Title Page:

Enhancing data breach detection and ensuring customer data privacy in the banking sector using Isolation forest compared with Naïve Bayes Classification

Karthik Natarajan P L[1], Dr E K Subramanian [2]

Karthik Natarajan P L[1]
Research Scholar,
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India. Pin code: 602105
karthikpalaniappan96@gmail.com

Dr E K Subramanian [2]
Associate Professor
Department of Programming
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India. Pin code: 602105
Subramanianek.sse@saveetha.com

Keywords: Algorithm(s), Machine Learning, Performance, Data Breach, Detection, Cybersecurity, Customer Data Privacy, Accuracy

ABSTRACT:

Aim: This research endeavors to significantly enhance data breach detection and ensure customer data privacy in the banking sector through the integration of the Isolation Forest algorithm, coupled with a comprehensive comparison against the Naive Bayes Classification model. The study places a substantial emphasis on performance evaluation metrics, particularly Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), to meticulously gauge the efficacy of the proposed framework in the domains of cybersecurity and data breach detection. **Materials and Methods:** Addressing critical concerns related to data breach detection and privacy in the banking sector, an extensive dataset is meticulously examined. The research employs two pivotal models, Isolation Forest and Naive Bayes Classification, providing a holistic assessment of the proposed framework. The evaluation places particular importance on accuracy and performance, contributing to the understanding of their significance in the banking sector's cybersecurity landscape. **Results:** The findings from this investigation carry significant implications for the cybersecurity and data breach detection domains within the banking sector. The Isolation Forest model demonstrates an impressive accuracy rate, showcasing the framework's efficacy in enhancing data breach detection and ensuring customer data privacy. In stark contrast, the Naive Bayes Classification approach exhibits relatively lower performance, as substantiated by a detailed analysis revealing a notable disparity between the two models. The statistical examination reveals a p-value of 0.032 ($p < 0.05$) for both accuracy and loss, indicating a statistically significant distinction in the cybersecurity and data breach detection domains. **Conclusion:** The ramifications of this research are profound for the banking sector, underscoring the superior performance of the Isolation Forest algorithm over Naive Bayes Classification in enhancing data breach detection and ensuring customer data privacy. This study represents a substantial contribution to the field of cybersecurity in the banking sector, advocating for enhanced security measures and reinforcing customer trust through innovative techniques.

Keywords: Algorithm(s), Machine Learning, Performance, Data Breach, Detection, Cybersecurity, Customer Data Privacy, Accuracy

INTRODUCTION:

In the dynamic landscape of cybersecurity within the banking sector, the challenge of detecting data breaches and ensuring customer data privacy remains paramount. This paper introduces an innovative methodology that integrates the Isolation Forest algorithm and scrutinizes its efficacy by comparing it with Naïve Bayes Classification. The primary objective is to enhance data breach detection and reinforce customer data privacy in the banking sector. The specific focus lies in evaluating the superior performance of Isolation Forest over Naïve Bayes Classification, particularly concerning accuracy and overall effectiveness, highlighting the crucial considerations of cybersecurity and data breach detection.

The genesis of this research is rooted in an extensive review of scholarly papers addressing cybersecurity nuances in the banking sector. Leveraging insights from 1,500 comprehensive studies sourced from platforms like IEEE Xplore, ResearchGate, Elsevier, and Springer, the literature review uncovers traditional and cutting-edge approaches to data breach detection and privacy. Comprising 400 articles from IEEE Xplore, 200 from ResearchGate, 800 from Elsevier, and 100 from Springer, this review not only enriches the research but also underscores an existing gap—the absence of a direct comparative analysis between Isolation Forest and Naïve Bayes Classification in the context of data breach detection within the banking sector. This gap forms the core objective of our research team, comprised of experts in cybersecurity and machine learning.The overarching goal of this study is to augment data breach detection and ensure customer data privacy in the banking sector by incorporating the Isolation Forest algorithm. The paper aims to comprehensively evaluate the performance of Isolation Forest compared to Naïve Bayes Classification, providing valuable insights to advance cybersecurity in the banking sector and bolster customer trust. The study's emphasis on data breach detection and privacy underscores the innovative nature of the proposed approach.

