International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

An Intelligent Model for Predicting Fintech Operational Health and Sustainability in Africa Using Machine Learning and Big Data Mining

Grace Olaronke Shobayo[1]; Patrick Njock Egbe [2]; Chibuzo, Anthony [3]; Fouodji Ngnoukaho Channelle[4]; Egbe Toochi Paul[5]; Jerome Christopher Atisu[6]; Rudolf Seth Afeku[7]; Onalaja, Samiat A [8]; Joshua Tangu Daniel[9] Victor Andikan [10]; Kofi Nyame Jnr [11]; Ojukwu Sylvia O[12]; Edidiong Hassan[13]; Osazuwa Joy[14]; Babatunde Oluwatunmise Michael [15], John Obasi[16]

1- Department of Economics, University of Jos, Nigeria
2 – Department of Economics and Management, University of Yaounde, Cameroon
3 - Department of Computer Science, Godfrey Okoye University, Nigeria
4 – Department of Accounting, University of Bamenda, Cameroon.
5 – Department of Computer Science, Federal Polytechnic, Nekede, Owerri, Nigeria
6 – Department of Accounting and Finance, Kwame Nkrumah University of Science and Technology, Ghana
7 – Department of Accounting and Finance, Accra Business School, Ghana
8 – Department of Business Education and management, Tai Solarin University of Education, Nigeria
9 – Department of Computer Engineering, Ghana Communication Technology University, Ghana
10 - Department of Economics, University of Lagos, Nigeria
11 – Department of Accounting, University of Education, Winneba, Ghana
12 – Department of Supply Chain Management, Nnamdi Azikiwe University, Nigeria
13 – College of Business, Lewis University
14 – Department of Economics, Ambrose Alli University, Nigeria
15 – Department of Business Administration, Ahmadu Bello University, Zaria, Nigeria.
16 – Department of Accountancy, Ecotes University, Benin Republic

Doi : https://doi.org/10.55248/gengpi.5.0824.2111

ABSTRACT

The rapid growth of Fintech in Africa poses significant operational and regulatory challenges, underscoring the need for effective risk management and sustainability assessment tools. This study develops and applies a Random Forest model to predict the operational health and sustainability of Fintech companies in Africa. Using a comprehensive dataset from ten sources spanning a period of 5-years (January 2019 to December 2023), the model integrates financial metrics, regulatory data, and macroeconomic indicators. The results show an overall accuracy of 75% and precision of 82% in detecting 'At Risk' companies, but a 0% recall in identifying 'Healthy' operations. The ROC-AUC score of 0.15 indicates limited discriminative power. Feature importance analysis highlights regulatory capital requirements, credit-to-GDP ratio, and cost-to-income ratio as critical predictors. The study reveals challenges of class imbalance and model conservatism, emphasizing the need for enhanced calibration and more diverse, dynamic data inputs. The findings have significant implications, suggesting that while the model provides a basis for risk identification, adjustments are necessary to improve predictive accuracy and real-world utility. Future research should focus on adaptive learning mechanisms and advanced ensemble techniques to better capture Fintech operational complexities.

Keywords: Big Data Analytics, Fintech Operational Health, Machine Learning, Predictive Modelling

INTRODUCTION

The fintech industry's rapid expansion globally, especially in emerging markets like Africa, holds the promise of enhancing financial inclusion (Campbell, 2024; Murinde et al., 2022; Arkanuddin et al., 2021; Ahmad et at., 2020). While fintech firms in Africa are striving to reach the unbanked and underbanked populations, challenges such as economic volatility, regulatory hurdles, and technological risks threaten their operational sustainability (Chowdhury, 2023; Stojanović et al., 2021; Kapaya, S.M. 2020).

