Big Data and Related Model Algorithms in Commercial Bank Credit Evaluation

Jiawei Zhang*

School of Finance, Southwestern University of Finance and Economics, Chengdu, China

*Corresponding author: 42004302@smail.swufe.edu.cn

Abstract. Contemporarily, with the development of big data and related machine learning models and algorithms, commercial banks no longer possess advantages in traditional risk control and need to introduce big data and its algorithmic models as a new means of risk evaluation. The introduction of big data and its model algorithms by some commercial banks and fintech companies has proven its outstanding effect in the field of risk control. Based on the evaluations of XGBoost model, OneClassSVM model and other models, the article analyses the application of related technologies in the fields of default prediction and public opinion identification. Besides, this study combines the analysis with the scenarios in which commercial banks can implement big data and model algorithms for risk evaluation. In this case, it provides a new feasible method for commercial banks' risk evaluation and control, and proves the possibility of the application of related technologies with great effect. These results play an important role in promoting the application of big data and its models for risk evaluation and control in commercial banks.

Keywords: Big data; commercial banks; risk evaluation; model algorithms.

1. Introduction

Since the 2008 financial crisis, the world banking industry has adopted the Basel III requirements as its own guidelines for risk management and liquidity control. It is a proposal to improve the regulatory capital framework based on Basel II to enhance the soundness of the banking sector concluded in 2014, imposing stricter requirements on the banking sector in terms of market discipline. Basel III imposes capital adequacy requirements on banks, requiring banks to increase their core capital adequacy ratio to 6%, core core capital adequacy ratio to 4.5% and maintain a minimum total capital adequacy ratio requirement of 8%. Meanwhile, it significantly increased the regulatory requirements for market risk by adopting the concept of stressed VaR, requiring regulatory capital for market risk to be the sum of three times the stressed VaR and three times the normal VaR, as well as requiring the measurement of the new wind direction of each trading account [1].

On this basis, commercial banks have had to use new technological means to achieve improved risk identification for users and within the bank in response to increasingly tightening regulatory measures and the impact of other internet financial companies. Based on historical data and the risk management process of commercial banks, commercial bank credit risk management can be considered as the process of identifying, measuring and assessing the existence of loan repayment process of delay or the possibility of default [2]. Typically, home mortgage loan is the main type of loans of commercial banks in countries including home mortgage loans, which are large and highly correlated with the country's economic development, having a profound and systemic impact on the financial system. However, in the actual credit risk measurement process of commercial banks, it is essentially impossible to effectively assess the likelihood of loan default and credit losses due to the difficulty of obtaining users' data, the inadequacy of users' credit data and the limitations of banks' internal risk identification and control tools [3]. Although banks can achieve coverage of recovery losses on some credit operations through the setting of expected losses, reducing expected losses can still significantly improve banks' profitability, reduce the likelihood of banks experiencing significant risks and significantly improve their stability. The ratio of non-performing bank loans to total loans, for example, has been empirically analysed and has been shown to be a key factor affecting bank failures [4].

Efficient and accurate credit risk management is critical to the stability and viability of banks. It is even more important for some commercial banks with high levels of expected risk to effectively measure their level and manage them through effective risk management tools. Therefore, for banks, the analysis of micro and comprehensive data based on various aspects of users, including but not limited to consumer data and credit data, is essential for effective bank credit risk management in the era of big data. Currently, big data analytics has been used in various areas of risk management and asset pricing for commercial banks or other related areas. According to research, the application of Big Data to the creditworthiness and repayment rates of home mortgages has been applied in several major countries (e.g., the United States and China). Under this circumstance, commercial banks use machine learning methods to effectively predict the repayment ability of their loan users by aggregating and comprehensively evaluating their income status, repayment status of other loan programs, FICO or other credit scores, in order to selectively and targeted loan disbursement, effectively improving loan recovery rates. Central banks in some countries have also used big data analytics to make comprehensive assessments of the loans granted by their domestic commercial banks, thereby achieving an effective assessment of the risk of their commercial banking systems [5]. In the consumer finance sector, big data technology is also widely used. The consumer finance sector is generally characterised by small disbursement amounts, flexible disbursement forms, difficulty in recovery, and difficulty in obtaining user data and portraits, which often have great limitations in determining the credit level of users and identifying the risk of a single loan. Nevertheless, based on the usage of big data, the traditional problem of difficult access to user data and difficulty in forming a user profile has been partially solved. Currently, some commercial banks are sharing data with multiple platforms and institutions to effectively assess the credit level of users and apply the results to the issuance of consumer finance loans, which has effectively contributed to an increase in the recovery rate and business volume of consumer finance business [6]. The application of big data in similar areas also reveals the promise of big data technology in bank loan pricing. A 2014 study was conducted based on millions of online sellers, and through big data and data analytics, sellers assessed the potential demand for products and matched prices with demand responses for different products. The same technology can be applied to banking credit business pricing, commercial banks based on big data to analyse the credit level of users, try to analyse the credit level and default rate of users. Based on the results of this effective pricing of different loans issued to different users, one can fully enhance the user's willingness to lend and repay, to achieve differentiated pricing, effectively reduce the default rate and loss rate, to enhance the bank's stability and profitability [7].

