

Credit Risk Analysis of Loans using Social media Information

P.P.Halkarnikar¹, H.P. Khandagale² and Amol Dhakne³

^{1,3}Dr. D. Y. Patil Institute of engineering, Management & Research, Pune, India.
Email: pp_halkarnikar@rediffmail.com

²Department of Technology, Shivaji University, Kolhapur, India.
Email: khandagalehp@gmail.com

Abstract—The core business of the banking sector is sanctioning loans to different individuals and industries. The credit risk analysis of these elements gives guarantee about regular repayment of loan. As a result, healthy business firms repay their loan regularly thereby increasing good return on investment to bank. It is possible to increase the accuracy of credit risk calculation using current technology like Big Data and different analytical tools. In our approach, along with traditional parameters like profit/loss, financial history, financial status of directors, cash flow, we also included non-formatted data like news and informal information for analysis. This information can be included as positive, negative and regular. This information can be collected using Big Data techniques from websites, news websites, government agencies and external agencies. This is used to construct the credit scoring models and to predict the borrower's creditworthiness and default risk. Looking at the uncertainty associated with judging the credit of borrower, it is necessary to add new tools and methods to get maximum correctness. Our approach to use Big Data analysis tools to input informal sources available on internet, will increase the accuracy of finding good borrower for banks.

Index Terms— Financial Analysis, Credit risk, Big Data, Data mining.

I. INTRODUCTION

The banking industry deals with capital flow and risk associated with it. The overall performance and profit of the bank depends upon the repayment of the loans distributed to different sectors. Bank distributes loans to individual and other businesses. An individual's credit can be calculated based on his income, tax paid, saving and assets. But for business firms it is a complicated process. Many banks are now using automated tools for risk calculation and credit determination. These tools take into consideration of profit/loss, sales history, and financial status of promoters, cash flow and other parameters. From these parameters, bank calculate the credit level of the firm. The bank's success greatly depends upon its decision of credit to firm. Banks are also exposed to different kinds of risk, but the most challenging risk is credit risk. The performance of loan contracts affects profitability and stability of a bank growth and development. The extent to which a borrower uses the credit facility efficiently will greatly impact the firm's repayment ability and performance, which in turn affects the lending institutions. Credit risk is the loss of bank's profit, since the customer does not adhere to his or her loan refund commitment. Financial institutions are facing the problem of loan proposals because of continuous changes in the business environment, credit regulations, marketing strategies and the competition in business

itself. The objective of credit scoring is to help credit providers to quantify and manage the financial risk involved in providing credit so that they can make better lending decisions quickly and more objectively. Various statistical and machine-learning techniques have been used to model company credit and bond ratings in the past. The present analysis depends upon the statistical figure gathered from various sources of business. Sanctioning of loans requires the use of huge and various data along with substantial processing time to process large number of variables. The development of Big Data technology has encouraged the researcher to add non statistical information as one of the parameters for credit risk analysis and default prediction. Today credit risk is concern by various stakeholders such as institutions, customers, regulatory bodies, depositors and investors. A lack of attention to changes in economic or other circumstances of business that can lead to a deterioration in the credit standing of a business will lead to bank's sustainability. Credit risk is interest topic for finance communities, researchers and banking sector.

The objective of this paper is to determine the utilization of Big Data techniques to develop a valid and useful mechanism for analyzing credit risk and estimating the allowance for credit default. The statistical data collected by bank is additionally supported by the data collected from social sites. This will help the loan distribution with less risk of failure. The loan applicant has been classified into two categories: good credit and bad credit. A good credit business is likely to repay the debt whereas a bad credit business is likely to default. An analysis of credit risk can provide indirectly an indication of whether bank's credit granting policies are proper.

II. RELATED WORK

Kwaku D. Kessey in his paper discussed the issue of increase in nonperforming assets in the banking sector of Ghana.[1] the author discusses the challenges of risk management in the changing scenario of technology and automation. The risk to the banking sector due to poor risk calculation using traditional methods, has resulted in many bad debts. The study is limited to Ghana but it reveals the changing evaluation process is need of time for banking sector. Proper portfolio design for credit risk analysis is highlighted in this paper. The use of the latest technology in risk calculation is underlined in this paper. The primary data required is taken from the bank and secondary data is collected from firms' website and annual reports. Trend analysis is applied for previous years to understand the company growth.