Machine learning (ML) has been recognized as a transformative tool in the fintech sector, with applications ranging from risk assessment to optimizing customer service (Shino et al., 2022). Despite the potential of ML, predicting the success of fintech businesses remains intricate due to the complex financial environments they navigate (Mulyk, 2022). While existing research has shown the potential of machine learning in revolutionizing various fintech aspects, including risk evaluation and customer service enhancement (Shang, Z. and Wang, Z. 2022; Liu et al., 2021). However, there is a scarcity of studies focusing on the operational health of fintech firms in African markets, where economic and regulatory landscapes differ significantly from developed nations (Lobozynska et al., 2023; Wang et al., 2022).

To address this gap, this study aims to develop an intelligent machine learning model utilizing the Random Forest algorithm to predict the operational health and sustainability of African fintech companies (Cambaza, 2023). The Random Forest algorithm is chosen for its ability to handle large datasets effectively and model intricate relationships without overfitting (Hudithi & Siddiqui, 2021; Xu et al., 2020). By focusing on operational health, which is crucial for long-term sustainability in the fintech sector, this research seeks to provide insights that can inform strategic planning and regulatory frameworks, fostering the growth of the fintech industry while ensuring financial stability and innovation (Shang & Wang, 2022).

The significance of this research lies in its emphasis on operational health, a critical factor determining the long-term viability of fintech firms. The findings from this study could guide strategic planning and regulatory frameworks, ensuring that the fintech sector in Africa not only prospers but also bolsters financial stability and innovation (Shin & Choi, 2019).

METHODOLOGY

Data Collection and Integration

The study utilized a comprehensive dataset aggregated from ten different sources, reflecting various aspects of Fintech operations in Africa. These sources included financial reports, regulatory databases, surveys, and industry publications. The data covered a period of five years, from January 2019 to December 2023, providing a robust temporal framework for analysis. This integrated dataset comprised financial metrics, bank-specific data, macroeconomic indicators, supervisory ratings, and other relevant variables. All datasets were merged based on the 'Month' attribute, ensuring a coherent and unified time series analysis framework.

Data Preprocessing

The integrated data underwent several preprocessing steps. Missing values were imputed using the mean for continuous variables and the mode for categorical variables. Categorical variables were encoded using one-hot encoding to facilitate computational processing. Numerical features were scaled using a standard scaler to neutralize the effect of differing scales. A thorough data quality check was conducted to identify and rectify any outliers, duplicates, and inconsistencies.

Feature Engineering

New features were engineered by combining existing variables, such as ratios of financial metrics to macroeconomic indicators (e.g., GDP, inflation rate), to capture complex relationships and enhance the model's predictive capabilities.

Feature Selection

Initial feature selection was based on variance, removing non-contributing features with zero variance. Subsequently, a Random Forest algorithm was employed to evaluate feature importance, guiding the retention of the most impactful predictors.

Model Development

A Random Forest classifier was developed to predict the operational health of Fintech companies, categorized into binary outcomes ('At Risk' or 'Healthy'). The model was trained using 80% of the data and tested on the remaining 20%. Class weightage was set to 'balanced' to account for the imbalance in the target variable. Model parameters were optimized based on cross-validation results to maximize accuracy and the area under the ROC curve.

Hyperparameter Tuning

Hyperparameters were fine-tuned using a grid search approach, focusing on the number of estimators (100-500), maximum depth (5-15), and minimum samples per leaf (1-5). The search strategy employed was exhaustive, with 10 iterations. The optimal parameters were chosen based on achieving the highest ROC-AUC score.

Model Evaluation

The model's performance was assessed using a suite of metrics including accuracy, ROC-AUC, precision, recall, F1-score, and a confusion matrix. Detailed results are presented in the subsequent section.

Software and Libraries

The analysis was performed using Python 3.8, with libraries including scikit-learn 1.0, pandas 1.3, numpy 1.20, and matplotlib 3.5, among others.

Results and Discussion

The Random Forest model's performance was evaluated using various metrics, as presented in Table 1.