However, it is clear is that the problems with Big Data for risk assessment and management in commercial banks are the same as the problems that exist in the application of Big Data in other areas. Big Data relies on extensive data that is not easily accessible, data validation and privacy protection issues, and low levels of data modelling learning, which together are the problems that limit commercial banks from using Big Data for credit assessment and credit management [8]. In addition, there are still very much expandable operations of Big Data in commercial banking business development and credit assessment and management. For example, in terms of user identification for banks, the results of traditional commercial bank business management assessments show that not all customers choose to obtain financial support from commercial banks, even if they do need it in their lives, work or business ventures. The application of big data for commercial banks can help them to quickly identify the current financial situation of their users, accurately analyse the potential needs they currently face and provide them with timely and customised needs from the bank. Likewise, banks can use their assessment of users to identify those who are at significant potential risk, even if they are not currently obtaining business from the bank. In terms of home mortgage origination and default prediction, the huge volume of home mortgages makes him particularly important. Therefore, commercial banks can provide users with suitable loan solutions by identifying their risks and home purchase needs, and provide them with a solution to the financial pressure in home purchase, while selecting users with good credit to reduce their own expected losses. In the consumer finance sector, the application of big data has absolute superiority over traditional customer relationship-based and manual issuance systems in terms of improving audit efficiency and accurately identifying risks. In terms of internal risk control, banks can easily identify potential risks, update user profiles and identify impending defaults through big data analysis for loans and funds already granted in the past, both during and after lending, thus contributing significantly to the stability and profitability of banks [9]. Henece, this paper is intended to provide an overview of traditional commercial banks.

Based on the analysis of traditional commercial banks' measures and problems in credit assessment and credit management, this paper intends to discuss and study the application of big data methods and models. Besides, this study will combine them with commercial banks' credit assessment and management business to explore the development areas and specific applications of big data analysis methods, and promote their combination, so as to improve the efficiency of credit risk management and the profitability and stability of commercial banks through big data analysis.

2. Big Data & Algorithm Models Analysis

With the widespread use of big data technology and machine learning technology in the financial field, numerous models for pricing and risk discovery have been widely used. In the field of risk control, Random Forest, XGB models, Logistic Regression models are widely used and are used to produce user credit score cards, predict default rates and other uses. Moreover, some deep learning models such as RNN are also widely used. In the field of account anomaly detection, models such as OneClassSVM, Isolation Forest and Local Outlier Factor are used extensively to effectively monitor unusual account changes and user security verification. In other areas, for example, deep learning NLP models are used for public opinion analysis and financial event recognition, while fully demonstrating the wide application of big data in the financial sector [10].

2.1 The Field of Risk Control

The application of machine learning algorithms based on big data in the field of wind control mainly includes random forest, XGB model and logistic regression model, among which the random forest model and XGB model have been proved to have significantly higher prediction effect than other models. The process of the XGBoost model is shown in Fig. 1.

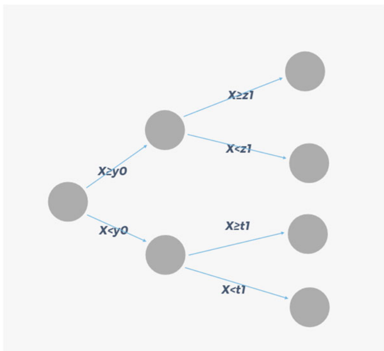

Fig. 1 A sketch of decision tree logic.

