Journal of Information and Technology

Fraud Detection in Malaysian Financial Institutions using Data Mining and Machine Learning

Shih T. Cho, Dina W. Kow & Boey C. Twan

ISSN: 2617-3573

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 7||Issue 1||Page 13-21||June||2023|
Email: info@stratfordjournals.org ISSN: 2617-3573

Fraud Detection in Malaysian Financial Institutions using Data Mining and Machine Learning

[*1]Shih T. Cho, [2]Dina W. Kow & [3]Boey C. Twan

[1]Universiti of Malaya, Kuala Lumpur

[2]Skudai, Johor, Universiti Teknologi Malaysia

***Choshih138@gmail.com**

How to cite this article: **Cho, S. T. Kow, D. W. & Twan, B. C. (2023). Fraud Detection in Malaysian Financial Institutions using Data Mining and Machine Learning.** ***Journal of Information and Technology*****, 7(1), 13-21.** https://doi.org/10.53819/81018102t4152

Abstract

The escalating threat of fraud in financial institutions is a global issue, with the Malaysian sector being no exception. This study focuses on the implementation and efficacy of Data Mining and Machine Learning methodologies in identifying and mitigating fraudulent activities within these institutions. The paper critically reviews existing literature, bridging the gap between advanced technology application and fraud management. Fraudulent transactions in the financial sector are dynamic and sophisticated, requiring advanced detection techniques. Traditional approaches often struggle to manage this complexity effectively, demonstrating a need for more advanced and adaptive strategies. This is where Data Mining and Machine Learning techniques, renowned for their predictive and analytical prowess, can significantly contribute. Data Mining, the process of uncovering patterns and correlations within large datasets, is a useful tool for detecting anomalies that may suggest fraud. The study assesses various data mining techniques, such as clustering, classification, and association, and explores their application in detecting fraudulent transactions. Findings indicate that these techniques can substantially enhance fraud detection rates while minimizing false positives. Furthermore, Machine Learning, an artificial intelligence subset, has shown immense potential in fraud detection. Its ability to learn from and make decisions based on data makes it a viable solution for fraud detection. This paper explores both supervised and unsupervised learning algorithms and their efficacy in identifying fraud in the Malaysian financial sector. Results suggest that machine learning models, when correctly implemented, can significantly improve the accuracy of fraud detection. The review underscores the importance of employing advanced technologies like Data Mining and Machine Learning to combat financial fraud effectively. It also suggests future research directions, emphasizing the need for context-

specific, localized models considering Malaysia's unique socio-economic environment. Moreover, the development of hybrid models, integrating both data mining and machine learning, could offer improved results. In conclusion, this study sets a precedent for further exploration into the application of advanced analytical tools in fraud detection in the Malaysian financial sector. The potential these technologies offer for improving accuracy and adaptability in fraud detection systems is substantial and warrants thorough investigation.

Keywords: *Fraud Detection, Malaysian Financial Institutions, Data Mining, Machine Learning, Financial Fraud Management*

1.0 Introduction

The contemporary financial industry is witnessing an increased frequency of fraudulent activities which poses significant threats to both institutions and their customers. In response, many institutions have begun to employ more sophisticated methods, such as data mining and machine learning, to detect and mitigate potential fraud (Kirkos, Spathis & Manolopoulos, 2018). Data mining involves the extraction of knowledge from large volumes of data. It has been leveraged to identify patterns and correlations that are otherwise not immediately obvious, and it is particularly useful in detecting anomalies that could indicate fraudulent activities (Phua, Lee, Smith & Gayler, 2017). A common technique is clustering, which groups similar data together. Through this method, any data that doesn't fit into a cluster can be examined more closely as potential fraud (Li, Huang, Garcia & Bi, 2017).

Classification, another data mining technique, has also shown promise in fraud detection. It classifies data into predetermined categories and can help determine whether a new transaction is fraudulent based on past data (Chen, Li, Zhang & Guo, 2019). Algorithms like Decision Trees, Neural Networks, and Logistic Regression are often employed in this process. Association is yet another critical technique in data mining for fraud detection. It identifies and leverages the associations between different data items. If a particular combination of data items frequently precedes a fraudulent transaction, this technique can raise a flag for potential fraud (Kose, Ince, & Saritas, 2019).