Table 1: Summary of Model Performance Metrics

Class	Precision	Recall	F1-Score	Support
0 (At Risk)	0.82	0.9	0.86	10
1 (Healthy)	0	0	0	2
Accuracy			0.75	
Macro Avg	0.41	0.45	0.43	12
Weighted Avg	0.68	0.75	0.71	12

Table 1 presents the key performance metrics of the Random Forest model, which predicted the operational health of Fintech companies with an overall accuracy of 75%. The model demonstrated a strong ability to identify 'At Risk' cases, achieving a precision of 82% and a recall of 90%. However, it struggled to detect 'Healthy' cases, resulting in a recall of 0% for this class. The F1-score for the 'At Risk' class was 86%, but the lack of detection in the 'Healthy' class yielded an F1-score of 0%. The ROC-AUC score of 0.15 indicates room for improvement in the model's ability to effectively distinguish between the two classes.

The results indicate a strong ability to predict 'At Risk' cases accurately, but a challenge in correctly identifying 'Healthy' cases, suggesting potential issues with class imbalance.

Table 2 presents the top ten features ranked by their importance in predicting the operational health of Fintech companies in Africa, as determined by the Random Forest model. The table reveals that regulatory capital requirements, credit-to-GDP ratio, and cost-to-income ratio are the three most influential factors, indicating a significant impact of financial stability and economic conditions on Fintech health.

Table 2: Top 10 Influential Features Determining Fintech Operational Health

Rank	Feature	Importance
1	Regulatory Capital Requirements	0.0542
2	Credit-to-GDP Ratio	0.0512
3	Cost-to-Income Ratio	0.0455
4	Customer Satisfaction	0.0431
5	Provision Coverage	0.0415
6	Return on Assets	0.041
7	Blockchain Adoption Rate	0.0352
8	Digital Payment Adoption Rate	0.0351
9	Banking Sector Stability Indices	0.0346
10	Asset Utilization	0.0302

The feature importance ranking in Table 2 offers valuable insights into the key drivers of Fintech operational health. The prominence of regulatory capital requirements highlights the vital role of regulatory compliance and financial robustness in sustaining Fintech operations. This finding aligns with regulatory frameworks that emphasize capital adequacy to mitigate operational risks. The importance of credit-to-GDP ratio and cost-to-income ratio further underscores the impact of economic conditions and financial management on Fintech health. These findings have significant implications for Fintech companies, regulators, and policymakers, emphasizing the need for robust financial management, regulatory compliance, and economic stability to ensure the sustainable operation of Fintech firms.

Figure 1 shows the feature importance plot derived from the Random Forest model, illustrating the relative importance of the top 10 features in predicting the operational health of Fintech companies. The plot ranks features based on their influence on the model's predictions, with regulatory capital requirements, credit-to-GDP ratio, and cost-to-income ratio emerging as the most critical factors.

