Proceedings of the 7th International Conference on Economic Management and Green Development
DOI: 10.54254/2754-1169/31/20231545

A Review of Alternative Data in Credit Risk

Yuhan Huang[1,†], Zhixuan Li[2,†,*], Sizhe Pan[3,†], Xinwei Wen[4,†]

[1] Economics and Management School, Wuhan University, Wuhan 430000, China
[2] School of Economics and Management, Xidian University, Xi'an 710000, China
[3] Economics and Management School, Wuhan University, Wuhan 430000, China
[4] Accounting school, Guangwai University, Guangzhou 511400, China
a. lizhixuan@stu.xidian.edu.cn
**corresponding author*
[†]These authors contributed equally

Abstract: With the increasing number of loans given to people who previously have no credit history, the current credit risk prediction models trained by conventional data, namely credit history, social capital, etc., have become less effective. The aim of this essay is to figure out how alternative data is used in credit risk evaluation from those published essays. Based on EBSCO and ScienceDirect serving as the main database sources, we filter out 24 most relevant essays. We conclude that the alternative data can considerably optimize the verdict risk prediction models, as well as solve current financing conundrum when it comes to using alternative data to train the prediction models. However, the quality of alternative data and its ethical issues should require further investigations.

Keywords: alternative data, credit risk, credit scoring, unstructured data, bank risk

1. Introduction

According to Anderson, the term "credit risk" refers to the likelihood of a legally binding agreement becoming worthless or substantially less valuable as a result of the counterparty's default or insolvency [1]. Nearly $5 trillion goes missing due to occupational frauds, showed by the ACFE 2021 released data, which manifests the significance of predicting consumer credit risks in advance of lending money. Nevertheless, traditional data scoring models currently in use usually rely on models that draw on credit-bureau data or users' credit history to identify potentially fraudulent activity. Therefore, models that rely solely on traditional data may not be effective, under the circumstance that the number of digital micro-loans given to individuals who don't have credit history has increased significantly recently. For instance, in many African countries, it has been approximated that over 11 million loans were provided to clients via mobile devices [2]. In hence, the widespread use of mobile devices and internet connectivity has created an unprecedented source of detailed user behavior data, namely location information, shopping patterns, social media activity, app usage and the like. This type of data is referred to as "alternative data," and is distinct from traditional data sources such as credit bureaus, credit applications, or a lender's own records [3].

Traditional agencies often use data such as repayment history, arrears, and credit history to make credit assessments. Customers with scarce credit data are unable to obtain loans under this evaluation mechanism. Alternative credit scores use a digital footprint, such as mobile payment bills, to assess a potential customer's credit. This is very beneficial to both lenders and borrowers. Customers with scarce credit data are more likely to get loans, and banks can also use alternative credit scores to increase penetration in untapped areas (such as suburban and rural areas) while maintaining low risk and preventing fraud, truly achieving financial inclusion and rural revitalization. Alternative credit scores make full use of a customer's digital footprint (such as online payment history, phone bills, travel history, etc.) to complement traditional credit data and paint a more complete picture of the applicant.

There are many papers that have conducted empirical studies on how different kinds of alternative data are applied to credit risk assessment. We have sorted out and summarized these papers to illustrate the role that alternative data plays in credit risk evaluation. It is essential to determine what has been achieved and what future research will entail.

In this paper, we first illustrate the procedure of listing the articles. After thorough analysis of these papers, we demonstrate the methods that these papers utilize and conclude the advantages and disadvantages of alternative data. Next, based on the impacts that using alternative data in constructing the models brings, we recommend some propositions on future researches concerning this discipline.

2. Systematic Literature Review Methodology

2.1. What is Systematic Literature Review

An essential tool for fintech research is a systematic literature review, a scientific research approach that differs from the conventional review. It uses a clear, scientific, and reproducible process. By offering an audit process of the searcher's judgments, practices, and conclusions, it seeks to conduct a thorough literature search of published and unpublished studies and eliminate bias [4].

Compared with the traditional review, the research scope of systematic review is more focused on a specific field, and the research methods are more scientific and standardized. This paper takes the form of a systematic literature review to facilitate a more scientific exploration and summary of the research topic.

2.2. Search Strategy

The search strategy of this review is based on the principle of conforming to the theme and scientific research standards. The goal of the search strategy is to retrieve papers combining alternative data and data science means with credit default and credit risk.