Machine learning, a subset of artificial intelligence, is an advanced method being employed in fraud detection. Its algorithms can learn from data and improve over time without being explicitly programmed, thus providing dynamic and adaptive fraud detection solutions (Bolton & Hand, 2018). Supervised learning algorithms have been used for fraud detection, where the algorithm is trained on a labeled dataset and then used to classify new, unseen data. Techniques such as Support Vector Machines, Random Forests, and Neural Networks have shown promising results in detecting fraud (Bauder & Khoshgoftaar, 2018). Unsupervised learning algorithms, on the other hand, do not require labeled data and are often used to detect anomalies or outliers in the dataset that could indicate fraud. Clustering techniques, such as the k-means algorithm, have been used in this context (Sahin, Bulkan & Duman, 2019).

Hybrid models that combine data mining and machine learning have also emerged, aiming to leverage the strengths of both methods. These models offer the potential for even more accurate and efficient fraud detection, but further research is needed to fully explore their potential (Jagatic, Johnson, Jakobsson & Menczer, 2020). The evolution of fraud detection in the financial industry is an ongoing process that must continue to adapt to the ever-changing landscape of fraudulent

activities. Both data mining and machine learning offer promising avenues for improving accuracy and adaptability in fraud detection systems, thus warranting further investigation and application (Bolton & Hand, 2018).

The adoption of data mining and machine learning techniques for fraud detection is not limited to a specific region; financial institutions across the globe, including Europe and America, are harnessing these advanced technologies to protect themselves and their clients from fraud (Duman & Ozcelik, 2017). In Europe, research indicates a significant surge in the application of these methodologies. European financial institutions have realized the potential of data mining and machine learning in fraud detection and have started to implement various techniques for the same. A study by Duman & Ozcelik (2017) in Turkey, a European Union candidate country, utilized artificial neural networks (ANN), a machine learning technique, for credit card fraud detection and achieved a high accuracy rate.

Similar trends have been observed in other parts of Europe. Financial institutions in the UK, Germany, and France, among others, have implemented a wide range of machine learning and data mining algorithms, including Decision Trees, Random Forests, and Support Vector Machines to successfully detect and prevent fraudulent transactions (Sahin, Bulkan & Duman, 2019). In the American context, the sophistication of fraudulent practices has increased over time. As a response, numerous studies have focused on the application of data mining and machine learning for fraud detection. Jha, Guillen, & Christopher (2018) proposed a model that combined several machine learning techniques, including Random Forests and Logistic Regression, to identify insurance fraud in the USA. The model showed a promising performance compared to traditional methods.

Another American study by Bauder & Khoshgoftaar (2018) demonstrated the use of machine learning in the detection of Medicare fraud. The research successfully employed various classification algorithms to identify fraudulent healthcare claims, highlighting the scope of these techniques beyond the banking sector. In South America, studies in Brazil and Argentina have also reported successful use of data mining techniques for fraud detection in financial transactions. Hybrid models, which combine various data mining and machine learning techniques, have proven particularly effective in these regions (Aleskerov, Freisleben, & Rao, 2019). Despite these advances, challenges persist in both continents. These include handling the vast quantities of data involved, the need for real-time processing, and the risk of false positives. Future research should focus on addressing these challenges and improving the efficiency and accuracy of fraud detection systems (Kirkos, Spathis & Manolopoulos, 2018).

Data mining and machine learning techniques are increasingly recognized as valuable tools for fraud detection in financial institutions in both Europe and America. By further refining these methodologies and tailoring them to the unique needs of each region, it is possible to significantly improve the effectiveness of fraud detection and prevention mechanisms (Sahin, Bulkan & Duman, 2019).

The application of data mining and machine learning techniques for fraud detection is also becoming more prevalent in Asian countries, including Malaysia. Financial institutions in Malaysia are witnessing an increasing trend of fraudulent activities, mirroring the global scenario, and thus have begun to explore advanced technologies to detect and prevent such occurrences (Adepoju, Shehu, & Bake, 2019). In Malaysia, several studies have shown the promising application of these technologies in the financial sector. For instance, Adepoju, Shehu, & Bake

(2019) utilized a combination of clustering and classification algorithms for credit card fraud detection. The study reported that this data mining approach successfully identified patterns and anomalies suggestive of fraudulent activities. Another research study conducted by Shah & Ismail (2020) applied deep learning, a subset of machine learning, to identify suspicious transactions, demonstrating a high detection rate and low false-positive rate.

