

Bigdata Technique Implementation in Bank Industry: Evidence from China

Yijie Zhao *

Department of Artificial Intelligence, Waterford Institute, Nanjing University of Information Science and Technology, Nanjing, China

* Corresponding Author Email: 202283910025@nuist.edu.cn

Abstract. With the continuous maturity and progress of cloud computing, intelligent tools, Internet of Things technology and other technologies, society has entered the era of big data, which is both a challenge and an opportunity for all traditional industries. As the pillar of national economy, the financial industry should comply with the trend of the time, comprehensively analyze the impact, use big data to analyze and mine effective information and reasonably apply it to the operation of banks. Finding the path of change is the key to foster the synchronous and long-term development of the banking industry in the new era. To achieve the aim of promoting the innovation and reform of the banking industry effectively, it's necessary to dig the answers from the examples of big data application. This study introduces the basic concepts and analysis methods of big data, shows the typical application of big data in Chinese banks, mines the convergence between banking and big data, analyzes the existing problems of big data finance and puts forward solutions. In addition, this research looks forward to the future intelligent and regular big data technique implementation in the financial industry. Overall, these results shed light on guiding further exploration of the big data implementation in financial industry.

Keywords: Financial technology; Big data analysis; Bank; China.

1. Introduction

Previously, scholars have pointed out the forward-looking view that the information storm brought by big data is gradually changing mankind's life, work and thinking. It can be said that big data has opened a major era transformation [1]. With the development of the Internet and digital life, advanced big data analysis has become the latest and highest application technology and tool for today's network and digitalization. Various search engines, network navigation and data statistics all rely on big data analysis. In addition, big data analysis has brought significant convenience to social production and people's daily life [1]. Over the preceding decade, the network and Web technology have developed at a rapid speed, the total amount of data created and copied in the world has exploded, the term big data has been born, and its hidden value has gradually emerged, which has attracted widespread attention and active discussion in the industry, research, and financial circles [2]. The assets of the banking industry account for 90% of the total assets of the financial industry. As the focus and pillar of the financial industry, the development of banking industry in the era of big data is related to the development of the whole financial market. Therefore, it's the proper time for traditional banking industry to use the business philosophy constructed by big data in order to carry out activities and make optimal business decisions [2].

In 2020, The State Council issued relevant plans to encourage commercial banks to actively use big data technology, promote online services, improve risk control capabilities, innovate business models, establish professional and intelligent financial service platforms, and complete digital transformation while developing their own advantages [3]. At present, big data technique is implemented in the following three aspects in commercial banks [2]:

Customer management. The data generated by social networks, terminal devices, e-commerce and other media is linked to the big data analysis platform of banks. After collecting, analyzing and screening the data, users can be classified and analyzed, so as to effectively classify and manage customers.

Marketing management. Through the big data dynamic targeting technology, customers can see the financial websites they have recently visited, the keywords they have viewed and the financial products they have purchased. According to the data, one can establish user models, push marketing service plans that fit customers' own personalized products, and directly meet customers' needs.

Risk management. After using big data to analyze the consumption behavior, credit and assets and liabilities of bank customers, a perfect risk prevention system can be built. The risk management ability of the banking industry will be improved as a whole and the social recognition of the financial industry's ability to control risks will be enhanced.

This paper will start from the introduction of the features and basic analytical methods of big data, take the implementation of big data in China's banking industry as the research object, penetratingly analyze the typical applications and existing problems of big data in the domestic banking industry, tap the potential of big data in the development of banking business. Subsequently, this study put forward relevant suggestions on the rational application and innovative development of big data according to the analysis content, promote the benign and sustainable digital transformation and development of the banking industry, and hope to provide some useful suggestions for future innovation and improvement of banks.

2. Basic Descriptions of Bigdata Techniques

2.1. Features

Big data refers to the collection of data which can hardly be captured, managed and processed by conventional software tools within a brief range of time. It is a kind of massive, high-growth and diversified information resources, which puts forward higher decision-making ability, insight, discovery and process optimality for the new processing model. It consists of huge data sets, which often exceed the amount of data that mankind can collect, vocabulary, manage and process in an acceptable amount of time. By 2012, big data's size had gone from one terabyte to tens of trillion bytes. In a survey and related report in 2001, DougLaney, an analyst at MetaGroup (now Gartner), divided the opportunities and opportunities for data growth into three major directions: quantity, speed, and diversification, which combined to be "3V" or "3V". Gartner and most of big data's industrial enterprises are using the word 3V to describe big data. In 2012, Gartner redefined big data as a new analytical method which is massive, high-speed, changeable and requires decision-making ability, insight and optimality." In 2008, IBM proposed the "5V" theory, which has been recognized by the whole related industry [4], as shown in Table 1.