For the literature search, we selected EBSCO and ScienceDirect as the main database sources for our literature search. It includes more than 7,000 journals, and there are more than 900 core journals included by SCI&SSCI, covering major disciplines and research fields. We used four stages to search the papers, which are topic screening, time screening, type screening, and relevance screening.

Step 1, based on the research purpose of this paper, after consulting relevant authoritative papers, we decide on the logical search string of literature according to our topic is shown as follows: "alternative data" OR "big data" OR "sentiment analysis" OR "data analysis" OR "text mining" OR "data science") AND ("credit scoring" OR "credit risk" OR "banking risk" OR "credit rating" OR "default risk".

The search statements were entered into the database for retrieval, and 307 relevant original articles that met the criteria were obtained after removing duplicate items. Figure 1 depicts the cumulative trend of the quantity of original papers over the previous 20 years, which demonstrates the yearly advancement of the most recent research in linked sectors.

Step 2, The fintech field is a very popular and emerging field in recent years with rapid development. Therefore, we limit our literature search to the literature in the past three years, that is, the literature from 2020 to 2022.

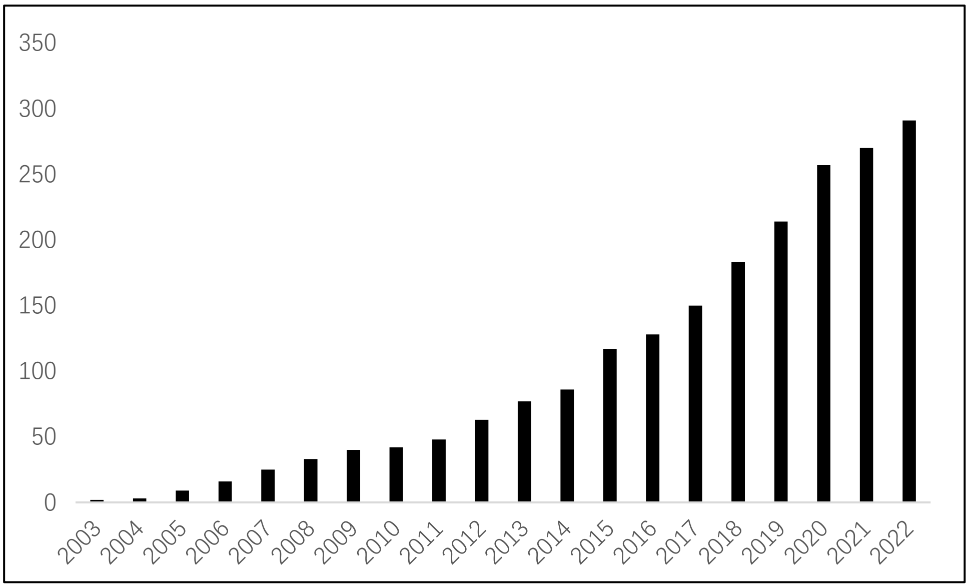

Figure 1: Accumulated quantity trend of literature.

Step 3, Due to the consideration of academic rigor and the nature of this paper, we only search for academic papers in authentic academic journals and only search relevant papers written in English. To make the data and information sources of this paper more authoritative.

Step 4, After the screening of the first three criteria, we read the title, keywords, abstract, and text of each literature respectively, and evaluated the relevance of each original paper with alternative data and credit risk topics one by one (Figure 2).

After obtaining the final 24 sample documents, we reviewed the sample documents against the research questions proposed in the first part and our research theme, extracted various research methods, research data, and research conclusions of alternative data and credit risk from them for integration, and conducted research on this topic according to the logic of this paper. The integration process includes updating and expanding theories or arguments proposed by past scholars, identifying complex concepts from different industries or market segments, and summarizing research conclusions from different articles for subsequent relevant analysis.

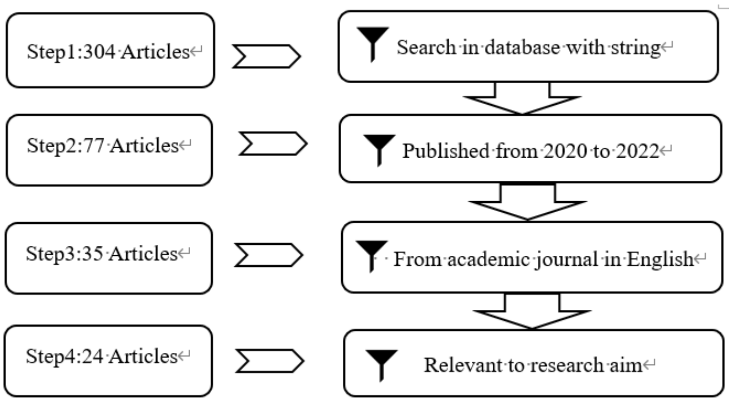

Figure 2: Retrieval flow chart.