Regarding to the Random Forest algorithm, based on a model shaped like a tree, the data given by the user is classified and divided into two groups at key nodes according to a predetermined certain criterion, and thereafter each time a key node is encountered the data is divided into two groups, eventually until all the data meet the conditions, forming the so-called nodes at the bottom of the decision tree, similar to the formation of leaves at the top of a tree branch. If the user wishes to use the decision tree to classify a task, different criteria can be set so that the data branches out in different directions, eventually achieving the classification criteria the user wishes. However, an important problem with the follow-in forest model is that the model can suffer from overfitting, meaning that if it focuses too much on the features of the training set, even if the model predicts the test set very accurately, it may not perform as well with a new test set [11]. The solution to the overfitting problem lies in setting up enough trees and that the model should focus only on a subset of the test set, making the random forest an algorithm that integrates trees [12]. As for the XGBoost model, it is a tree-like scalable machine learning model that allows the user to add different other models (e.g., neural networks) to the XGboost model, and the combination of other models with the XGBoost model can significantly improve the overall accuracy of the prediction results. XGBoost uses regularisation techniques to significantly improve the prediction accuracy of traditional gradient trees, while the model uses shrinkage techniques to reduce the weight of the numbers, thus preventing overfitting problems [13].

2.2 The Field of Account Anomaly Detection

In the field of account anomaly detection, the applications mainly include OneClassSVM, Isolation Forest, Local Outlier Factor and other models, and the following will introduce the OneClassSVM model. the algorithm of the OneClassSVM model is fundamental, in with first to find a hyperplane, after that by the sample. By performing time-sharing, as depicted in the Fig. 2, a positive example circle in the sample is calculated, and eventually predictions and decisions are made through this positive example circle, and samples within the positive example circle are positive samples. However, as this calculation can be time consuming, the OneClassSVM model is not widely used nowadays when faced with large data [14].

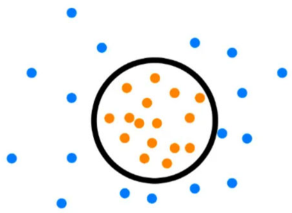

Fig. 2 Positive example circle of OneClassSVM.

2.3 Opinion Analysis & Financial Event Recognition

In the field of opinion analysis and financial event recognition, natural language processing models are mainly used for analysis. Language consists of sentences, which are made up of words that are stitched together, and a finite number of letters to build a finite number of words. Even though it is believed that a sentence can be infinite, sentences have an upper limit to their length in practice, so the research is limited to a finite set. Natural language processing is a range of computer techniques that capture human language and use theory to analyze it automatically.

