issues	Boolean(Yes/No)	No
30	AR:Non Cash Transaction	Boolean(Yes/No)	No
31	SL:Purchase history	Boolean(High,moderate, low)	Moderate/low
32	SL:Investment:	Boolean(High,moderate, low)	Moderate/low
33	SL:Social Life	Boolean(High,moderate, low)	Moderate/low
34	SL:Travel	Boolean(High,moderate, low)	Moderate/low
35	SL:Apparels	Boolean(High,moderate, low)	Moderate/low
36	SL:Social/reference groups	Boolean(High,moderate, low)	Moderate/low

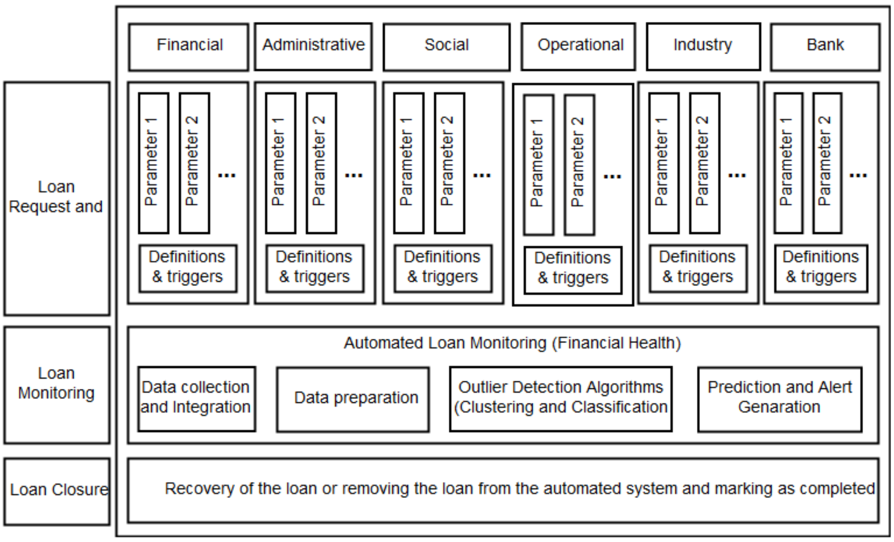
Financial
Administrative
Social
Operational
Industry
Bank
Loan Request and
Parameter 1
Parameter 2
...
Definitions & triggers
Loan Monitoring
Automated Loan Monitoring (Financial Health)
Data collection and Integration
Data preparation
Outlier Detection Algorithms (Clustering and Classification
Prediction and Alert Genaration
Loan Closure
Recovery of the loan or removing the loan from the automated system and marking as completed

Figure 5. Framework for NPA/Willful default identification

Big data technology is suitable to deal with data that is not only structured but in any format, in real time. The early detection of willful default needs analysis of unstructured data and generate alert. Hence utilization of Big data technology is essential. Such early warning system with Big Data capability help in identifying stress in banks and improve loan management life cycle. Further Big Data can be leveraged in loan underwriting decision making and NPA management.

Classification algorithms are required to build prediction model for NPA/willful default. Hence prediction algorithms are implemented using machine learning utilizing various structured and unstructured parameters. These machine learning prediction models are designed using map reduce logic on hadoop big data platform [19-21]. The classification algorithms considered are Naive Bayes [22], Logistic Regression [23], Support Vector Machine [24], Neural Network [25] and Random forest [26]. These are implemented using Map-Reduce technique of Big Data on Hadoop Cluster. The models are compared based on accuracy obtained and the algorithm with best accuracy is considered for prediction.

5. EVALUATION OF PREDICTION MODELS

The evaluation of the models is done considering structured and unstructured data. Structred data fields include Loan ID, Customer ID, Current Loan Amount, Term Credit Score, Annual Income, Years in current job, Home Ownership, Purpose, Monthly Debt, Years of Credit History, Months since last delinquent, Number of Open Accounts, Number of Credit Problems, Current Credit Balance, Maximum Open Credit, Bankruptcies, Tax Liens etc.The dataset comprises of around two lacs of rows. Unstructured data considered includes synthesized social media data. Sentiment analysis using Apache Hive is done on this data to get social outlook value [27, 28]. If this value is positive it indicates positive life style. Payment data is also considered to get spending patterns values. These values are added to the data. Spending pattern and Social outlook are the parameters from Table 2 and are synthesized for the purpose of validation. The aim of the model is Loan default prediction. For this purpose prediction models are built using machine learning on Hadoop and spark. Multiple machine learning models are implemented. These models are evaluated based on accuracy obtained. Machine learning algorithms considered for building prediction models are Logistic regression, Neural Network, Random Forest, and Naive Bayes. The results obtained for the model are shown in the Figure 6. As depicted in the figure Neural network has the highest accuracy, hence is used in the process of prediction of NPA and there by willful default.