3. Research Methods and Characteristics of Alternative Data

3.1. Categories of Alternative Data

Alternative Data is different from the traditional exchange disclosure and company announcement disclosure, and is valuable information conducive to investors' investment decisions. Such as personal consumption information, regional weather conditions, company sales records, etc.

Although there are various sources and forms of alternative Data, the existing alternative data can be divided into three categories. Firstly, data generated by individuals, which contains social network information, product evaluation, search records, shopping preferences, etc. Secondly, data generated by business process, which include measurements made at the stage of application and include things like years in employment, years at address, age, income, credit bureau data, and the percentage of the population in the postcode that defaults on their payments [5]. Thirdly, data generated by sensors, which contains satellite data, GPS positioning data, vehicle track, personal motion track, etc.

3.2. Characteristics of Alternative Data

The alternative Data is typical Big Data, which represents its three characteristics. Firstly, huge data scale and transmission volume. Secondly, high velocity. Data transmission and acquisition happen in real-time or quite close to it. Thirdly, alternative data comes in a variety of forms. The data either already has its own data structure, or it has no data structure.

3.3. Scoping Review Methodology

The first approach is modeling. According to Djeundje et al, by comparing the data set without alternative data and the data set combined with alternative data, and modeling and forecasting the credit risk based on the two data sets, they found that using alternative data can significantly improve the model accuracy of the data set [5]; According to Iakimenko et al, they use the panel GMM approaches and evaluate the dynamic model [6]; According to Rozo et al, thay claim that they use a sizable sample from a well-known online retailer and financial services provider to show that these unique features increase the prediction accuracy of probability of default (PD) models at the account level. The large sample contains information of the repayment credit accounts and also the consumption record [7]; According to Lu et al, they implement multiclass classification (one-versus-all) and numerical regression algorithms respectively for the pre-defined categorical and numerical outcome variables. To be more specific, they used a variety of widely used machine learning models and two ensemble methods [3].

Next, the second method is to use an index or rating. According to González-Fernández and González-Velasco, they gather investor sentiment on the internet by using Google data, which can develop a sentiment index that measures bank credit risk [8]; According to Chen and Chen, they develop a method for anticipating corporate credit ratings by looking at how people feel about companies on their social media. This can help financial institutions assess and manage business risk. These following steps are used to help attain this goal: (i) creating methods for corporate credit rating forecasting, (ii) making the corporate credit rating forecasting mechanism implemented and assessed, and (iii) designing a corporate credit rating forecasting process based on social media big data [9]; According to Giudici et al, they add ‘alternative data’ to typical credit scoring models, which include centrality measures generated from borrowers’ financial ratios and similarity networks [10]; According to Jiang et al, their research is based on a double-blinded comparison between the internal rating of an anonymous traditional lender and the score of big data created by a Chinese service provider in the field of big data. As a result, they may evaluate the prediction ability of big data credit score both independently of the traditional internal score and in tandem with it [11].

3.4. The Advantages and Disadvantages of Alternative Data

3.4.1. Advantages of Alternative Data

The key advantages of alternative data are as follow. It is a better credibility evaluation of comprehensive data, promote efficiency of management of risk assessment. Using alternative credit scores, banks can get a more accurate picture of an applicant's credit and reject loan applications from people with poor credit scores. By analyzing data such as monthly bill payments, you can determine which applicants are less likely to default or fall behind on payments. Furthermore, more personalized customer experience through Banks and other lenders alternative credit scores to provide supplementary data information can quickly examine and approve a loan. More importantly, it can also help banks avoid subjective bias and wrong judgment in processing applications, eliminate discrimination in credit granting, and promote financial fairness and financial freedom. Moreover, greater customer base farmers, students, and small owners and free professionals under the traditional institutions of credit rating system is very difficult to get loans.