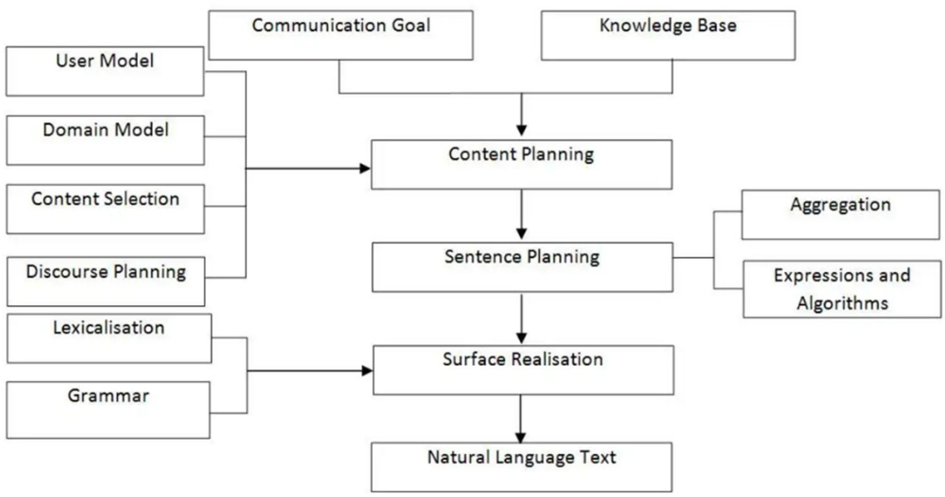

Fig. 3 Natural Language Processing Process

This technology is widely used today for online information retrieval and question and answer services, a service that relies on both algorithms for the representation of web text and low-level natural language processing techniques. However, such basic natural language processing can only perform basic word analysis, but not effective analysis when the content becomes a sentence. Current natural language processing methods are still based on the syntactic representation of sentences in text and rely heavily on the frequency of word occurrences, but such algorithmic models can only process the information in each text and cannot judge the context and process the information like a real human being. The technique has now been sufficiently improved to achieve increased precision and a significant increase in accuracy in language processing by analyzing sentence structure, sentence meaning and grammatical usage patterns, allowing it to be widely used. People create public opinion through social networks or other means, and financial events can also be revealed through words or descriptions that predict some of the emotional biases expressed by users. Hence, this technique can be used to carefully analyses public opinion and related texts to eventually discover public opinion and financial events hidden behind public opinion and complex information [15]. The process of the natural language processing is illustrated in Fig. 3 and Fig. 4 presents the application of the NLP.

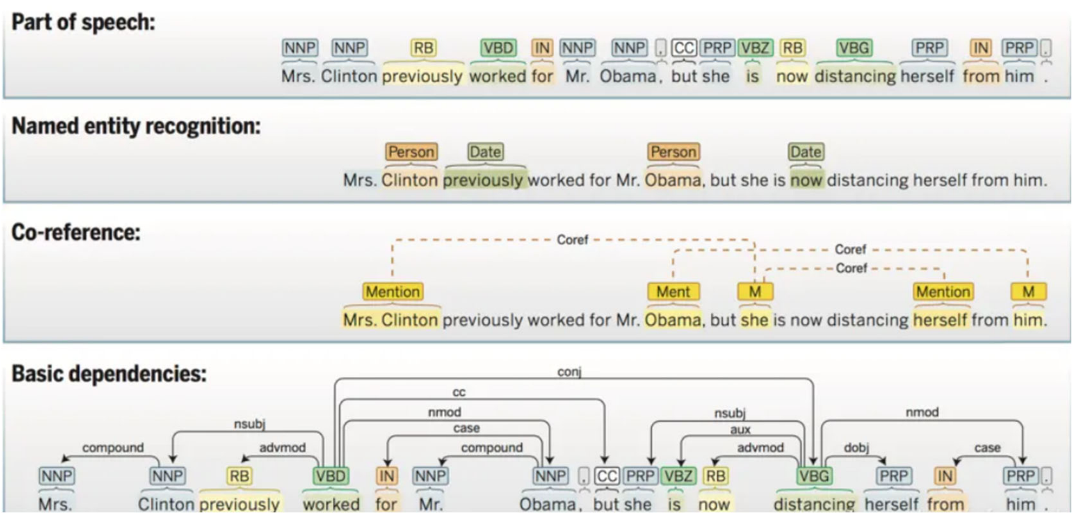

Fig. 4 Application of Natural Language Processing

3. Application Scenarios

In the development of the banking industry, machine learning algorithm models based on big data have been widely used in various businesses in the banking industry, including business risk management, customer risk identification, product development and other areas. As the main business of banks, the relevant applications in the field of lending is also very common, in the three major areas of pre-lending, lending and post-lending to improve efficiency and risk control level to make a significant boost. The following will elaborate on the specific applications of big data in the areas of consumer finance and lending in terms of identifying risky users, identifying potential users and internal risk control of banks.

3.1 Identifying Risky Users in Consumer Finance and Lending

Over the past few decades, the demand for credit in various countries has increased rapidly due to the rise in economic development. In the United States, the level of consumer credit has exploded over the past three decades, with total loans currently exceeding US$4 trillion and the level of household debt doubling. Against this backdrop, the use of big data technology in consumer credit has become necessary and, in fact, decision tools based on algorithmic results have been widely used in the consumer credit sector, achieving important developments in the identification of risky users and potential users.

In the area of identifying risky users, there is already research that demonstrates the important role of technology in risk control. By analysing data on users' credit cards and past loans, commercial banks identify risky users among them and eventually reduce the loan amount of the relevant users, effectively controlling the bank's overall risk exposure and thus reducing the probability of using bank capital to cover expected losses. They base their estimation on techniques such as decision trees and random forests, and analyse users' data over two quarters through a model that plots the risk ranking of users and the credit score of their accounts, classifying users who are 0-90 days overdue at intervals of every thirty days. The results show that while credit scores have been somewhat effective in distinguishing good from bad lenders, logistic regression and other machine learning models are very capable of predicting account defaults and risky account rankings, effectively identifying the level of account delinquency [16].