Table 1. 5V's Features [4].

5Vs		Features				
VOLUME	Terabyte	Records/Arch	Tables,	Distributed	Linke	
VARIETY	Structured	Unstructured	Files	Probabilistic	d	
VALUE	Statistical	Events	Multi-	Hypothetical		
VELOCIT	Batch	Real/near-time	factor	Stream		Dynami
Y	Availabilit	Trustworthines	Correlation	Accountabilit		c
VERACIT	y	s	s	y	Origin	
Y			Processes			
			Authenticit			
			y			

2.2. Analytical Method

With the advent of the era of big data, extracting the key information from the mass data and bringing benefits to enterprises or individuals has become the focus of attention. Specific processing methods of big data mainly include [5]:

Hashing. Its essence is to transform the data into a shorter length of fixed length value or index value method. The superiority of the method is that it has a rapid speed of reading, writing and query language, but the inferiority is that it's hard to define a perfect Hash function.

Bloom Filter. It is essentially a bit array and a series of Hash functions. The principle of Bloom filter is to use a bit array to store the Hash of the data. The bitmap index, which is stored by lossy compression of data using Hash function, has the advantage of high space efficiency and query rate, but the inferiority is that it has a certain false recognition rate and is difficult to delete. Bloom filter is suitable for occasions where low error recognition rate is allowed.

Indexing. both traditional relationships in managing structured data. Index is an effective way to diminish the cost of disk reading and writing and improving the rate of adding, deleting, updating, and searching. The disadvantage of index is that it needs extra cost to store index files. Besides, it needs to be maintained dynamically according to the update of data.

Parallel computing. Compared with conventional serial computing, parallel computing refers to the simultaneous use of multiple computing resources to complete the operation. The basic concept is to separate the problems into several independent processors for the purpose of completing their own tasks, so as to achieve the goal of collaborative processing. At present, the typical parallel computing models include MPI, MapReduce, Dryad etc.

3. Implementation

3.1. Applications in Credit Default

For a long time, most countries have a mature system for credit research and analysis of bank users, and the credit evaluation model can be roughly divided into three categories. In the United States, the collection and evaluation of bank user credit data are market-oriented, while in European countries, the collection and evaluation of bank user data are basically led by the government, while in Japan, the collection and evaluation of user credit data are mainly carried out by bank associations and bank organizations. In our country, the bank user credit evaluation began roughly in the 1970s and 1980s with rapid development [6]. At present, it has mainly formed a pattern that the credit information center of the central bank takes the lead and conducts the collection and evaluation of user credit data together with private credit investigation institutions. The credit information system of the central bank covers part of the tax data of Chinese tax authorities and the payment data of many commercial banks, and has corresponding credit evaluation data and credit evaluation methods for Chinese enterprises and individuals [7]. In addition, many Internet institutions (e.g., Ant Financial and Jingdong Finance) also have their own credit evaluation system [6]. Credit risk is very common in the credit payment business of the financial industry, and it is also a key factor affecting the development of the financial industry. The research on the early credit payment business shows that in the scenario where the carrier is mainly credit cards, the borrower or the bank cannot perform the loan or repay on time due to internal reasons of the bank and the user itself [6].

This paper introduces the previous literature and research model on big data and credit risk scoring: using the unique internal data of online microfinance companies, this paper discusses the role of a new personal credit risk scoring system (based on online big data information, using machine learning and artificial intelligence to generate and update personal credit risk scores) in predicting borrower default risk [8]. The study randomly selected nearly 100,000 loan records from the company in 2018, including borrower demographic information, credit risk score information (PL score and fraud score), loan amount, term, purpose and repayment, default information, etc. In order to quantitatively evaluate the role of personal credit risk scoring system based on Internet big data in predicting personal online loan default risk, the following quantitative regression model (LPM) was established:

$$Default_i = \beta Score_i + \gamma Z_i + \delta X_i + \theta p + \mu_i \quad (1)$$