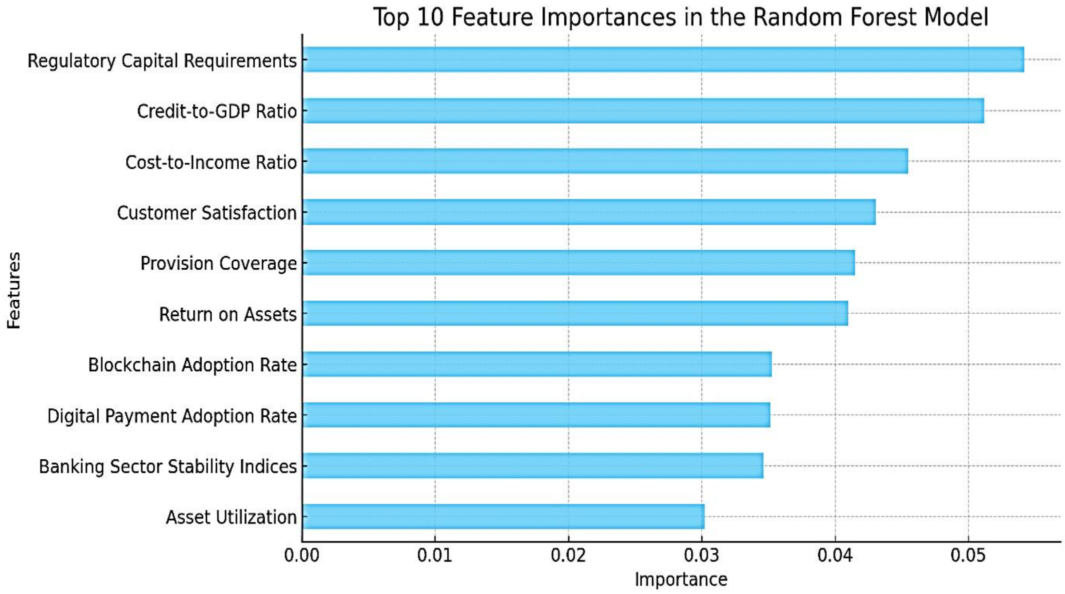

Figure 1: Feature Importance Plot

The feature importance plot in Figure 1 provides crucial insights into the key drivers of Fintech operational health. The prominence of regulatory capital requirements as the top feature highlights the vital role of regulatory compliance and financial stability in the Fintech sector. This finding aligns with regulatory emphasis on maintaining adequate capital to mitigate potential risks associated with financial operations.

The importance of credit-to-GDP ratio and cost-to-income ratio underscores the sensitivity of Fintech operations to broader economic conditions and internal efficiency metrics, respectively. These insights are invaluable for Fintech companies seeking to optimize their operational strategies in a highly competitive and regulated market.

The plot also suggests that technological factors, such as blockchain adoption rate, while significant, rank lower than traditional financial indicators. This implies that innovation, though crucial, should be balanced with foundational aspects of financial health to ensure operational viability.

The implications of these findings are broad, suggesting that efforts to enhance Fintech sustainability should focus on strengthening regulatory compliance, economic resilience, and technological innovation. Policymakers and regulators can use these insights to develop supportive policies that foster a conducive environment for Fintech growth and sustainability

Figure 2 displays the ROC curve for the Random Forest model, plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. This curve visually assesses the model's ability to distinguish between 'At Risk' and 'Healthy' operational statuses

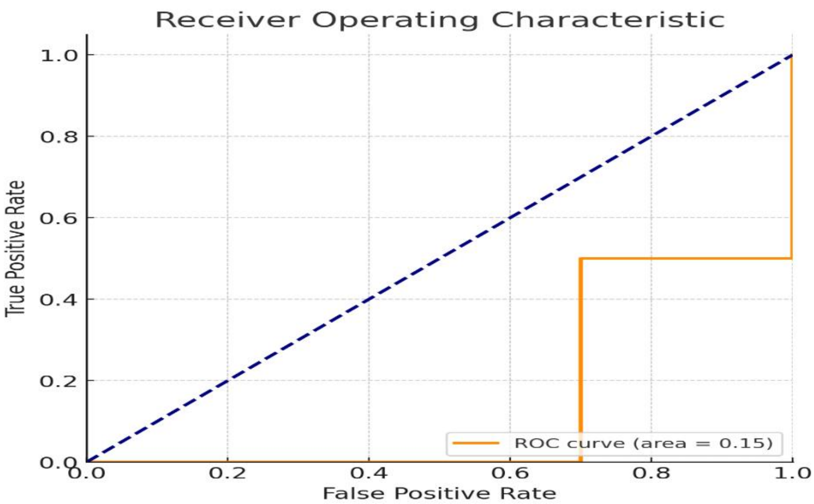

Figure 2: Receiver Operating Characteristic (ROC) Curve

The ROC curve in Figure 2 critically evaluates the discriminative power of our predictive model. The AUC of 0.15 indicates that the model performs better than a no-skill classifier but shows room for improvement. The curve's shape and position reveal that the model identifies a high number of true positives for the 'At Risk' class but also incurs a proportionally high number of false positives.

The relatively low AUC highlights challenges in model performance, particularly in differentiating between 'Healthy' and 'At Risk' categories. This limitation may be attributed to overfitting to the 'At Risk' class or dataset imbalance.

