An Analytic Approach of Bank Account Fraudulent Prediction

CSE6242 Data and Visual Analytics

Final Report

Team 22

Team Members:

Chin-Hsien Tsai, Yanhui Li, Mina Liu, Cheng Feng, Xiaofan Jiao, Xin Chen

April 19th, 2024

Professor: Dr. Duen Horng

I. INTRODUCTION

In the realm of financial security, detecting and preventing fraudulent activities in bank account transactions is paramount [4]. Leveraging the BAF dataset for its advanced privacy measures, our study aims to improve fraud detection accuracy in bank transactions by addressing gaps in current models[2] [9][17][18]. We utilized Python for both modeling and interactive visualizations, enhancing our understanding and analysis of financial fraud complexities.

II. PROBLEM DEFINITION

In tackling the complex realm of bank account fraud, our project, inspired by other studies [10] [16][19], employed advanced machine learning algorithms to enhance detection accuracy and adaptability. Three key innovations drive our approach: refined data transformation techniques, polynomial feature introduction, and strategic clustering using K-means and Gaussian Mixture models. These strategies, complemented by interactive 3D visualization, offer a dynamic framework to identify fraudulent behavior effectively. By translating intricate analytical outcomes into actionable measures, our methodology equips financial institutions with robust tools to combat fraud, marking a significant advancement in bank account security.

III. SURVEY

The current landscape of fraud detection in banking predominantly relies on machine learning models coupled with rule-based systems for robust defense mechanisms [1]. While other studies [5][11] [13][15] have explored the detection of fraudulent transactions using neural networks, classification algorithms, logistic regression, and k-fold machine learning techniques, gaps remain in adapting these methodologies to address the dynamic nature of fraud. Perttilä [14] underscores the potential of integrating text mining with existing models to overcome limitations associated with traditional data sources and methods. However, there is a notable lack of comprehensive understanding regarding how text mining can effectively adapt to evolving tactics employed by fraudsters. Furthermore, Asmar and Tuqan [3] provided insights into machine learning applications in digital bank cybersecurity, although their focus did not delve deeply into specific machine learning techniques tailored for fraud detection, leaving ample room for our project to explore and expand upon these methodologies. Therefore, the problem at hand requires more robust and adaptive fraud detection techniques. These techniques should effectively address the dynamic nature of fraudulent activities, while leveraging the strengths of machine learning models in the context of banking cybersecurity.

IV. PROPOSED METHOD

1. Variable Selection

Inspired by prior research [12][15], our project employs a meticulous approach to enhance the accuracy and adaptability of fraud detection models in banking transactions. We commence by refining our dataset through rigorous data cleansing, mitigating multicollinearity through correlation analysis, and reducing dimensionality via Principal Component Analysis (PCA). These preprocessing steps are pivotal in optimizing the dataset for subsequent analysis.

2. Classification Methods

Leveraging classification methods, our analysis presents an innovative approach to identify any band fraudulent transactions. Our **Base Method** laid the groundwork by fitting scaled data to various machine learning models. Through the **Transformation Method,** the first approach involves scaling and transformation to improve normalization and linear relationships, as illustrated in Figures 1 and 2. This step establishes a performance baseline which helps prepare for the advanced machine learning algorithms later on. Additionally, our model boosts model accuracy by exploring data transformation techniques like Box-Cox and logarithmic transformations.

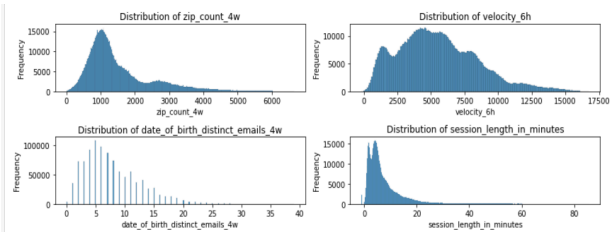

Figure 1: features before transformation

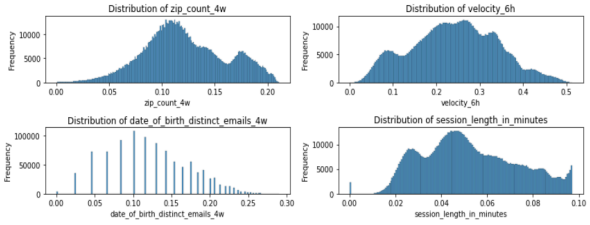

Figure 2: features after transformation

The Polynomial Method added another layer of depth, introducing polynomial and squared variables that allowed our models to detect complex, non-linear patterns and subtle nuances within the financial data. These are key indicators of potential fraud. These three methods collectively sharpen our analytical toolkit, providing an effective strategy for fraud detection in the banking industry.