The dependent variable Default of this model indicates whether the borrower defaults on the due date (default on the overdue date =1). Since the degree of judging the borrower's dishonesty is whether

the borrower defaults on the overdue date, which is a relatively strict definition. In the extended regression analysis in the third part, the research defines "overdue" and "number of telephone collections required to settle a loan" to represent a loose measure of borrower default risk for alternative analysis. The core variable Score represents the borrower's credit rating score (PL score and fraud score). Z_i in the control variable represents the borrowing characteristics (borrowing amount, borrowing purpose, borrowing experience, etc., since the loan term and interest rate of this product are almost fixed (14 days and annual interest rate of 24%), the regression equation does not need to control for these two borrowing characteristics). X represents borrower characteristics (including age, gender, education background, income level, marital status, household registration and industry, etc.), θ_p represents province fixed effects (or city fixed effects), μ_i represents random interference items, which are clustered at the city level in this study.

Table 2. Baseline regression results [8].

Variable	Overdue default	Overdue default	Overdue default	Overdue default
PL Score	-0.00074*** (-0.00003)	-0.00062*** (0.00004)	-0.00051*** (0.00004)	-0.00057*** (0.00004)
Fraud score		0.00095*** (0.00008)	0.00110*** (0.00008) s	0.00120*** (0.00008)
Loan Amounts			-0.00003*** (0.00000)	-0.00003*** (0.00000)
Number of loans			-0.03517*** (0.00137)	-0.03455*** (0.00139)
Gender (Male =1)				0.01119*** (0.00300)
Age				-0.00382*** (0.00029)
Marriage (Married =1)				-0.00105 (0.00324)
Household Registration (rural =1)				0.00919*** (0.00322)
Income				0.00036 (0.00105)
Education				0.00305 (0.00190)
Observations	94559	94383	94383	89938
R-squared	0.00913	0.01086	0.02228	0.02601

Note: All regression standard errors were clustered at the city level, ***p<0.01, **p<0.05, *p<0.1.

Table 2 reports the estimation results of the benchmark regression equation. The first regression equation contains only the PL score of the company's independent R&D, which is the basis for the company to make loan decisions. The second column controls for the PL score and the credit risk score purchased from a third party -the Tianyu fraud score. The third column controls loan characteristics such as loan amount, loan frequency and loan purpose on the basis of including credit risk score. The fourth column further controls the borrower's demographic characteristics such as age, gender and household registration, as well as other information such as industry and income level.

Thus, the rese can reach the conclusions as follows. Primarily, the personal credit risk scoring system based on Internet big data can effectively predict the borrower's overdue default risk. The borrower with lower online credit rating (lower PL score, higher fraud score) has higher overdue default risk. In addition, only based on Internet credit risk score is not enough to predict the borrower's default risk. The characteristics of online loans (loan amount, purpose and number of loans) and the personal characteristics of borrowers (gender, age and household registration attributes, etc.) can help to predict the borrower's default risk more comprehensively.

Moreover, each bank has its own credit rating methodology. Among them, the personal credit evaluation mode of the central bank is an important financial institution that applies big data mining technology to financial institutions. The credit rating system is called China Scores. The index system includes seven types of credit rating index systems, and has been well applied in several large commercial banks. This scoring system combines the historical information of all personal credit accounts of major financial institutions in China, such as housing loans, auto loans and credit cards. The number of users has reached more than 60 million, and the data accumulation time has reached more than 3 years. Systematic research is carried out on the massive data of users, such as demographic characteristics, credit history, behavior record and transaction record, and the behavior pattern contained in it is discovered.