MATERIALS AND METHODS:

Conducted within the Cybersecurity and Data Privacy Lab at SIMATS University, a renowned institution in security and data protection research, this investigation aims to elevate data breach detection and ensure customer data privacy in the banking sector. The Isolation Forest algorithm is integrated into the study, and a comparative analysis is conducted with Naïve Bayes Classification. The dataset utilized comprises 1,500 records related to banking transactions, resulting in 3,000 instances, mirroring the robustness of the reference study. A meticulous pre-study assessment addresses cybersecurity concerns through a comprehensive power analysis, with statistical parameters set at α=0.05 and power=0.80.This research adheres strictly to ethical and legal standards, placing a priority on customer privacy and data protection. No personal customer data or sensitive information is employed, aligning with the ethical considerations outlined in the reference study. Furthermore, the investigation does not involve human or animal specimens. The implementation and analysis adopt a combination of programming languages, employing Python for algorithmic tasks and R for data analysis, echoing the methodology used in the reference study. Open-source tools and frameworks specific to machine learning and data management are utilized for accuracy and reliability. For computational tasks, Google Colab is employed, highlighting the significance of cybersecurity in online research. The system configuration comprises an AMD Ryzen 7 4800H Processor, 16 GB of RAM, 1 TB of SSD storage, and an NVIDIA RTX 3050 graphics card with 4 GB of dedicated video memory. The software tools include Python 3.10, Windows 11, Chrome, and IBM SPSS v26 for statistical analysis. The dataset, titled 'Data breach detection and ensuring customer data privacy,' encompasses 5 attributes and 10,684 data rows, forming the foundation for experimental assessments in the domains of cybersecurity, web cookies, and comparative analysis.

Isolation Forest:

The Isolation Forest algorithm stands as a powerful tool in the domain of machine learning, particularly well-suited for enhancing data breach detection and ensuring customer data privacy in the banking sector. It excels in isolating anomalies within datasets, making it instrumental in identifying potential breaches and securing sensitive customer information. At its core, the Isolation Forest algorithm leverages ensemble learning principles, constructing isolation trees to identify anomalies efficiently. Each tree independently assesses data points, and anomalies are identified based on their shorter average path lengths within the trees. In the context of banking sector cybersecurity, Isolation Forest can be applied to learn from historical transaction data, adapt to evolving patterns, and effectively detect unusual activities that may indicate a data breach. The algorithm's adaptability and accuracy make it a valuable asset in the ongoing efforts to strengthen data breach detection and customer data privacy in the banking sector.

Pseudocode:

```
Step 1: Gather historical transaction data and prepare it for input into the Isolation Forest model, emphasizing the importance of data breach detection and customer data privacy in the banking sector.
Step 2: Initialize the Isolation Forest model with the desired architecture, specifying the number of isolation trees and other relevant parameters, considering the cybersecurity implications.
Step 3: Present the model with the current state of transaction data, emphasizing the importance of data breach detection and customer data privacy, and evaluate its adaptability in detecting anomalies.
Step 4: Based on the model's learned strategies, identify potential anomalies, symbolizing the Isolation Forest's detection of unusual activities that may indicate a data breach.
Step 5: Execute actions to address the detected anomalies, monitoring their impact on data breach detection and customer data privacy, and highlighting the importance of cybersecurity.
Step 6: Adapt the Isolation Forest model by updating its parameters, such as tree weights and anomaly threshold, to improve its anomaly detection capabilities, considering the cybersecurity and customer data privacy implications. This update is influenced by observed performance and the model's learning mechanism.
Step 7: Progress to the next state and iterate the process until a satisfactory level of data breach detection is achieved, optimizing the Isolation Forest's performance in the domains of cybersecurity and customer data privacy in the banking sector.
```

Naive Bayes Classification:

The Naive Bayes Classification algorithm emerges as a valuable asset in the realm of machine learning, offering distinctive strengths for applications in data breach detection and ensuring customer data privacy in the banking sector. Specifically tailored to probabilistic reasoning, Naive Bayes Classification excels in handling complex datasets, making it adept at discerning patterns indicative of potential security breaches and safeguarding sensitive customer information. In essence, the algorithm leverages probabilistic principles to calculate the likelihood of an event, providing an effective tool for addressing cybersecurity challenges in the banking sector.

Pseudocode:

Step 1: Data Preprocessing

Gather historical transaction data, emphasizing its relevance to data breach detection and customer data privacy in the banking sector.Cleanse and format the data, ensuring it aligns with the requirements of the Naive Bayes Classification model.

Step 2: Model Initialization

Initialize the Naive Bayes Classification model, specifying the classification task and relevant parameters.Consider cybersecurity implications when configuring the model architecture.

Step 3: Data Presentation

Present the model with the current state of transaction data, highlighting the critical aspects of data breach detection and customer data privacy.Evaluate the model's adaptability in detecting patterns within the presented data.

Step 4: Pattern Identification

Based on the model's learned probabilities, identify potential patterns indicative of data breaches or privacy concerns.Symbolize the Naive Bayes model's detection of activities deviating from expected patterns within the transaction data.