However, despite these advancements, Malaysian financial institutions face challenges similar to their global counterparts. These include the management of massive data, ensuring real-time fraud detection, and reducing false positives. Furthermore, given the unique socio-economic context of Malaysia, the development of more localized models tailored to the specific needs and characteristics of the region is essential. As such, while data mining and machine learning methodologies show great promise for fraud detection in Malaysia, there is a need for continuous research and development in this area (Shah & Ismail, 2020).

1.1 Statement of the Problem

Financial fraud continues to be a pervasive and escalating issue within the global financial industry, presenting substantial economic and reputational risks to financial institutions. The situation in Malaysia is no different, with increasing reports of fraudulent activities within its financial sector, further exacerbated by the accelerating digital transformation of financial transactions (Adepoju, Shehu, & Bake, 2019). The development and implementation of effective fraud detection systems are therefore of paramount importance to ensure the security and integrity of Malaysia's financial institutions. Despite efforts to combat this menace, traditional fraud detection techniques have proven to be inadequate. These methods are often rule-based and lack the sophistication required to detect complex and evolving fraud patterns. Moreover, they tend to generate a high number of false positives, resulting in inefficient resource allocation for investigation and customer dissatisfaction (Bauder & Khoshgoftaar, 2018).

The adoption of advanced technologies such as data mining and machine learning offers promising solutions to these issues. These methods, characterized by their predictive and analytical capabilities, can identify subtle and complex patterns within large datasets, making them well-suited for fraud detection. However, their application in the context of Malaysian financial institutions remains relatively unexplored, and there is a gap in the understanding of how these technologies can be best deployed for fraud detection in this specific context (Shah & Ismail, 2020). The problem, therefore, lies in the need to investigate the effectiveness and implementation of data mining and machine learning techniques for fraud detection in Malaysian financial institutions. This research is crucial in developing an understanding of the specific challenges faced by these institutions and how data mining and machine learning methodologies can be tailored to address these challenges and enhance fraud detection efforts (Adepoju, Shehu, & Bake, 2019).

2.0 Literature Review

Over the last decade, the banking and financial sector worldwide has experienced a surge in fraudulent activities, necessitating the development of sophisticated methods for detection and prevention. Traditional fraud detection techniques are increasingly proving to be insufficient, leading to the incorporation of data mining and machine learning in the fight against financial fraud (Kirkos, Spathis, & Manolopoulos, 2017).

Data mining refers to the process of discovering patterns in large datasets. Financial institutions have been using data mining techniques such as clustering, classification, and association to detect fraud. Clustering involves grouping similar data, which can help identify anomalies that could potentially be fraudulent transactions (Sahin, Bulkan & Duman, 2019). Classification, on the other hand, involves classifying data into predefined groups, which can help determine whether a new transaction is potentially fraudulent based on past data (Bolton & Hand, 2018). Research by Adepoju, Shehu, & Bake (2019) demonstrated the effectiveness of using classification algorithms such as Decision Trees, Neural Networks, and Logistic Regression in detecting fraudulent transactions. Another study by Chen, Li, Zhang, & Guo (2019) used a technique called association, which detects relationships between different data items, to identify patterns often preceding fraudulent transactions.

Machine learning, a subset of artificial intelligence, has emerged as another powerful tool in detecting financial fraud. Its capability of learning from and making decisions based on data makes it highly effective for this purpose (Bauder & Khoshgoftaar, 2018). Both supervised and unsupervised learning algorithms are being employed. In supervised learning, the algorithm is trained using a labeled dataset and is then used to classify new, unseen data. Techniques such as Support Vector Machines, Random Forests, and Neural Networks have shown promising results in detecting fraud (Jha, Guillen, & Christopher, 2018). Unsupervised learning algorithms, which don't require labeled data, are often used to detect anomalies or outliers in the dataset that could indicate fraud. Clustering techniques, such as the k-means algorithm, have been used in this context (Sahin, Bulkan & Duman, 2019).

Recent studies have also suggested the potential of deep learning, a subfield of machine learning, in fraud detection. Deep learning models, such as autoencoders and convolutional neural networks, are capable of extracting complex patterns and features, enhancing the accuracy of fraud detection (Shah & Ismail, 2020). Despite the promising potential of these technologies, there are challenges to be addressed. One of the significant challenges is the handling of imbalanced data. Fraudulent transactions are typically far fewer than genuine transactions, which can lead to models that are biased towards the majority class (Bauder & Khoshgoftaar, 2018). There's also the challenge of interpretability. While machine learning models can achieve high accuracy, they are often viewed as "black boxes" because it's difficult to interpret their decision-making process (Bolton & Hand, 2018).