3.2. Applications in Combating Money Laundering

The early anti-money laundering analysis technology adopts a rule-based method to find abnormal transactions. On the basis of analyzing and summarizing previous money laundering cases, this technology finds out the regularity and refines it into behavioral rules. By presetting a threshold for transaction characteristics, the transaction behavior exceeding the threshold is judged as suspicious transaction, and then compares the capital transaction data one by one. Tracing money flows has long been used to determine whether money is being laundered. After the money laundering capital flow develops into the form of big data, the growth rate has become extremely fast, and the requirements for data processing speed are extremely high. As real data develops from quantitative change to qualitative change, traditional data screening methods and processing procedures often cannot be calculated within a foreseeable time when facing big data, and their effectiveness is greatly reduced [9]. Suspicious transaction clues in anti-money laundering are usually based on the combination of customer information and transaction information to form suspicious transactions [10]. The suspicious case screening process is shown in Fig. 1. Building a support system of big data analysis, application model and intelligent system can enhance the construction of intelligent management, so that data collection can have more accuracy, integrity and effectiveness. Meanwhile, through parameter control, improve the hit rate of monitoring model and realize intelligent control. With the support of big data, high-quality and efficient anti-money laundering work provides a strong guarantee for preventing and resolving money laundering risks and sanctioning compliance risks [11].

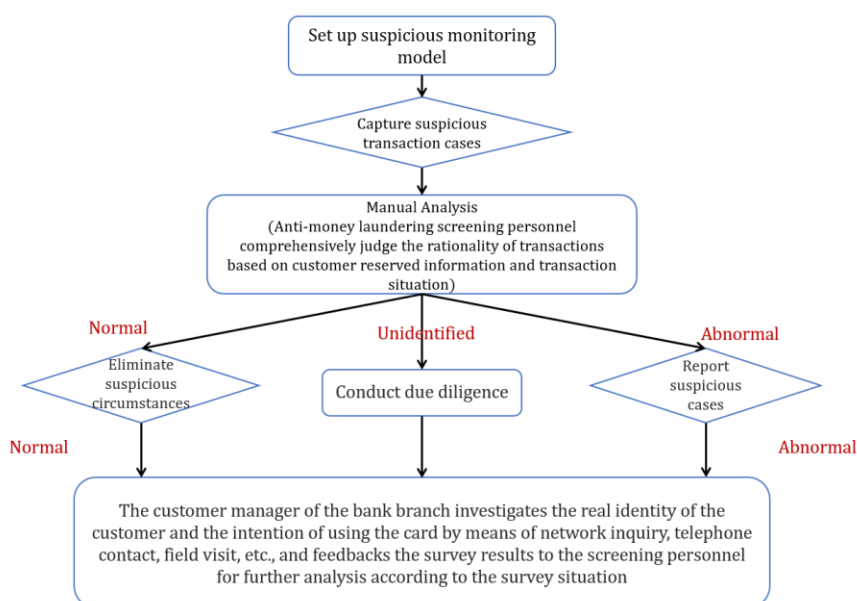


Fig 1. Suspicious case screening process [10]

For instance, Wuhan Branch of the People's Bank of China adopts Spark big data computing model for its entire anti-money laundering work to ensure its computing capacity for massive data(Figure 2). Distributed NoSQL database is used for data storage to cope with complex and changeable

business data. Its core work focuses on the following aspects: first, it designs a feature extraction model to extract the characteristics of capital activities of various money laundering crimes and portrait different accounts; The second is to design a gang identification model, find out different criminal gangs, and find out the relationship between the gangs; Third, in money laundering crime, different roles have different behavior characteristics. On the basis of detailed data, feature vector set and gang characteristics, a role recognition model is designed to identify the roles filled by different people (accounts); Fourth, analyze the channel of capital flow within and between gangs; Fifth, analyze and refine past cases to find out the characteristics and patterns of different types of crimes, and then abstract into the class crime recognition model [9].