Alternative credit scores could solve their conundrum in getting credit. For example, if they pay household expenses and bills on time, they are more likely to get a loan even if they have a low traditional credit score. What’s more, more real-time data in the process of traditional credit scores, bad credit record may be 2 to 3 years of effects on the borrower. It costs a long time for an applicant to clean up a bad credit history. As a result, an applicant will be considered a potential risk even if he has defaulted only twice three years ago, and even though his financial position may have improved considerably over the past three years.

Alternative credit scores analyze implementation data that provides a complete assessment to determine the creditworthiness of the current applicant. Last but not the least, more perfect loan process Through the use of alternative credit scores, can greatly improve the traditional lending process, to reduce the credit risk and to establish a more perfect risk control model and monitoring process. All in all, alternative credit scores are expected to have a significant impact on China's financial services industry. On the one hand, it can bring more customers into the credit ecosystem. On the other hand, banks, microfinance institutions and other financial institutions can be helped to improve loan procedures, approval efficiency and risk management.

3.4.2. Disadvantages of Alternative Data

However, there are still many problems in the practical application of alternative data, which is worth further consideration. The first thought about alternative data is that new data types require corresponding analysis techniques. When we only have volume and price data, traditional technical analysis such as moving averages and bollinger bands can be very useful. However, these technical analyses have little effect on structured accounting data. For this reason, the corresponding analytical methods have emerged, such as multi-factor model. Today, technologies like natural language processing and generalized artificial intelligence are needed for the analysis of unstructured text data as well as more general multimedia data. Clearly, this is placing increasing demands on both managers and investors. Moreover, the moderate improvement of integrating multilevel MVs limits some research findings [12].

With the explosion of alternative data volumes, another problem to confront is the dimension disaster. In the case of predicting stock returns, alternative data represent different independent variables. The surge in parameters causes the overfitting risk of the prediction model to be higher, while the anticipated bias and variance will rise. In addition, the use of different alternative data construction factors can also lead to the problem of multiple testing. Besides, a problem with most alternative data is that it tends to be short. In limited research, the found alternative datasets are typically less than 5 years old (two to three years is common); More than 5 years is a long time. Too short historical data will exacerbate the harm of multiple tests and increase the overfitting problem. Yang et al. were working on it, their model, lasso, had the advantage of allowing variables to be chosen and the complexity to be changed during fitting. Instead of including every variable in the model for fitting, the selection of variables involves identifying crucial elements, which are aimed at selectively fitting variables into the model and achieving accurate performance parameters. Adjusting complexity prevents the model from overfitting. And simultaneously logistic model multicollinearity issue can be solved by the lasso-logistic model, which also accurately identifies the most important explanatory factors. This broadens the model's potential applications. The lasso-logistic model's discrimination power and prediction accuracy are superior to those of the ridge regression and BP neural network models [13].

Another consideration regarding alternative data is whether the data generation (collection) process is unbiased and whether it represents the whole population well. According to Acevedo-Viloria et al., the main limitation of their results is that they cannot extrapolate the results to the entire population. Further research should be conducted to see if the sources of alternative data are applicable to different demographic groups [14]. For another example, the minor sample divergence between the projected and expected results, which shows that the BP neural network's accuracy needs further improvement [15].

It can be seen from the sorted papers that, in the modeling process, the analysis of user credit risk with alternative data can give greater predictive accuracy. Similarly, the construction of an index to assess credit risk can also find that the value obtained by using alternative data is similar to that obtained by using traditional data. In this regard, numerous methods for evaluating credit performance are created. Nevertheless, alternative data provides a warranty for the precision and applicability of the research findings.

However, further research is still much needed to address the main drawback we have identified. For instance, it must be demonstrated whether alternate data sources are effective for certain demographic segments. In addition, the professionalism of practitioners and the timeliness of alternative data need to be fully considered.

Table 1: The advantages and disadvantages of alternative data.

Alternative Data	Advantages	For Customers	Solve their problem in getting credit
		For Financial Institution	Have better credibility evaluation
			Promote efficiency
			Promote financial fairness and financial freedom
	Disadvantages	Limited Corresponding analysis techniques	
		The dimension disaster	
		Biases data	

4. Significance and Current Situations of Alternative Data in Credit Risk Assessment

4.1. Summary of the Impact of Alternative Data

4.1.1. Enhance the Accuracy of Credit Risk Prediction Models

According to our review, a large amount of empirical researches on alternative data constantly develop and modify risk-assessment models, contributing to the improvement of the accuracy of the prediction of default behavior as well as the early warning and prevention.