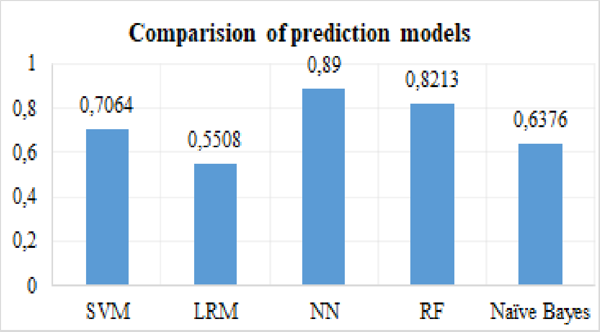

Figure 6. Evaluation of classification algorithms for prediction

6. CONCLUSION

Banking is the major service sector to balance the economy of the country. The loans going bad intentionally are not only affecting the bank’s profitability but also causing setback for the economy of the country as a whole. The technological assessment and support for early identification of such willful default is the need of the hour. It is imperative that customers’ entire profile including behavioral, financial, social parameters have to be considered and monitored. In this paper a process for identification of critical parameters is designed for early identification of willful default. This parameterization process needs to be integrated into the process of loan. Hence a novel framework which takes in to account starting from loan sanctioning till completion is designed. The framework is built using big data technology as it needs to deal with both structured and unstructured parameters. In order to choose the best prediction model in the framework an experiment is conducted. It is carried out on the loan data set which is structured and the generated synthetic unstructured data. Various classification models are built using map reduce and compared based on the accuracy. The results show that neural network has the best performance, and hence it is implemented in the framework. The results also indicate that in order to identify willful default unstructured components play a major role.

REFERENCES

[1] PWC Survey, “Current fraud trends in the financial sector”. http://www.pwc.in/assets/pdfs/publications/2015/current-fraud-trends-in-the-financial-sector.pdf, 2015.

[2] Deloitte Survey, “India banking fraud survey edition II,” https://www2.deloitte.com/content/in/Documents/finance/in-fa-banking-fraud-survey-noexp.pdf, 2015.

[3] RBI Report of committee on data standardization. https://rbi.org.in/scripts/PublicationReportDetails.aspx, available online: March 2015.

[4] Chetan, R., “The vicious cycle of NPAs,” http://wtdnews.com/the-vicious-cycle-of-npas/, Accessed on:24 May, 2018.

[5] Charan, S., Deepanshu, P., Divyesh, D., Kiran, A., Mohit, A., Ravi, K., S, M., Siddharth, N., Suryaansh, M., Tamanna, S., Vipul, M., “Frauds in the Indian banking industry,” *IIM Bangalore Research Paper*, No. 505, pp. 1-24, 2016.

[6] RBI Reports, “Master circular on willful defaulters”. https://www.rbi.org.in/scripts, January 2015.

[7] Bardan, S., Mukherjee, V., “Willful default in developing country banking system: A theoretical exercise,” *Journal of Economic Development*, vol. 38(4), pages 101-121, December, 2013.

[8] Chetan, R., “India vs. the world: Is the npa problem that bad everywhere?!,” "http://wtdnews.com/india-vs-world-npa-problem-bad-everywhere/" , October, 2017.

[9] Sabnavis, M. “Indias npas and the global scenario,” https://www.thehindubusinessline.com/opinion/indias-npas-and-the-globalscenario/article24145872.ece, June 2018.

[10] Sabnavis, M., “Npa problem: India ranked 5th in bad loans in world Eu's 4 tumbling economies top list,” https://www.businesstoday.in/current/policy/npa-problem-indiaranking-bad-loans-economies-with-huge-npa-bank-recapitalisation/story/266898.html , December 2017.

[11] Charan, S., Jagvinder, B., “Stressed assets and banking in India,” *IIM Bangalore Research Paper*, No. 507, pp. 1-20, April 2016.

[12] Sinkey, J.F., Greenawalt, M.B. “Loan-loss experience and risk-taking behavior at large commercial banks,” *Journal of Financial Services Research*, Vol. 5, No.1, pp. 43-59, 1991.

[13] Van Lai, S, “An analysis of private loan guarantees,” *Journal of Financial Services Research*,. Vol. 6, No.3, pp.223-248, 1992.

[14] Glennon, D., Nigro, P., “An analysis of sba loan defaults by maturity structure,” *Journal of Financial Services Research*, Vol. 28, No.1, pp. 77-111, 2005.

[15] Kaul, J.B., Keenan, D.C., “Catastrophic default and credit risk for lending institutions,” *Journal of Financial Services Research*, Vol. 15, No. 2, pp. 87-102, 1999.