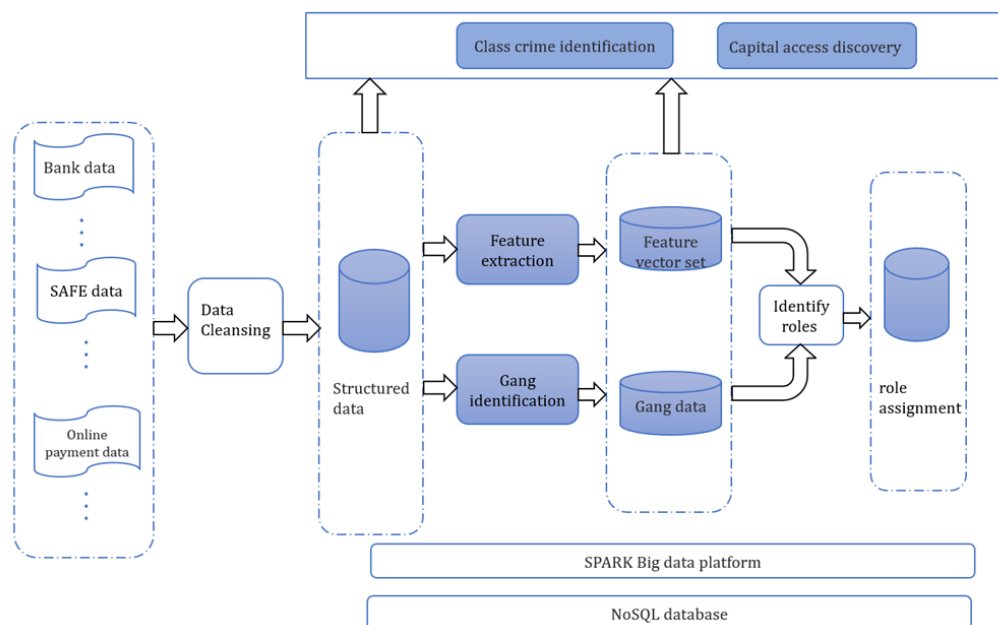


Fig 2. Anti-money laundering analysis framework based on big data [9].

The advantages of big data in anti-money laundering suspicious funds monitoring are as follows. First, it helps financial institutions to understand customer identities more comprehensively. Second, it is helpful for financial institutions to effectively analyze customer transaction behavior. Third, it is beneficial to promote financial institutions to improve the efficiency of anti-money laundering. Forth, it helps to improve the customer risk management level of financial institutions.

3.3. Other Applications

Big data technology has been widely used in financial industry and plays an significant role in the innovation and development of the financial industry. Big data technology can quickly and accurately analyze financial data, sort out, analyze and evaluate customer information, and change the problem of single, one-sided and limited data in the traditional financial industry. In the development of financial marketing, big data technology is used to analyze user data and business data, understand user purchase intention and product preference, carry out accurate marketing, and make accurate judgment. Big data can be integrated with financial product innovation. As user classification becomes more and more refined and user needs become more and more diversified, financial products also need to be constantly innovated and enriched. Big data technology is used to understand users' consumption and develop financial products that meet market demand [12]. For example, Ali Finance, the most representative of platform finance, uses e-commerce platform and online payment, and adopts cloud computing technology to process the financial data of enterprises and individuals quickly and accurately. Jd.com's supply chain finance also applies big data technology to provide consumers with installment financial services based on their precise data [12].

Moreover, big data technology can also be applied to comprehensively gather information from different channels. Securities companies can use big data modeling and analysis to analyze many

factors affecting the trend of the stock market, establish quantitative analysis models, predict stock prices, and provide investors with consulting suggestions, so as to create intelligent investment of big data and differentiated competitive advantages [13]. Moreover, big data intelligent customer service system set up by each platform can classify and sort out the problems customers have asked, consulted and complained, identifying keywords and answering customers' questions, saving labor time and cost [13].

4. Limitations & Future outlooks

Nevertheless, the state-of-art applications of bigdata still have some shortcomings and defects. First thing is lack of compound talents or teams in high-tech financial industry. Big data fintech knowledge belongs to the frontier knowledge of the discipline, the theory is obscure and difficult to understand, and the disconnect between theory and practice is serious. Putting theory into practice is a major difficulty. Capable financial industry talents and teams are in short supply, and the application of science and technology develops slowly [13]. Hence, in the financial industry, the enterprise should establish a special team of human resource management personnel, and strengthen the ability to process the data of human resource management personnel. Employees in the financial industry should have sufficient understanding of industry dynamics, advanced technologies, and platform construction, and provide professional training in the financial industry and technical training in big data technology for employees, so as to improve talent capacity and technical level, and promote the integration of financial industry and big data technology [12].

In addition, laws and regulations are not maturely enacted. With the development of the financial industry, its competition is increasingly intensified, and financial data has become the main object of competition between major banks. Massive financial data is not only related to the commercial interests of banks, but also related to the personal privacy of banks' customers. Therefore, how to strengthen the information security between banks in the process of combining the financial industry and big data technology is a very critical issue. When using big data, it is needed not only to strengthen the protection of the Internet, but also to improve the laws and regulations related to it, so as to provide protection for its legal security to a certain extent. In information security, personal privacy information should be classified in accordance with relevant laws and regulations, and the corresponding protection measures and operation methods should be adopted. Meanwhile, the data supervision of financial institutions should also be strengthened. Regulatory authorities should ensure the security of users' personal privacy data in the financial field, and establish corresponding norms and authorization for effective management. To strengthen the supervision of financial data, it is necessary not only to set up user authorization, but also to build a system of financial data security, confidentiality and internal control, so as to strengthen the supervision and management of data [12].