For examples, according to Zhou et al, the phone usage data, falling under the category of alternative data can exhibit remarkable predictive capabilities, owing to its universality, abundance and authenticity. [16]. Besides, traditional evaluation is labor-and-time-consuming and prejudiced. While alt data can be an effective tool for e-commerce to manage sparsely and imbalancedly distributed data [16].

4.1.2. Optimize the Financing Circumstance

In terms of emerging economies, those papers regarding to alternative data have important policy implications. With higher accuracy, alternative data can help financial institutes to make decision about whether to grant financial facilities to an applicant or not, and to gain more profits. Therefore, those methods based on alternative data assist financial institutions in expanding their offerings to individuals with lower incomes and less education [3]. Thus, consumers have more chance to credit by having alternative data using to evaluate them.

4.1.3. Offer a Solution to the Thin Files Problem

Alternative data can offer a solution to the thin files problem. In developed economies, various forms of information asymmetry, such as financial misreporting, exist. It is no doubt that the alternative data offers a novel source of information value in the financial services industry and enriches conventional business practices.

Overall, the utilization of using alternative data can potentially mitigate information asymmetry in financial lending and other related fields.

4.1.4. Get Better Understanding of Existing Risk Assessments

Alternative data provide more explanation of how various kinds of alternative data affect credit risk, helping us identify some key factors affecting the credit risk. One example is that certain research has demonstrated how individuals' character can be reflected through their mobile phone usage behavior, thereby aiding in the prediction of credit risk. [16]. By examining the interconnections between multiple factors and credit risk, models utilizing alternative data can more effectively elucidate and forecast credit risk.

4.1.5. Improve Social Welfare

Those studies demonstrate there is enormous possibility in utilizing alternative data to enhance social welfare within the financial market. This type of data has the capability to mitigate inequalities in the financial service sector. Besides, various credit scoring models based on alternative data can be employed, not only by borrowers and lenders, but also by regulators and supervisors to oversee peer lending. This can help to shield consumers and uphold financial stability.

4.2. Examples of Application

Recent availability of alt data has triggered a significant transformation in both economic researches and practice. In terms of consumer lending, financial institutions have started utilizing alternative data to assess the creditworthiness.

There are lots of concrete instances concerning the application of alternative data in credit risk that has been realized in reality. In the terms of rental payment history: Rent reporting companies like RentTrack and Kharma, report a borrower's rental payment history to credit bureaus, helping renters build credit. Rental payment history can also be used to evaluate a borrower's creditworthiness. Moreover, rental payment data can be used to assess the creditworthiness of renters who may not have a traditional credit history. For example, FICO, the credit scoring company, offers a scoring model that uses rental payment data to predict a borrower's likelihood of default. In the terms of social media data: LenddoEFL, a credit scoring company, uses social media data to assess the creditworthiness of borrowers who do not have a traditional credit history. By analyzing their social media profiles and activity, the company can gain insights into their behavior, such as their level of social connectedness, which can be used to predict their ability to repay loans. In the terms of utility payment data: Experian offers a scoring model that uses utility payment data to predict a borrower's likelihood of default. In the terms of e-commerce: Ant Financial, the financial services arm of Alibaba, uses data from its e-commerce platforms to assess the creditworthiness of borrowers. In the terms of phones usage data: Tala, a mobile lending platform, uses data from borrowers' mobile phones, such as call and text history, to assess their creditworthiness. Some other forms of alternative data are also be used wisely: LendingClub, a peer-to-peer lending platform, uses alternative credit data, such as education and employment history, to assess the creditworthiness of borrowers. Nevertheless, there has been relatively little research on alternative data within real-world business contexts

5. Conclusions

5.1. Existing Challenges and Concerns

While using alternative data has generally been lauded, it has not been without controversy. First, the quality of alt data can be inconsistent, and it may not always be reliable or accurate. For instances, incorporating public records which contain information such as minor criminal offenses, unemployment, provides a lasting record of temporary personal situations. Besides, the costs of obtaining alt data may be rather high and the calculations are complicated and costly with the requirement of large amount of data samples. Alternative data is often unstructured and may require significant cleaning and normalization before it can be used.

Ethical issues such as privacy concerns and data bias also need to be addressed. Use of data such as online social connections and contact lists has sparked concerns. Assessing individuals according to their friend networks has been considered unjust and potentially unethical. Moreover, many studies discussing alternative data is centered around technologically enabled and generated data, which holds relevance in today's technologically advanced society. This exclude those who have been already marginalized, thus exacerbating the disparity in financial inclusion.