Hamid EslamiNosratabadi et al. used different data mining tools for credit risk analysis of loans.[2]. Different data mining techniques like KNN, decision tree and others are applied to loan parameters using Clementine data mining software. After applying data mining tools, loans are classified into three groups: bad, medium and good. A fuzzy expert system is proposed for better risk prediction. Authors propose the fuzzification of risk classes for analysis of loans. Asrin Karimin in [3] develops similar techniques.

Ghatge and Halkarnikar, in this paper have proposed automated analysis of bank parameters for finding credit risk in sanctioning loans to firms.[4]. They proposed artificial neural network technique for credit risk analysis. The ANN technique is advance data mining tool which has self-learning capability. The parameters selected for credit risk are automatically weighted according to the training set. The work in this paper is further extended to incorporate non formatted data collected from websites and social media for calculation of credit risk. Khaled Alzeaiden also proposes the artificial neural network approach for credit risk analysis in his article.[9].

Sudhakar M et al. propose the use of datamining in the field of banking for prediction of credit risk for loans. [6]. In this paper credit risk is calculated using famous Weka tool of data mining. The decision tree technique is used for classification and predication of customers. Two stages are implemented for analysis. CIBIL credit score maintained by external organization is considered for risk calculation. In first step tree is built and pruning is done. In second stage prediction of loan is done for risk stage. Loans are recommended for selection or rejection of proposal. Sudhamathy G applies similar approach in his paper for risk prediction.[7] The tool used for tree development is R. He has used this approach of developing tree structure to predict possible risk in sanctioning individual loans.[10].

Somayeh Moradi et al. proposed dynamic model for credit analysis which is monthly trained for prediction.[11]. They also added fluctuating politico-economic factors for development of dynamic model. These factors work along with financial parameters set by bank, making robust prediction possible. For model they have considered the Iranian bank system. Fuzzy rules are formed to take into consideration of all internal and external factors.

The statistical methods are considered in many papers we referred. [3][5][8]. These papers provided valuable parameters used for credit risk analysis. Looking to the present state of credit risk analysis, data mining techniques are becoming popular for analysis. Salihu, Armend et al. provided a comprehensive survey of different data mining algorithms used for credit analysis. [12] In this paper advantage of use of modern tools is

highlighted by author. The data generated by social; site is unformatted and is termed as Big Data. The BigData tools are currently used for analysis of data generated by social sites and interne for credit risk analysis. Wenshuai Wu described the advantage of using Big data analysis for credit risk calculation. [13] The complex methods and parameters involved in risk calculation always need advance tools to incorporate in this process. Author has focused on modern tools and future trends in this complex process of credit risk calculation. The volume of data is huge; hence independent research is carried out for the effect of social information on risk calculation in our proposed system. The social site's information and news provided are valuable information that can provide the financial state of firms and their business domain. If this information is utilised for the calculation of credit risk, a bank may avoid a possible loss to its assets.

III. DATA COLLECTION AND METHODOLOGY

A. Methods of Data Collection

Financial analysis of a potential borrower begins with an understanding of the firm, its business, sales figure, its key risks and success factors. The financial ratios are calculated from these values. Some qualitative variables are also derived from available data. For the development of our proposed model, data is collected from a nationalized bank with restricted area of commercial loans. These loans are disbursed to various firms in past five years. Their performance is known during this period. The decision of loan sanction is based on a perception of sanctioning persons. These authorities check the fact presented to them at the time of loan application. There is no other tool is used to collect the non- informal information about the business domain except the personal interview at the time of sanctioning of the loan. The data collected from the bank is as under:

- 1) History/Application of the Borrower.
- 2) Financial Statements/Balance sheet for 3 years before loan sanctioning.
- 3) Account Statements for each financial year.

B. Data Collection from Social sites

The data required for analysis is collected from other sites also. The following sources are used with a filter to specific firm name and period.