From a practical perspective, this finding underscores the need for additional or alternative metrics and features to enhance the model's sensitivity and specificity. Improving the model may involve integrating nuanced data related to customer behaviors, market conditions, or non-traditional financial indicators.

Figure 3 presents the confusion matrix for the Random Forest model, visually representing the model's performance in classifying Fintech companies' operational health statuses into 'At Risk' and 'Healthy' categories. The matrix quantifies true positives, true negatives, false positives, and false negatives, providing an intuitive display of the model's predictive accuracy and errors

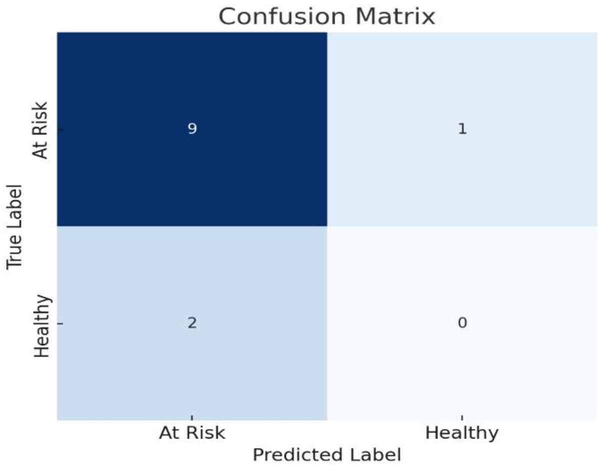

Figure 3: Confusion Matrix of Model Predictions

The confusion matrix in Figure 3 is helpful in understanding the practical implications of the model's predictive capabilities. The high number of True Positives suggests effective identification of Fintech companies at risk, crucial for preventative strategies and risk management. However, False Negatives raise concerns about the model's sensitivity and potential risks of undetected financial instability. False Positives indicate unnecessary scrutiny or resource allocation, potentially diverting attention from genuine cases of risk. The absence of True Negatives highlights a critical area of improvement, as the model fails to reassure the operational health of companies genuinely performing well.

Figure 4 shows a histogram of the predicted probabilities assigned by the Random Forest model for Fintech companies being classified as 'Healthy'. This visualization provides insights into the distribution of probabilities across the tested data, indicating the model's confidence in its predictions for each class.

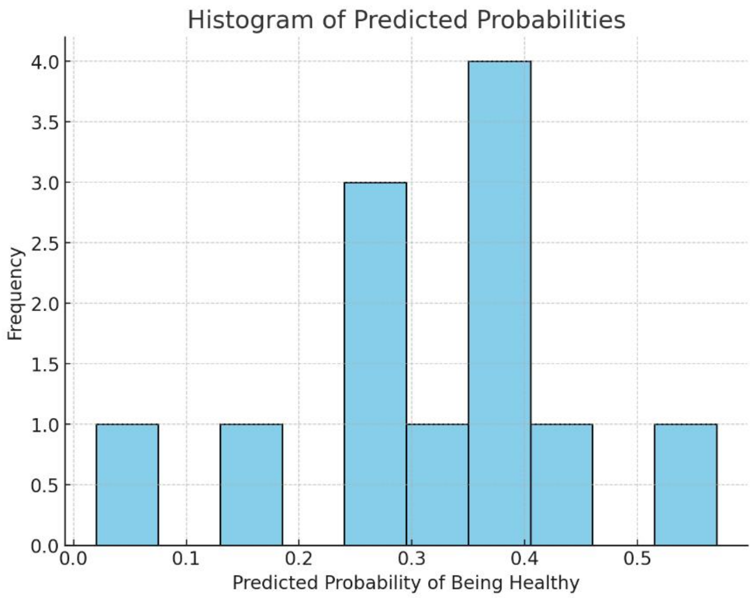

Figure 4: Histogram of Predicted Probabilities for 'Healthy' Classifications

The histogram in Figure 4 reveals the model's conservative nature in predicting the 'Healthy' class. Most probability scores cluster at the lower end, suggesting the model rarely assigns high confidence to this outcome. This tendency indicates a model that may be overly cautious or biased towards predicting higher risk.