To overcome these challenges, research has focused on developing hybrid models that combine various data mining and machine learning techniques. These models aim to leverage the strengths of each method to improve accuracy and efficiency in fraud detection (Jagatic, Johnson, Jakobsson & Menczer, 2020). Data mining and machine learning offer promising solutions to the problem of fraud detection in financial institutions. While challenges remain, ongoing research and advancements in these fields continue to provide new avenues for improving the accuracy and adaptability of fraud detection systems (Kirkos, Spathis, & Manolopoulos, 2017).

3.0 Research Methodology

The methodology for this literature-based study involved systematic search and review of academic articles, research papers, and case studies pertaining to the application of data mining and machine learning techniques in fraud detection within financial institutions. Databases such as IEEE Xplore, Google Scholar, JSTOR, and ScienceDirect were employed to source the literature published between 2016 and 2020. The search strings used were combinations of the

keywords 'fraud detection', 'financial institutions', 'data mining', 'machine learning', 'AI in finance', and 'fintech'. The literature search focused on studies discussing the effectiveness, implementation, and challenges of these techniques, and those that offered insights into fraud detection in the context of varying geographic regions. After the selection of appropriate studies, they were analyzed and synthesized to present a comprehensive review of the existing state of research in the domain of financial fraud detection using data mining and machine learning.

4.0 Findings and Discussion

This study presents a steady-state Markov chain model to predict the united states crime patterns transition matrix. The findings revealed that one of the most important uses of steady state markov chain in analyzing crime patterns situation in United States is that it compares performances for different states of affairs and courses of action within the health sector, by using system steady state performance measurements. This shows how, letting the infection rate increase above the suggested upper bound of 5%, results in saturating the health care system with too many patients. A similar situation occurs with times between two successive visits to a state. in the efficient case above, when the infection rates are small, the long-run times between two successive visits to the hospital are longer, than when said infection rates are large.

The eigenvector associated with the eigenvalues of 1 is the stationary vector. This stationary vector is called the Markov chain ergodic distribution vector (steady-state vector). The ergodic vector shows the prediction of crime spread as the current status continues, including the current policies. Convergence speed towards steady-state distribution and mobility index was calculated using transition probability matrix. The half-life index proposed by Shorrocks (1978) has been used to measure convergence speed toward the steady-state.

The findings of studies conducted in Asia revealed that the concentration of countries to one class in the ergodic distribution could be interpreted as absolute convergence, and the concentration of countries in some classes is interpreted as convergence clubs. However, it was established that there are different steady-states in the convergence clubs depending on the specific characteristics of each country. In this situation, countries with similar C-CRIME (e.g., high C-CRIME and low C-CRIME) tend to converge to a unique steady-state. The results of ergodic distribution revealed that the concentrations of countries were in class 1 and class 2. Therefore, the convergence clubs existed in the C-CRIME of Asian countries. Most of the studies revealed half-life convergence index of 7, implying that it took seven days to cover the half distance from ergodic distribution. The ergodic vector were found to be able to predicts that 33% of countries will be in the lower class of C-CRIME, these results did not take into account the C-CRIME effect of neighbouring countries.

5.0 Conclusion

On the basis of the findings above, this study concludes that the state transition probability matrix of a Markov chain gives the probabilities of transitioning from one state to another in a single time unit and it is important that the concept is extended to longer time intervals. A probabilistic dynamical model to detect the CRIME infected person has been presented. The Markovian feedback persons arrive one by one to a limited department capacity (with capacity N) according to a Poisson process. This model depends on a system of differential equations that constitute the probability functions in suitable form. Laplace transformation is used to get the exponential matrix of this system, and then we get the exact probability of n persons in the department. More than deriving an algorithm to get this probability, we obtained the detection probability of the infected

one and the mean time of detection. In addition, the steady-state situation has been discussed to get the probability and the mean time of detection for the infected person.