3. Clustering Methods

In the clustering process, we focused on using numerical columns to group similar records because they provide precise quantitative data. We used Principal Component Analysis with three components to simplify our features. This helped us select the most important aspects of the data, reducing information loss and better preparing the data for other clustering models.

4. Algorithms

Logistic Regression: We chose his model as it is great for binary outcomes, like fraud detection. Logistic regression model forecasts the outcome of a dependent variable by examining how it relates to one or more independent variables already in existence[20].

Random Forest: We used this model as it is strong and doesn't overfit easily. Random Forest is our choice; it constitutes an ensemble of decision trees, typically ranging from 500 to 1000 trees, employed for predictive analytics and classification tasks[21].

Naive Bayes: We decided to use this model because it is fast and performs well for large datasets, like our dataset for fraud detection, despite assuming feature independence[23].

Ensemble Learning: We decided to use ensemble learning which allows us to combine multiple models' strengths to improve overall performance. It reduces individual biases and errors, potentially offering more accurate predictions[24].

In conjunction with these models, our Jupyter notebook provides a thorough comparative analysis. We encompassed data preparation with SMOTE for class balance and included detailed model training codes to present the process. We provided evaluation across key metrics such as ROC curve analysis with AUC, according to Chan.C [22], an ROC curve is a visual representation utilized to illustrate the diagnostic performance of binary classifiers. This exhaustive approach allowed for a nuanced understanding of each model's capabilities and their collective contribution to effectively identifying fraudulent activities in our dataset.

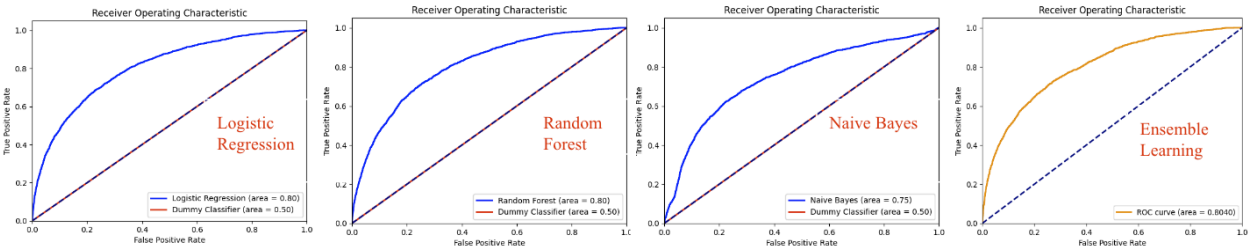

Figure 3: Base Method ROC curves

In our clustering strategy, we initially applied the **K-means** algorithm, renowned for its straightforward and efficient approach to grouping data. For both the original dataset and its

PCA-transformed counterpart, K-means clustering was executed with various numbers of clusters, ranging from 2 to 10. To determine the most suitable number of clusters, we employed the Elbow method, analyzing the within-cluster sum of squares (WCSS) against the number of clusters to pinpoint the "elbow point," where an increase in clusters no longer yields significant gains in variance explained.

Simultaneously, we explored the **Gaussian Mixture Model (GMM)** for its adeptness in handling diverse cluster shapes and sizes. GMM clustering was similarly assessed with different cluster counts, utilizing the Bayesian Information Criterion (BIC) for both the original and PCA-transformed data to identify the optimal configuration. This combination of K-means with the Elbow method and GMM with BIC provided a comprehensive view of potential clustering patterns, which is crucial for detecting nuanced fraudulent activities. Enhanced by interactive visualizations, our approach allowed for dynamic exploration of clusters, offering a more transparent analysis of clustering outcomes in the context of fraud detection.