Big data technology is still under development, and its application, reform and innovation in the financial industry has great prospects. It is believed that in the future, with the continuous development and popularization of big data platforms and enterprises, in-depth technical research, mature talent training, and gradual optimization of various existing or potential problems, the application of big database in the financial industry will be more in-depth and comprehensive, promoting and promoting the development of the banking industry. It is believed that the whole banking industry will continue to deepen reform and steadily innovate in regular transactions, marketing, user portraits, risk prevention and control, securities and other aspects driven by big data.

5. Conclusion

In summary, this paper introduces the basic concepts and characteristics of big data, analyzes the typical application of domestic existing big data financial technology in China's banking industry, and explores the future potential and prospects of big data in banking industry. With the growth of population and rapid technology development, the era of big data is coming and the data set in global

financial industry is growing geometrically, which puts forward higher requirements for the supervision and business of banks. The introduction of big data technology into financial industry can solve the problems of single data, limited scope and difficult to deal with in the past can ensure the bank to summarize and process huge transaction data efficiently, judge trends accurately, and come up with more comprehensive and appropriate solutions and decisions to facilitate the development and transformation of banking institutions. Nevertheless, this paper has certain limitations. Since big data has a wide range of implementations, variety, and great potential, the common examples selected by this paper have certain occasions of applications and cannot fully represent other situations. In addition, big data technology is not fully mature, its application still has several known but unresolved or unknown potential risks. Besides, it is difficult to predict the new application conditions caused by human intervention or future environmental changes. For further studies, researchers can strengthen the implementation of big data technique in a specific financial field, effectively avoid the risks existing in data application, analyze the current demand, and improve the value created by bigdata in the banking industry. Overall, these results offer a guideline for the integration of big data technique and fintech.

References

- [1] Hu Minzhong. The understanding of the characteristics of big data analysis. *Journal of dialectics of nature research*, 2018, 1: 112-117.
- [2] Chen Yiya. Research on Problems and Countermeasures of banking industry in the era of Big Data. *Times Finance*, 2016, 14: 101-102.
- [3] Long Jing, Deng Hanlin. Research on the application of Big Data in the intermediate business of commercial banks under the background of Fintech. *Time-honored Brand Marketing*, 2022, 10: 78-80.
- [4] Blog of CSDN, Retrieved from: <https://blog.csdn.net/arsaycode/article/details/70847184>.
- [5] Zhang Yin, Chen Min, Liao Xiaofei. Current situation and Prospect of Big Data Application. *Journal of Computer Research and Development*, 2013, S2: 216-233.
- [6] Hu Yajun. Bank credit default user prediction method based on machine learning research. *Huazhong university of science and technology*, 2021.
- [7] Wang Hao. Research on Credit Risk Evaluation of Personal Consumer Credit in Chinese commercial Banks. *Changchun University of Science and Technology*, 2011.
- [8] Li Yanwen, Jiang Wenhua, Wang Chute. Can Network Big Data credit risk Score Effectively Predict Credit Default Risk? *Economic problems*, 2021, 7: 70-77.
- [9] Nie Tao, Wan Jianghan. Application research of Big data in anti-money laundering case analysis. *Financial Technology Times*, 2022, 30(03): 56-61.
- [10] Li Jiao, Cai Ningwei. Application of big data in anti-money laundering suspicious transaction report. *Modern Finance Guide*, 2021, 12: 30-35.
- [11] Wang Ruxun. Discussion on the practice of anti-money laundering intelligent method of commercial banks. *Financial Technology Times*, 2023, 3: 30-34.
- [12] Wang Changyi. Deep integration of big data technology and financial industry. *Business Culture*, 2021, 16: 108-109.
- [13] Lu Yunli, Chai Wenjiao. Financial technology helps the development of financial industry. *Cooperative Economy and Technology*, 2022, 13: 54-55.