- CIBIL site.
- Web site of the firm.
- Facebook responses.
- Twitter responses.
- News site of financial newspapers.

All the information is filtered through Big data tools limited to a given firm and period. The data collected from these sources is non-structured type. So needs more complicated processing to bring it into a useful format. Information collected from these sources is processed using natural language processing and clustering algorithms for keywords. Based on keywords in contents of sites are graded as positive, negative and normal. The status and quantity of these grades will influence the final decision based on the weight assigned to these parameters. The final decision about the status of loan is calculated by adding these social site parameters with regular class decided by facts presented in the bank.

C. Probable Methods of Data Analysis

Normal parameters are derived by the bank from the previous balance sheet. The analysis of the balance sheet of the firm is done on the following risk factors:

D. Credit Scoring

Credit scoring is defined as a statistical method that is used to predict the probability that whether a loan applicant will default or become delinquent. This helps to determine the amount of credit that should be granted to a borrower. Credit scoring can also be defined as a systematic method for evaluating credit risk that provides a consistent analysis of the factors that have been determined to cause or affect the level of risk .The objective of credit scoring is to help credit providers to quantify and manage the financial risk involved in providing credit so that they can make better lending decisions quickly. [4] Credit scoring helps to increase consistency of the loan application process and allows the automation of the lending process. Based on the consumer's credit scores, the financial institutions are also able to determine the credit limits to be set for the business firms. For the calculation of credit score, the parameters discussed above are considered with different weights assigned by the bank. The pattern of loan repayment of previous or existing loans is also considered and called as behavioral

TABLE I. TRADITIONAL RISK FACTOR ANALYSIS

SR. NO.	RISK FACTORS	VARIABLES	DESCRIPTION
1	Leverage and Solvency Indicators	Capital	Funds raised by the firm/ borrower
		Net worth	Capital + reserves
		Debt Equity ratio	It is a proportion between firms total debt and total equity
2	Liquidity Indicators	Current Liabilities	Creditors, Loans to be repaid within one year, provisions of taxes and expenses.
		Current Assets	Cash in hand and bank balances, Inventory of the firm.
		Current Ratio	It measures the proportion of a party's current assets to its current liabilities and thus gives a measure of the short term liquidity of the firm.
3	Profitability Indicators	Sales	Sale of the Goods by firm
		Profit	Total sales minus Total expenses.
		Net Profit	Profit after depreciation of Building/machinery and furniture.
		Profit to sales ratio	It is the percentage of profit to total sales of the year.

scoring. The deductive credit scoring system awards points (weights) to particular relevant attributes of the credit parameters. The weightage value of attributes is aggregated to a total score. The relevant attributes and their weights are determined by the credit decision-makers based on their experiences. The cut-off for the score is considered by bank for rejection of the loan or determining risk associated with the credit sanctioned.

IV. PROPOSED SYSTEM

The proposed system is based on present parameters and parameters collected from internet. The data mining techniques are used to find the credit score using financial and internal parameters of firm, while the Big data tools are used to capture information from social sites and web sites. Web sites used by Big data tools are government sites and company web sites. The change in government rules or change in business patterns are detected which may affect the profitability of firm. The specific key words based on business of firm is selected to filter the information from social sites. The collected information is classified as positive, negative and normal. The accumulated score is again normalised by weight assign to them by bank. Both blocks give accumulated credit scores used to decide the firm's credit risk. The same system can be used for predicting credit default based on current parameters collected from both the stages. The details are shown in figure 1.