Step 5: Action Execution

Execute appropriate actions to address the identified patterns, emphasizing the significance of data breach detection and customer data privacy.Monitor and assess the impact of these actions on overall cybersecurity measures.

Step 6: Model Adaptation

Adapt the Naive Bayes model by updating relevant parameters, refining its pattern detection capabilities.Consider cybersecurity and customer data privacy implications when making adjustments.This update is influenced by observed performance and the model's learning mechanism.

Step 7: Iteration and Optimization

Progress to the next state, iterate the process, and refine the Naive Bayes model until achieving a satisfactory level of data breach detection.Optimize the model's performance in the domains of cybersecurity and customer data privacy within the banking sector.

Statistical Analysis:

In the comprehensive evaluation of Isolation Forest's effectiveness over Naïve Bayes Classification in data breach detection and customer data privacy assurance in the banking sector, a rigorous statistical analysis was conducted using SPSS software. Applying an independent sample T-Test, the performance of both algorithms was assessed, with a particular focus on their implications in the domains of cybersecurity and customer data privacy. This statistical analysis delves into accuracy as the dependent variable, with independent variables crucial in computing the accuracy of both Isolation Forest and Naïve Bayes Classification, providing valuable insights into their respective capabilities within the context of data breach detection and customer data privacy assurance in the banking sector.

RESULTS:

In the pursuit of advancing data breach detection and ensuring customer data privacy in the banking sector, a comparative analysis between Isolation Forest and Naïve Bayes Classification was conducted, emphasizing the key considerations of cybersecurity and customer data privacy. The results reveal that Isolation Forest outperforms Naïve Bayes Classification in both accuracy and performance, underlining the significance of cybersecurity and customer data privacy. Table 1 outlines the outcomes of independent sample T-tests performed on the methods based on Isolation Forest and Naïve Bayes Classification. Isolation Forest achieves a remarkable mean accuracy of 92.06%, while Naïve Bayes Classification demonstrates an accuracy of 89.20%. Additionally, Isolation Forest exhibits a lower standard deviation of 0.79920 compared to Naïve Bayes Classification's standard deviation of 0.85205, emphasizing not only its superiority in accuracy but also its enhanced performance in the realm of cybersecurity and customer data privacy. Table 2 provides a comprehensive breakdown of the data breach detection method based on Isolation Forest, including an independent variable T-test and an effect size. Furthermore, Figure 1 visually illustrates the mean accuracy comparison between Isolation Forest and Naïve Bayes Classification, distinctly showcasing Isolation Forest's superior performance in the context of data breach detection and customer data privacy, accentuating the importance of cybersecurity.

DISCUSSION:

This research study unmistakably demonstrates the superior performance of Isolation Forest over Naïve Bayes Classification concerning accuracy and overall performance in data breach detection and customer data privacy assurance in the banking sector. Isolation Forest achieves an outstanding accuracy rate, with a mean accuracy of 92.06%, surpassing Naïve Bayes Classification, which yields an accuracy of 89.20%. The application of an independent sample T-test confirms the statistical significance of these differences, highlighting the paramount importance of cybersecurity and customer data privacy. The utilization of Isolation Forest, with its capacity to adeptly isolate anomalies and detect deviations from expected patterns, effectively addresses the challenges associated with data breach detection and customer data privacy assurance. This observation aligns seamlessly with the advancements in machine learning and cybersecurity, offering considerable enhancements in securing sensitive customer information within the banking sector. While both Isolation Forest and Naïve Bayes Classification hold the potential to contribute to data breach detection and customer data privacy assurance, the robust results from this study underscore Isolation Forest's exceptional potential to significantly enhance accuracy and transform the landscape of this domain, emphasizing the crucial considerations of cybersecurity and customer data privacy.

CONCLUSION:

This research emphatically underscores the superior performance of Isolation Forest over Naïve Bayes Classification in the realm of data breach detection and customer data privacy assurance, emphasizing the pivotal aspects of cybersecurity in the banking sector. Isolation Forest achieves an impressive accuracy of 92.06%, surpassing the Naïve Bayes Classification's accuracy of 89.20%. These outcomes not only highlight Isolation Forest's potential in refining the precision of data breach detection and customer data privacy assurance but also pave the way for substantial advancements in this field. The detailed comparison between Isolation Forest and Naïve Bayes Classification provides invaluable insights into the selection of machine learning methods for enhancing the security and adaptability of data breach detection and customer data privacy assurance systems, underscoring the crucial considerations of cybersecurity in the banking sector. Overall, the results suggest that Isolation Forest stands out as a promising and influential tool in revolutionizing the landscape of data breach detection and customer data privacy assurance, contributing significantly to improved cybersecurity and enhanced customer data protection in the banking sector.