[16] Yuwono Abdillah, Suharjito, “Failure prediction of e-banking application system using Adaptive Neuro Fuzzy Inference System (ANFIS),” *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 9, No. 1, pp 667-675, February 2019.

[17] Reuters: “Unpublished rbi data shows fraud problems extend far beyond PNB,” https://www.thehindubusinessline.com/money-and-banking/unpublished-rbi-datashows-fraud-problems-extend-far-beyond-pnb/article22773169.ece, Accessed on: February 16, 2018.

[18] Survey, E., “Unmasking Indias NPA issues can the banking sector overcome this phase?” https://www.ey.com/Publication/vwLUAssets/ey-unmasking-indias-npaissues-can-the-banking-sector-overcome-this-phase/, 2018.

[19] Xu, J.J., Lu, Y., Chau, M., “P2P Lending Fraud Detection: A Big Data Approach,” *Springer International Publishing*, Cham, pp.71-81.2015.

[20] Bathla G, Aggarwal H, Rani R. “A Novel Approach for Clustering Big Data based on MapReduce”. *International Journal of Electrical & Computer Engineering*, Vol. 8, No.3. pp. 2088-8708, June 2018.

[21] Sachin Arun Thanekar, K. Subrahmanyam, A. B. Bagwan. “Big Data and MapReduce Challenges, Opportunities and Trends,” *International Journal of Electrical and Computer Engineering (IJECE)*, Vol. 6, No. 6, pp.2911-2919, December 2016.

[22] Zheng, S., “Naive Bayes Classifier: A MapReduce Approach,” North Dakota State University, 2014.

[23] Bell, J., “Machine Learning for Big Data: Hands-On for Developers and Technical Professionals,” Wiley, 2014.

[24] Catak, F. O., Balaban, M.E., “A mapreduce-based distributed svm algorithm for binary classification,” *Turkish Journal of Electrical Engineering & Computer Sciences*, Vol. 24, No. 3,pp. 863-873, 2016.

[25] Binhan, Z., Wang, W., Zhang, X, “Training backpropagation neural network in mapreduce”, *In: Proceedings of International Conference on Computer, Communications and Information Technology* (CCIT 2014), pp. 22-25, 2014.

[26] Rahman, Md Armanur, J. Hossen, C. Venkataseshaiah, Ck Ho, Kim Geok Tan, Aziza Sultana, M. Z. H. Jesmeen, and Ferdous Hossain. “A Survey of Machine Learning Techniques for Self-tuning Hadoop Performance,” *International Journal of Electrical and Computer Engineering*, Vol 8, no. 3, pp.1854-1862, June 2018.

[27] Attigeri, Girija V., Manohara Pai MM, Radhika M. Pai, and Aparna Nayak. "Stock market prediction: A big data approach,*" In TENCON 2015-2015 IEEE Region 10 Conference*, pp. 1-5. IEEE, 2015.

[28] Yassine Al Amrani, Mohamed Lazaar, Kamal Eddine El Kadiri, “A Novel Hybrid Classification Approach for Sentiment Analysis of Text Document,” *International Journal of Electrical and Computer Engineering,* Vol 8, no. 6, pp.4554-4567, December 2018.

BIOGRAPHIES OF AUTHORS

Girija Attigeri, She is currently Assistant Professor-Selection Grade in the department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India. She received B.E. and M.Tech. Degrees from Visvesvaraya Technological University, Karnataka, India. She has 12 years of experience in teaching and research.

Manohara Pai M. M., He is a Professor and Associate Director of Research and Consultancy at Manipal Institute of Technology (MIT), Manipal Academy of Higher Education (MAHE), Manipal. He has an experience of 26 years in research, academics and industry. He received his Ph.D. from University of Mysore at Karnataka, India. His research interests span big data analytics, wireless sensor networks, internet of things, cloud computing and intelligent transportation system. He has wide publications in reputed international conferences and journals. He has six patents to his name and has authored two books. He has supervised four Ph.D. and 80 plus Post Graduate students. He was visiting professor of ESIGELEC-IRSEEM at University of Rouen, France. He is the investigator for several projects funded by Government of India and Industries. He is IEEE Senior member and Chair of IEEE, Mangalore Sub Section.

Radhika M Pai, She is a Professor at Manipal Institute of Technology (MIT), Manipal Academy of Higher Education (MAHE), Manipal, in the Department of Information and Communication Technology. She has an experience in research, academics and industry for about 25 years. She received her Ph.D. from National Institute of Technology (NIT), Karnataka, India. Big data analytics, database systems, data mining and warehousing and operating system are her major research interests. She has wide publications in reputed international conferences and journals. She has received grants from Government of India.