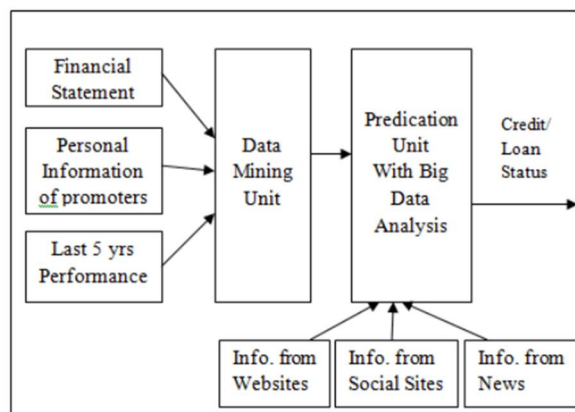


Figure 1. The architecture of Proposed System

Process flow diagram is shown in figure 2. Basic steps of data mining are not shown in the figure, but data cleansing and attribute pruning is necessary for proper application of data mining for prediction. Different

algorithms can be implemented here. We used decision tree technique for analysis of regular parameters. Similarly proper key words suitable to firms business are selected so that big data tools used for filtering data from social sites can give correct score. The Natural Language Toolkit is used for text tokenization of data collected from social site. Key words are separated and different clustering techniques are applied to classify the information into positive, negative and normal. These classes are provided different weights by bank depending upon the policies and processes adopted by bank. So total credit score is calculated from the two stages presented to decide on loans. Our system is capable of indicating the status of loan after it is also sanctioned. As it is continuously monitor the internet for information. Any change in domain policies, and company business news are picked up by the proposed system to predict the sanctioned loan condition in future.

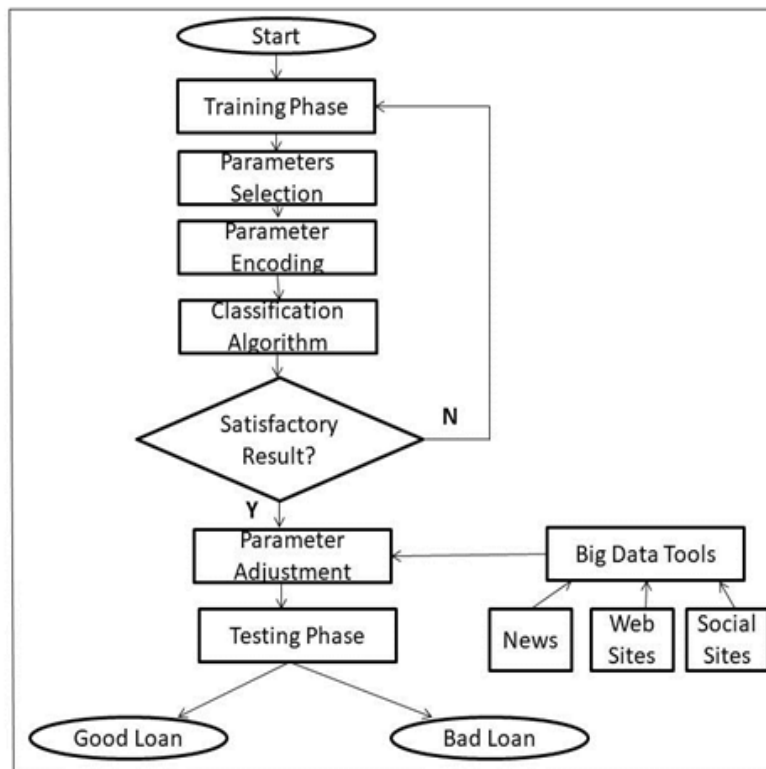


Figure 2. Flowchart of Proposed System

IV. THE IMPLEMENTATION AND RESULT

The dataset collected from the nationalized bank is considered as an input to the system. It contains 43 instances; with 30 cases of sanctioned loan proposals and 13 cases of rejected loan proposals. The database consist of attributes with each case is characterized by 13 decision attributes. Out of these, 10 were numerical and 3 categorical. From the given dataset 14 inputs are used as a training sample and remaining 29 are used as a testing samples.

For first step of data mining, decision tree classification C4.5, is used for classification of loan proposals based on their credit score. In the second stage social site information around the firms name is collected using HADOOP map reduce. The combined credit score is used to classify loan proposals in either good or bad credit. The results are compared with result using the existing system as shown in table 2. The proposed system shows more correct results as compared to the existing system.