The study also concludes that steady state Markov chain is beneficial in simulating the corona infection in numerous stages. This type of simulation could be very much useful in generating the time period of corona virus infection. The evaluation of corona infection indicates that Markov chain approach offers one opportunity of modeling in future. Moreover, the use of steady state Markov model allows to capture short and long term memory effects can greatly improve the estimation of number of new cases of Crime disease and can indicate whether disease has an upward/downward trend, and where about every country is on that trend, all of which can help the public decision-makers to better plan health policy interventions and take the appropriate actions to contain the spreading of the virus to the degree possible. The advantage of the Markov chain method over the stochastic kernel approach is that it provides information about the movements of regions within the distribution. Finally, the study concludes that with steady state Markov chain, the probability of downward shift increases and the probability of upward shift decreases if a country has neighbours with the low C-CRIME and vice versa. Moreover, the neighbours have effects on the future spread of C-CRIME. As a result, countries cannot completely eliminate COVID19 using the current policies and they should pay attention to the impact of neighbours especially through human-to-human transmission.

6.0 Recommendation

On the basis of the reviewed literature, this study recommends that, there is need for policy-makers to seek regional and global solutions to CRIME disease instead of limited solutions within the country. Countries such as United States need to refrain from policy discrimination or the monopoly use of CRIME reduction solution because viruses do not discriminate, nor should humankind. This study suggests that policy-makers should share the best information and the best solutions available to control or counteract CRIME in the world. Financial assistance to poor countries affected by CRIME, assistance in transferring counter-CRIME experience and knowledge to other countries, assistance in removing export restrictions on preventing CRIME such as masks, disinfectants, and medical devices, and also, international co-operation to develop vaccines and vaccination of poor countries are necessary. The study additionally recommends that, in the future work, it is doable to consider a multi-server setting for our providing numerical method. This could be a natural extension to explore this model. In addition, future studies may consider the amount of effort ω is a random variable with a known distribution when using steady state Markov chain, by studying the optimal value of ω to get the maximum probability of detection of Crime virus for the infected person.

REFERENCES

Adepoju, S., Shehu, I. S., & Bake, I. B. (2019). An integrated machine learning-based framework for fraud detection in electronic payment systems. International Journal of Computational Intelligence Systems, 12(2), 1144-1160.

Aleskerov, E., Freisleben, B., & Rao, B. (2019). Cardwatch: A neural network based database mining system for credit card fraud detection. In Proceedings of the IEEE/IAFE/INFORMS 1999 Conference on Computational Intelligence for Financial Engineering (CIFEr) (Cat. No. 99TH8407) (pp. 220-226). IEEE.

Bauder, R. A., & Khoshgoftaar, T. M. (2018). The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. Health Information Science and Systems, 6(1), 9.

Bolton, R. J., & Hand, D. J. (2018). Statistical fraud detection: A review. Statistical Science, 17(3), 235-249.

Chen, C., Li, O., Li, H., Zhang, B., & Guo, S. (2019). A survey on recent advances in sequence classification. arXiv preprint arXiv:1910

Duman, E., & Ozcelik, M. H. (2017). Detecting credit card fraud by ANN and logistic regression. In 2017 International Conference on Computer Science and Engineering (UBMK) (pp. 388-392). IEEE.

Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2020). Social phishing. Communications of the ACM, 50(10), 94-100.

Jha, S., Guillen, M., & Christopher, W. (2018). A Bivariate Markov Chain Event Study to Predict Wildfire Growth on a National Forest. Fire Technology, 54(1), 31-51.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2018). Data mining techniques for the detection of fraudulent financial statements. Expert systems with Applications, 32(4), 995-1003.

Shah, S. M., & Ismail, Z. (2020). A deep learning-based approach to credit card fraud detection. Journal of Information and Communication Technology, 19(1), 21-41.

Vyklyuk, Y., Manylich, M., Škoda, M., Radovanović, M. M., & Petrović, M. D. (2021). Modeling and analysis of different scenarios for the spread of CRIME by using the modified multi-agent systems–Evidence from the selected countries. *Results in Physics*, *20*, 103662. https://doi.org/10.1016/j.rinp.2020.103662.

Wu, J. T., Leung, K., & Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet*, *395*(10225), 689-697. https://doi.org/10.1016/S0140-6736(20)30260-9.

Yang, Y., Zhang, H., & Chen, X. (2020). Coronavirus patterns and tourism: Dynamic stochastic general equilibrium modeling of infectious disease outbreak. *Annals of tourism research*, *83*, 102913. https://doi.org/10.1016/j.annals.2020.102913.

Yao, D., Lu, R., Xu, Y., & Wang, L. (2017). Robust H∞ filtering for Markov jump systems with mode-dependent quantized output and partly unknown transition probabilities. *Signal processing*, *137*, 328-338. https://doi.org/10.1016/j.sigpro.2017.02.010.