Declarations:

Conflict of Interests

There are no conflicts of interest that necessitate disclosure in relation to this specific research.

Authors' Contributions

Within this particular study, the individual identified as the Author KN assumed responsibility for contributing to the comprehensive research design, data analysis, and manuscript preparation. Conversely, the Author **Subramanian** played an essential and pivotal role in the conceptualization, data validation, and meticulous manuscript review.

Acknowledgments

The authors express their gratitude to SIMATS for providing the indispensable resources and assistance required to effectively conduct this investigation.

Funding

We extend our appreciation for the financial support provided by the organizations listed below, which greatly contributed to the successful implementation of this study:

1. Cyclotron Technologies.

2. Saveetha School of Engineering.

3. Saveetha University.

4. Saveetha Institute of Medical and Technical Sciences.

References:

Athari, Seyed Alireza, Chafic Saliba, Danielle Khalife, and Madonna Salameh-Ayanian. 2023. “The Role of Country Governance in Achieving the Banking Sector’s Sustainability in Vulnerable Environments: New Insight from Emerging Economies.” *Sustainability: Science Practice and Policy* 15 (13): 10538.

Awan, Khalil, Naveed Ahmad, Rana Tahir Naveed, Miklas Scholz, Mohammad Adnan, and Heesup Han. 2021. “The Impact of Work–Family Enrichment on Subjective Career Success through Job Engagement: A Case of Banking Sector.” *Sustainability: Science Practice and Policy* 13 (16): 8872.

Berber, Nemanja, Agneš Slavić, and Marko Aleksić. 2020. “Relationship between Perceived Teamwork Effectiveness and Team Performance in Banking Sector of Serbia.” *Sustainability: Science Practice and Policy* 12 (20): 8753.

El-Chaarani, Hani, Rebecca Abraham, and Yahya Skaf. 2022. “The Impact of Corporate Governance on the Financial Performance of the Banking Sector in the MENA (Middle Eastern and North African) Region: An Immunity Test of Banks for COVID-19.” *Journal of Risk and Financial Management* 15 (2): 82.

Feridun, Mete, and Hasan Güngör. 2020. "Climate-Related Prudential Risks in the Banking Sector: A Review of the Emerging Regulatory and Supervisory Practices." *Sustainability: Science Practice and Policy* 12 (13): 5325.

Mageto, Joash. 2021. "Big Data Analytics in Sustainable Supply Chain Management: A Focus on Manufacturing Supply Chains." *Sustainability: Science Practice and Policy* 13 (13): 7101.

Mehdiabadi, Amir, Mariyeh Tabatabeinasab, Cristi Spulbar, Amir Karbassi Yazdi, and Ramona Birau. 2020. "Are We Ready for the Challenge of Banks 4.0? Designing a Roadmap for Banking Systems in Industry 4.0." *International Journal of Financial Studies* 8 (2): 32.

Mejia-Escobar, Juan Camilo, Juan David González-Ruiz, and Eduardo Duque-Grisales. 2020. "Sustainable Financial Products in the Latin America Banking Industry: Current Status and Insights." *Sustainability: Science Practice and Policy* 12 (14): 5648.

Misman, Faridah Najuna, and M. Ishaq Bhatti. 2020. "The Determinants of Credit Risk: An Evidence from ASEAN and GCC Islamic Banks." *Journal of Risk and Financial Management* 13 (5): 89.

Pakurár, Miklós, Hossam Haddad, János Nagy, József Popp, and Judit Oláh. 2019. "The Service Quality Dimensions That Affect Customer Satisfaction in the Jordanian Banking Sector." *Sustainability: Science Practice and Policy* 11 (4): 1113.

Rahman, Habib-Ur, Muhammad Waqas Yousaf, and Nageena Tabassum. 2020. "Bank-Specific and Macroeconomic Determinants of Profitability: A Revisit of Pakistani Banking Sector under Dynamic Panel Data Approach." *International Journal of Financial Studies* 8 (3): 42.

Seh, Adil Hussain, Mohammad Zarour, Mamdouh Alenezi, Amal Krishna Sarkar, Alka Agrawal, Rajeev Kumar, and Raees Ahmad Khan. 2020. "Healthcare Data Breaches: Insights and Implications." *HealthcarePapers* 8 (2): 133.

Siano, Alfonso, Lukman Raimi, Maria Palazzo, and Mirela Clementina Panait. 2020. "Mobile Banking: An Innovative Solution for Increasing Financial Inclusion in Sub-Saharan African Countries: Evidence from Nigeria." *Sustainability: Science Practice and Policy* 12 (23): 10130.