TABLE II. COMPARISON OF RESULTS USING THE CONFUSION MATRIX

$\begin{bmatrix} 14 & 6 \\ 6 & 3 \end{bmatrix}$	$\begin{bmatrix} 18 & 2 \\ 4 & 5 \end{bmatrix}$
Existing System	Proposed System

V. CONCLUSIONS

Credit risk is major risk parameter in the banking sector. Wrong calculation of credit score leads to credit defaulters. Irregular repayment leads to poor health of bank. This affects the interest of stock holders of the financial sector. For a good and healthy economy of the country, it is essential to have a strong banking sector. It is expected that banks should give loans by proper analysis of the firm's financial health. In this paper we have proposed the system which relies on the present automatic system using data mining techniques and adding social site information. The system takes into account non informal and non-structured data for analysis. This data works as sensor for detecting the health of firm and its business sector. The bank can decide using this system the credit limit, interest rate and repayment capacity of the firm. The implementation using decision tree and Hadoop shows good results at the primary work carried out by us. The correctness of system can be judged by evaluating it further using different evaluation matrix. This system not only helps the bank during the sanctioning process of loans but also helps monitor the firm's capability to repay the loan during its loan period.

REFERENCES

- [1] Kwaku D. Kessey, "Assessing credit risk management practices in the banking industry of ghana: processes and challenges", *Global Journal of Management and Business Research*, Vol. 15 Issue 6 Version 1.0 pp. 201-212, 2015.
- [2] Hamid EslamiNosratabadi, SanazPourdarab and Ahmad Nadali, "A new approach for labeling the class of bank credit customers via classification method in data mining", *International Journal of Information and Education Technology*, Vol. 1, No. 2, pp 151- 156, June 2011.
- [3] Asrin KARIMI, "Evaluation of the Credit Risk with Statistical analysis", *International Journal of Academic Research in Accounting, Finance and Management Sciences*, Vol. 4, No.3, pp. 206–211, July 2014.
- [4] Ms. A. R. Ghatge, Mr. P. P. Halkarnikar, "Estimation of credit risk for business firms of nationalized bank by neural network approach", *International Journal of Electronics and Computer Science Engineering*, Vol. 2, No. 3, pp. 828-834, 2012.
- [5] Maubi Andrew Mokaya, Dr. Ambrose Jagongo, "Corporate loan portfolio diversification and credit risk management among commercial banks in kenya", *International Journal of Current Business and Social Sciences*, Vol.1, Issue 2, pp.81-111, 2014.
- [6] Sudhakar M, Dr. C. V. K Reddy, "Two step credit risk assesment model for retail bank loan applications using decision tree data mining technique", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, Vol. 5 Issue 3, pp. 705 – 718, March 2016.
- [7] Sudhamathy G., "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R", *International Journal of Engineering and Technology (IJET)*, Vol. 8 No. 5, pp. 1954 -1966, Oct-Nov 2016.
- [8] Naoyuki Yoshino and Farhad Taghizadeh-Hesary, "A comprehensive method for the credit risk assessment of small and medium-sized enterprises based on asian data", *ADB Working Paper Series*, December 2018.
- [9] Khaled Alzeaiden, "Credit risk management and business intelligence approach of the banking sector in Jordan", *Cogent Business & Management*, 6:1, 1675455, DOI: 10.1080/23311975.2019.1675455, 2019.
- [10] Anchal Goya, Ranpreet Kaur, "Loan Prediction Using Ensemble Technique", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 5, Issue 3, March 2016.
- [11] Somayeh Moradi, Farimah Mokhtab Rafiei, "A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks", *Financial Innovation*, 2019.
- [12] Salihu, Armend; Shehu, Visar, "A Review of Algorithms for Credit Risk Analysis", *Proceedings of the ENTRENOVA - ENTERprise REsearch InNOVation Conference*, , IRENET - Society for Advancing Innovation and Research in Economy, Zagreb, Vol. 6, pp. 134-146, 10-12 September 2020.
- [13] Wenshuai Wu, "Credit Risk Measurement, Decision Analysis, Transformation and Upgrading for Financial Big Data", *Hindawi Complexity*, Vol. 2022, Article ID 8942773, <https://doi.org/10.1155/2022/8942773>, 2022.