For Fintech companies and regulators, these insights are valuable for calibrating risk assessment tools to reflect actual risk levels accurately. Ensuring predictive models are neither overly pessimistic nor unduly optimistic in assessing health can help in efficient resource allocation and more targeted regulatory oversight.

Figure 4 underscores the need for a nuanced understanding of model outputs and serves as a basis for further refining predictive analytics in Fintech. It highlights the importance of focusing on how the model expresses confidence across its predictions, crucial for practical application and trustworthiness.

Critical Discussion

The study's findings highlight the potential of machine learning and big data mining in predicting Fintech operational health and sustainability in Africa. The Random Forest model's performance, as evaluated by the ROC-AUC score and confusion matrix, demonstrates its ability to identify at-risk Fintech companies. However, the model's conservative nature in predicting the 'Healthy' class and the presence of false negatives suggest areas for improvement.

Implications

1. Risk Management: Fintech companies and regulators can leverage the model's predictive capabilities to proactively identify and mitigate potential risks, ensuring operational sustainability.

2. Model Refinement: Further refinement of the model, including threshold adjustment and calibration, can enhance its accuracy and reliability.

3. Data Quality: The study emphasizes the importance of high-quality, diverse, and representative data in developing robust predictive models.

4. Regulatory Oversight: Regulators can use the model's insights to inform targeted oversight, ensuring that Fintech companies operate within a safe and sound framework.

Limitations

1. Class Imbalance: The model's skewed accuracy, favoring the 'At Risk' classification, is largely due to class imbalance. This imbalance affects the model's utility in confidently declaring firms as 'Healthy' and impacts the ROC-AUC score, reflecting poor discriminative ability between operational states.

2. Predictive Conservatism: The model demonstrates a conservative approach, overly favoring riskier outcomes. This tendency, while potentially safeguarding against financial oversight, could lead to inefficiencies and unnecessary caution in operational management.

3. Feature Dynamics: The reliance on financial and regulatory metrics highlights an underutilization of potential predictive indicators like technological innovation metrics, crucial for a sector driven by cutting-edge technology.

1. Incorporate Advanced Modeling Techniques: Explore complex ensemble methods or hybrid models for refined predictions.

2. Adaptive Learning Models: Develop models that adapt to changing economic and market conditions.

3. Regulatory Frameworks: Examine the impact of regulatory frameworks on model performance, guiding policy development that supports innovative yet stable Fintech operations.

Conclusion

This study demonstrates the feasibility of using machine learning and big data mining to predict Fintech operational health and sustainability in Africa. The Random Forest model's performance highlights its potential as a valuable tool for risk management and regulatory oversight. However, the study's findings also underscore the need for ongoing model refinement, data quality improvement, and nuanced understanding of model outputs. By addressing these areas, stakeholders can harness the full potential of predictive analytics to foster a resilient and sustainable Fintech ecosystem in Africa. While the model presents a promising foundation for predicting Fintech health in Africa, ongoing refinement and adaptation are essential. This study advances our understanding of Fintech operational dynamics and sets the stage for developing more robust predictive models that support sustainable growth in emerging markets.

References:

[1]. Cambaza, E. (2023). The role of fintech in sustainable healthcare development in sub-saharan africa: a narrative review. Fintech, 2(3), 444-460. https://doi.org/10.3390/fintech2030025

[2]. Campbell, D. (2024). Machine learning in fintech., 206-217. https://doi.org/10.4324/9781032644165-13

[3]. Liu, C., Fan, Y., & Zhu, X. (2021). Fintech index prediction based on rf-ga-dnn algorithm. Wireless Communications and Mobile Computing, 2021, 1-9. https://doi.org/10.1155/2021/3950981