In order to test the validity of applying the XGBoost model into commercial banks risk management, this paper used the XGBoost model for training. Based on the previously mentioned risk measurement models and application areas, the XGBoost model is selected for default probability prediction in this paper. The paper selects user account data and housing information from within commercial banks in its tests, and the prediction value set is whether a customer will fully repay his or her home mortgage. The customer data used came from Xinwang Bank in Sichuan, China, an emerging Internet commercial bank that effectively collects many data in its business services, and it provided the paper with 12,488 user data for training the model and about 6,000 data for validation, with each data table containing more than 200 user features.

Although this amount of data is far from the amount of data that commercial banks really face to process, it is still very effective for proving the effectiveness of the XGBoost model. Priamrily, in terms of data processing, this paper removes the values with more than 90% missing proportion from the training set data and the test set data respectively, and also remove the user information with more missing data, and finally leave 10,000 data in the training set and 5,300 data in the test set by filtering, while only 170 features for each account are retained. It is also calculated the evidence weights and information ratios of character-based variables, and coarsely classified and merged the loan types and housing types of users with similar and logically reasonable values. Secondly, the paper was populated with the remaining data, and a mean fill approach was taken to ensure that there are no missing features in model training and prediction in this paper to prevent problems in model fitting. After setting the maximum learning step depth to 5 steps and other related metrics, the test result of this paper is 80.5% accuracy. As a comparison, the same model was fitted using logistic regression, and this paper processed the data in the same way and applied the trained model to make the same predictions, but unfortunately, the measured accuracy was also 76.2%. The accuracy of the prediction is summarized in Table 1. Although the prediction accuracy did not reach a high level, it is still evident that the large amount of data and machine learning algorithms still have unparalleled advantages in identifying user risks, etc. According to the results, this paper demonstrates the importance of relevant models for user risk identification.

Table 1. Model and accuracy of predictions for at-risk users.

Model	Prediction Accuracy
XGBoost	80.5%
Logistic Regression	76.2%

3.2 Potential User Identification

In the area of identifying potential users, commercial banks have used extensive data analysis and machine learning algorithms to identify users with potential loan needs and recommend loan products to them. The specific procedure is as follows. Firstly, it is necessary to analyse the financial status of the user based on the bank's internal accumulated data on user consumption, income and deposits. Secondly, one needs to analyse the current potential needs of the user based on the information collected from multiple scenarios of the user, e.g., social platform data and advertising browsing data.

If the user's browsing content indicates that he has a need to purchase a home but his income cannot support his needs, then this effectively achieves. This is very important for the development of commercial banks in the current situation of rapid development of Internet financial platforms and insufficient demand in the traditional commercial banking sector, which can more fully promote the market potential tapping, while bringing more profits to the banks.

3.3 Bank Internal Risk Control

Commercial banks have significantly increased their investment in financial technology, or set up their own fully or partially controlled ownership of financial technology companies to provide services to commercial banks. One of the important service areas is internal bank risk control. Some commercial banks have also chosen to partner with large Internet companies, which currently have more advantages in the area of user data and financial technology. Other options also include commercial banks choosing to partner with other fintech service providers to access relevant services. The important element is that commercial banks can accurately identify the risk of issuing loans before and during the loan process through analytical data based on extensive data. On this basis, expected losses can be set in a timely manner to cover possible future loan defaults; at the same time, based on the collection of data throughout the loan process, commercial banks no longer need to make judgments on whether to issue loans through empirical analysis as they did in the past, but can use the data of the whole loan process can be used to modify the bank's internal risk control model to make the model more optimised. Taking into account the unpredicted events that occur in the loan by the user, so that the future risk control model can be more accurate and the level of internal risk control can be improved. The principles of risk control for other bank businesses are the same as those for loans, in that they use massive amounts of data for analysis to identify the risk points of each business, effectively identify the risks and adopt relevant risk management tools [17].