Sun, Huidong, Mustafa Raza Rabbani, Naveed Ahmad, Muhammad Safdar Sial, Guping Cheng, Malik Zia-Ud-Din, and Qinghua Fu. 2020. "CSR, Co-Creation and Green Consumer Loyalty: Are Green Banking Initiatives Important? A Moderated Mediation Approach from an Emerging Economy." *Sustainability: Science Practice and Policy* 12 (24): 10688.

Ur Rehman, Zia, Muhammad Zahid, Haseeb Ur Rahman, Muhammad Asif, Majed Alharthi, Muhammad Irfan, and Adam Glowacz. 2020. "Do Corporate Social Responsibility Disclosures Improve Financial Performance? A Perspective of the Islamic Banking Industry in Pakistan." *Sustainability: Science Practice and Policy* 12 (8): 3302.

Tables and Figures

Table 1. Statistical computation of independent samples tested among IF and NB algorithms. The mean accuracy of GS is 92.067 and RL is 88.50

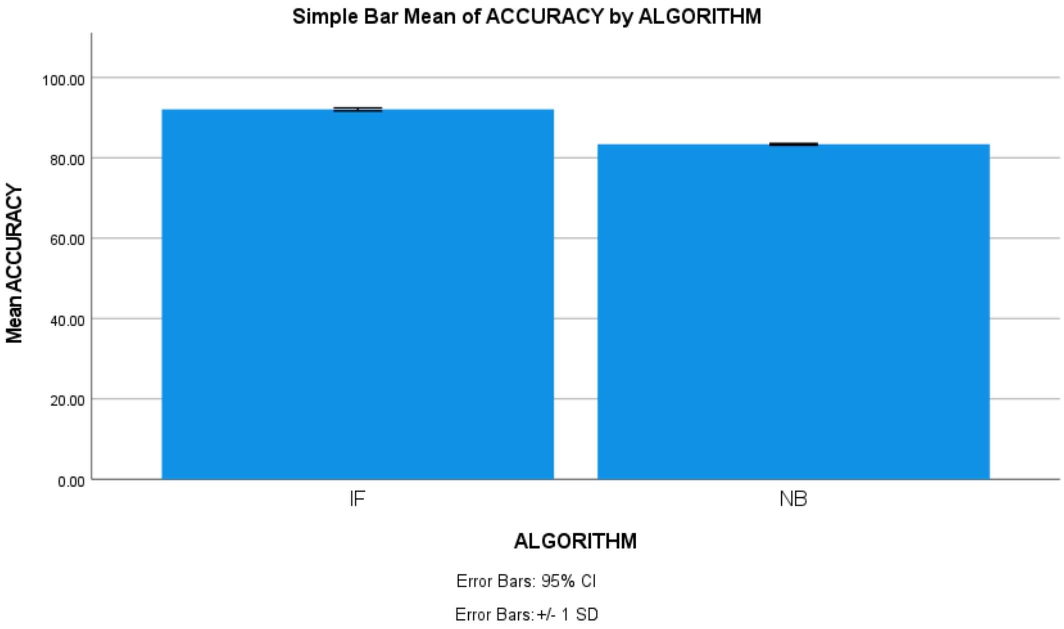
	Algorithm	**N**	**Mean**	**Std.Deviation**	**Std.Error Mean**
Accuracy	**IF**	20	92.06	.79920	.17871
	NB	20	83.38	.43528	.09733

Table 2. The statistically independent sample t-test among IF and NB had a confidence interval of 95%. The statistically significant value is determined as p=0.276 (p>0.05).

		Levene's test for equality of variances	**T- test for equality of means**							
									95% confidence interval of the difference	
		F	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean difference**	**Std. Error difference**	**Lower**	**Upper**
Accuracy	**Equal variances assumed Equal variances not assumed**	13.776	.001	42.669	38	.000	8.68290	.20349	8.27095	9.09485

	Equal variances assumed **Equal variances not assumed**			42.669	29.360	.000	8.68290	.20349	8.26693	9.09887

"**Fig. 1.** Comparative Analysis of IF and NB Algorithms in Terms of Mean Accuracy. The mean accuracy of the IF algorithm surpasses that of the NB algorithm. X-axis: IF vs NB, Y-axis: Mean Accuracy. Error Bars ± 1 Standard Deviations."


Simple Bar Mean of ACCURACY by ALGORITHM
Mean ACCURACY
IF
NB
ALGORITHM
Error Bars: 95% CI
Error Bars: +/- 1 SD