[4]. Shang, Z. and Wang, Z. (2022). Machine learning in finance: a brief review., 258-264. https://doi.org/10.2991/978-94-6463-010-7_27

[5]. Stojanović, B., Božić, J., Hofer-Schmitz, K., Nahrgang, K., Weber, A., Badii, A., … & Runevic, J. (2021). Follow the trail: machine learning for fraud detection in fintech applications. Sensors, 21(5), 1594. https://doi.org/10.3390/s21051594

[6]. Wang, H., Chen, X., Du, J., & Lai, K. (2022). Classification of fintech patents by machine learning and deep learning reveals trends of fintech development in china. Mathematical Problems in Engineering, 2022, 1-15. https://doi.org/10.1155/2022/1852447

[7]. Xu, L., Lu, X., Yang, G., & Shi, B. (2020). Identifying fintech innovations with patent data: a combination of textual analysis and machine-learning techniques., 835-843. https://doi.org/10.1007/978-3-030-43687-2_70

[8]. Ahmad, A.H., Green, C.J., & Jiang, F. (2020). Mobile Money, Financial Inclusion and Development: A Review with Reference to African Experience. *Econometric Modeling: International Financial Markets - Emerging Markets eJournal*.

[9]. Kapaya, S.M. (2020). Financial System and Economic Development in Africa: A Review and Policy Recommendations. *International Journal of Academic Research in Accounting, Finance and Management Sciences, 9*.

[10]. Murinde, V., Rizopoulos, E., & Zachariadis, M. (2022). The impact of the FinTech revolution on the future of banking: Opportunities and risks. *International Review of Financial Analysis*.

[11]. Shang, Z. and Wang, Z. (2022). Machine learning in finance: a brief review. Atlantis Highlights in Intelligent Systems, 258-264. https://doi.org/10.2991/978-94-6463-010-7_27

[12]. Arkanuddin, M., Saragih, F., & Nugroho, B. (2021). The key role of the financial regulation in fintech ecosystem: a model validation. Studies of Applied Economics, 39(12). https://doi.org/10.25115/eea.v39i12.6239

[13]. Chowdhury, M. (2023). Assessing the evolutionof fintech adoption in traditional banking: a bibliometric analysis using r. Turk Turizm Arastirmalari Dergisi. https://doi.org/10.26677/tr1010.2023.1349

[14]. Hudithi, F. and Siddiqui, K. (2021). Designing the guidelines for fintech curriculum. Journal of Entrepreneurship and Sustainability Issues, 9(1), 633-643. https://doi.org/10.9770/jesi.2021.9.1(39)

[15]. KILIÇ, G. (2023). The emergence of islamic fintech and its applications. International Journal of Islamic Economics and Finance Studies. https://doi.org/10.54427/ijisef.1328087

[16]. Lobozynska, S., Vladychyn, U., & Skomorovych, I. (2023). Digital banking transformation through cooperation with fintech startups in ukraine. Finanse I Prawo Finansowe, 1(37), 85-101. https://doi.org/10.18778/2391-6478.1.37.05

[17]. Mulyk, T. (2022). Formation and development of fintech in ukraine. Economics & Education, 7(1), 25-33. https://doi.org/10.30525/2500-946x/2022-1-3

[18]. Shin, Y. and Choi, Y. (2019). Feasibility of the fintech industry as an innovation platform for sustainable economic growth in korea. Sustainability, 11(19), 5351. https://doi.org/10.3390/su11195351

[19]. Shino, Y., Lukita, C., Rii, K., & Nabila, E. (2022). The emergence of fintech in higher education curriculum. Startupreneur Business Digital (Sabda Journal), 1(1), 11-22. https://doi.org/10.33050/sabda.v1i1.71

[20]. Stojanović, B., Božić, J., Hofer-Schmitz, K., Nahrgang, K., Weber, A., Badii, A., … & Runevic, J. (2021). Follow the trail: machine learning for fraud detection in fintech applications. Sensors, 21(5), 1594. https://doi.org/10.3390/s21051594