4. Conclusion

Based on the content of the above analysis, it is believed that machine learning and other algorithms based on big data technology have a wide range of application areas in commercial banks, as they can have a huge impact in a number of areas such as pricing, risk discovery, opinion analysis, financial time identification and detection of unusual account movements. This paper have likewise based the analysis on the specific applications of XGBoost in the field of risk control and potential user identification and risk control. Big data-based algorithms, models and other technologies that may emerge in the future will fundamentally change the problems that exist in the development of traditional commercial banks, profoundly affecting the internal logic and business models of commercial bank operations in the coming period and significantly improving the quality and effectiveness of bank business management.

Apparently, this study has some shortcomings and drawbacks. Although Big Data and its algorithms have significant advantages for the business of commercial banks, not all commercial banks can afford the cost of developing Big Data business, as the collection, storage and analysis of data requires a lot of computing power, as well as a lot of technical talent, which means that what may have been handled in the past by operational staff according to experience now requires significant costs and the quality of business and profitability levels have not improved significantly. There is also the issue of data availability, i.e., not all users are willing to share their data with the bank, and not all users' data is sufficient. Finally there are the shortcomings of Big Data and its algorithms; the technology is not yet able to achieve accurate predictions, making Big Data somewhat less attractive. In the future, the application of big data and its algorithms in commercial banks will inevitably require its own development to enhance the accuracy and quality of forecasting through model development. At the same time, the relevant authorities need to develop mechanisms that both protect consumers and achieve effective data sharing, to identify the rights to the data and effectively manage its use. Last but not least, it is necessary to reduce the cost of using big data technology through technological development. Only when a large number of small commercial banks are able to only when a large number of small commercial banks have access to such technology can the overall banking system's capabilities be effectively enhanced. Nevertheless, these problems can be solved and this paper remain optimistic about the future of Big Data and its algorithms in commercial banking. Overall, these results provides guidance for commercial banks to improve their risk management capabilities through big data and other technologies, and points the way to improve the accuracy of model prediction and better predict users' special behaviors to accomplish risk management better afterwards.

References

[1] Pan Rui. Research on risk measurement and management of commercial banks in China based on the New Basel Accord. Dissertation of Shandong University, 2018.

[2] Al-Shakrchy E J. The impact of credit risk managing on bank profitability an empirical study during the pre-and post-subprime mortgage crisis: The case of Swedish commercial banks. Journal of business and finance, 2017, 3(1): 31-42.

[3] Magnifique U J. The effect of credit risk management on the financial performance of commercial banks in Rwanda. Unpublished doctoral dissertation), University of Nairobi, Kenya, 2013.

[4] Kumbirai M, Webb R. A financial ratio analysis of commercial bank performance in South Africa. African Review of Economics and Finance, 2010, 2(1): 30-53.

[5] Bholat D. Big data and central banks. Big Data & Society, 2015, 2(1): 2053951715579469.

[6] Jiang J, Liao L, Lu X, et al. Deciphering big data in consumer credit evaluation. Journal of Empirical Finance, 2021, 62: 28-45.

[7] Einav L, Levin J. Economics in the age of big data. Science, 2014, 346(6210): 1243089.

[8] Chen J, Tao Y, Wang H, et al. Big data based fraud risk management at Alibaba. The Journal of Finance and Data Science, 2015, 1(1): 1-10.

[9] Wang Y, Xiuping S, Zhang Q. Can fintech improve the efficiency of commercial banks?—An analysis based on big data. Research in international business and finance, 2021, 55: 101338.

[10] Mahesh B. Machine learning algorithms-a review. International Journal of Science and Research (IJSR), 2020, 9: 381-386

[11] Schonlau M, Zou R Y. The random forest algorithm for statistical learning. The Stata Journal, 2020, 20(1): 3-29.

[12] Breiman L. Random forests. Machine learning, 2001, 45(1): 5-32.

[13] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[14] Erfani S M, Rajasegarar S, Karunasekera S, et al. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recognition, 2016, 58: 121-134.

[15] Hirschberg J, Manning C D. Advances in natural language processing. Science, 2015, 349(6245): 261-266.

[16] Butaru F, Chen Q, Clark B, et al. Risk and risk management in the credit card industry. Journal of Banking & Finance, 2016, 72: 218-239.

[17] Wang R, Liu J, Luo H. Fintech development and bank risk taking in China. The European Journal of Finance, 2021, 27(4-5): 397-418.