5.2. Recommendations and Expectations

To overcome the above challenges, initially, companies must be transparent about their data sources and methods, and have rigorous quality control processes in place.

In order to optimize the learning performance and final outcomes of credit prediction models, it is crucial to ensure the authenticity, reliability, and comprehensiveness of data sources, as the number and authenticity of data samples have a significant impact. This emphasizes the need to prioritize data quality in the future. Furthermore, the future of alt data is closely linked to the artificial intelligence and machine learning technologies. These technologies can help address some of the current challenges of alt data, such as data quality and bias, by automating data cleaning and normalization processes. Lastly, efficient integration of alt data with traditional sources can lead to profit optimization and reduced possibility for prediction bias.

However, overall, despite these challenges, the future of alt data is promising. As technology continues to advance, the quality of alt data will continue to improve, making it more reliable and accurate. Moreover, as more companies adopt alt data analysis, the market will become more competitive, leading to increased innovation and more refined data analysis techniques. This will help address some of the current challenges of alt data, such as data bias and quality issues. In conclusion, alt data is a promising field with significant potential for improving financial analysis.

References

[1] Anderson, E.J.: Business risk management: models and analysis. John Wiley & Sons (2013).

[2] Francis, E., Blumenstock, J.,Robinson, J.: Digital credit: A snapshot of the current landscape and open research questions. CEGA White Paper, pp.1739-76 (2017).

[3] Lu, T., Zhang, Y.,Li, B.: The value of alternative data in credit risk prediction: Evidence from a large field experiment (2019).

[4] Tranfield,D.,Denyer,D.,Smart,P.: Towards a methodology for developing evidence- informed management knowledge by means of systematic review. British Journal of Management, 14(3): 207- 222 (2003).

[5] Djeundje, V.B., Crook, J., Calabrese, R., Hamid, M.: Enhancing credit scoring with alternative data, Expert Systems with Applications, 163, p. N.PAG (2021).

[6] Iakimenko, I., Semenova, M., Zimin, E.: The more the better? Information sharing and credit risk. Journal of International Financial Markets, Institutions and Money 80, 101651 (2022).

[7] Rozo B. J. G., Jonathan Crook, Galina Andreeva: The role of web browsing in credit risk prediction, Decision Support Systems 164, 113879 (2023).

[8] González-Fernández, M., González-Velasco, C.: An alternative approach to predicting bank credit risk in Europe with Google data, Finance Research Letters 35, 101281 (2020).

[9] Yuh-Jen Chen, Yuh-Min Chen: Forecasting corporate credit ratings using big data from social media, Expert Systems with Applications Volume 207, 118042 (2022).

[10] Paolo Giudici, Branka Hadji-Misheva,Alessandro Spelta: Network based credit risk models, Quality Engineering, 32:2, 199-211 (2020).

[11] inglin Jiang, Li Liao, Xi Lu, Zhengwei Wang, Hongyu Xiang: Deciphering big data in consumer credit evaluation, Journal of Empirical Finance Volume 62, 28-45 (2021).

[12] Yufei Xia, Yinguo Li, Lingyun He, Yixin Xu, Yiqun Meng: Incorporating multilevel macroeconomic variables into credit scoring for online consumer lending, Electronic Commerce Research and Applications Volume 49, 101095 (2021).

[13] Yang, Y., Chu, X., Pang, R., Liu, F., Yang, P.: Identifying and predicting the credit risk of small and medium-sized enterprises in sustainable supply chain finance: Evidence from China. Sustainability 13(10), 5714 (2021).

[14] Acevedo-Viloria, J.D., Pérez, S.S., Solano, J., Zarruk-Valencia, D., Paulin, F.G., Correa-Bahnsen, A.: Feature-Level Fusion of Super-App and Telecommunication Alternative Data Sources for Credit Card Fraud Detection, In 2021 IEEE International Conference on Intelligence and Security Informatics pp. 1-6, IEEE (2021).

[15] Du G., Liu Z., Lu H.: Application of innovative risk early warning mode under big data technology in Internet credit financial risk assessment. Journal of Computational and Applied Mathematics Volume 386, 113260 (2021).

[16] Zhou J., Wang C., Ren F., Chen G.: Inferring multi-stage risk for online consumer credit services: An integrated scheme using data augmentation and model enhancement. Decision Support Systems 149, 113611 (2021).