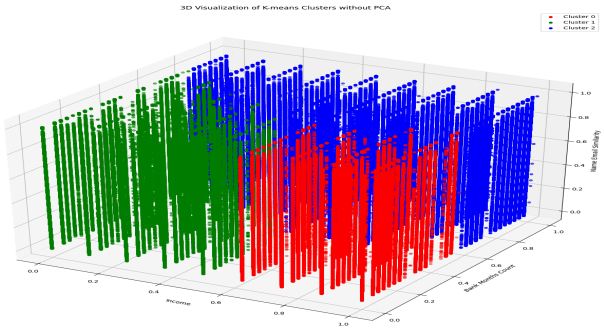

Figure 4: Visualization of K-Means Models

V. EVALUATION

In our evaluation against the imbalanced BAF dataset, we employed Synthetic Minority Over sampling Technique (SMOTE) during training to bolster the reliability of our models in detecting less frequent fraudulent cases[10]. In the **Base Method**, Logistic Regression achieved an accuracy of 74.46%, contrasted with the Dummy Classifier's 98.89%. Despite a lower accuracy, Logistic Regression had a notable recall of 70.71%, highlighting its strength in identifying true positives. Random Forest showed an accuracy of 79.24%, with a recall of 65.87%, while Naive Bayes stood out with the highest recall at 71.25%. Ensemble Learning followed closely, with a 71.34% recall rate. These results reveal the models' potential, but also emphasize the need for balance in precision and recall. The **Transformation Method** saw improvements: Logistic Regression's recall was 70.48% with a slightly higher accuracy of 75.28%. Random Forest's recall was 65.51%, and its accuracy edged up to 79.86%. Naive Bayes and Ensemble Learning demonstrated similar trends. This method enhanced true positive detection, a crucial factor in fraud detection. Using the **Polynomial Method**, Logistic Regression balanced accuracy (76.83%) and recall (72.20%) effectively. Random Forest's accuracy was slightly higher at 77.39%, with a recall of 68.49%. Naive Bayes, while achieving a high recall of 78.75%, did so at the cost of accuracy. Ensemble Learning's recall topped at 74.01%. These models, collectively, outperformed the baseline in discerning fraudulent cases. The inclusion of the Dummy Classifier serves as a baseline to illustrate the meaningfulness of our models' predictions[6]. Its high accuracy but zero recall and precision reflect its inability to correctly identify fraudulent cases, emphasizing the necessity of our more sophisticated approaches. These comparative insights guide us in refining our models for optimal fraud detection in banking transactions.

Summarize Table (Base, Transformation, Polynomial)

Method	Logistic Regression				Random Forest				Naive Bayes				Ensemble			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Base	0.7446	0.0301	0.7071	0.0577	0.7924	0.0345	0.6587	0.0656	0.6716	0.0237	0.7125	0.0458	0.7484	0.0308	0.7134	0.059
Method1 - Transformation	0.7528	0.031	0.7048	0.0593	0.7986	0.0354	0.6551	0.0671	0.682	0.0241	0.7034	0.0466	0.7576	0.0316	0.7061	0.0605
Method2 - Polynomial	0.7683	0.0338	0.722	0.0645	0.7739	0.0329	0.6849	0.0628	0.5267	0.0182	0.7875	0.0355	0.745	0.0314	0.7401	0.0603

Figure 5: Summarize Table

Our feature importance analysis revealed distinct insights across different machine learning models. For the Ensemble Learning model, the feature 'income' showed the most significant impact, implying its strong predictive power for fraud detection. In the Logistic Regression model, 'credit_risk_score' surfaced as a critical indicator, highlighting its relevance in assessing the likelihood of fraud. When examining the Random Forest model, both 'income' and 'credit_risk_score' maintained their prominence, but additional features like 'current_address_months_count' also emerged as influential, illustrating the varied importance of features depending on the model used.

In evaluating the cohesion and separation of the clusters generated by our algorithms, we measured the Average Silhouette Score—a metric that gauges how similar an object is to its own cluster compared to other clusters [7][8]. K-Means without PCA achieved the highest score of 0.60838, suggesting well-separated and cohesive clustering in comparison to K-Means with PCA, which scored slightly lower at 0.60466. On the other hand, the Gaussian Mixture Model (GMM) with PCA exhibited a score of 0.339224, indicating a less distinct clustering. However, GMM without PCA improved significantly to 0.44215, reflecting better-defined cluster structures. These scores are instrumental in our selection of the clustering method, as they provide insight into the intrinsic structure of our data and the effectiveness of our dimensionality reduction through PCA.

	Model	Average Silhouette Score
0	K-Means with PCA	0.60466
1	K-Means without PCA	0.608038
2	Gaussian Mixture Model with PCA	0.339224
3	Gaussian Mixture Model without PCA	0.44215

Figure 6: Comparison of Clusters

VI. VISUALIZATION

1. *Interactive ROC, AUC*

To enhance the utility and clarity of our model comparisons, our research incorporated interactive ROC AUC plots, which allow users to selectively view and analyze the performance of different classification models. By employing `plotly.express` and `plotly.graph_objects`, coupled with `ipywidgets`, users can interactively select one or multiple models to display on the ROC curve. Each model is delineated by a distinct color for easy comparison, and the AUC values are annotated directly on the plot to provide immediate insight into each model's predictive power. Below the ROC curves, a comprehensive table summarizes key performance metrics like accuracy, precision, recall, F1 score, and AUC, enabling a side-by-side evaluation of the models. This level of interactivity is crucial as it not only engages users in the analytical process but also facilitates a deeper understanding of the models' operational characteristics. These interactive plots ensure that the insights derived from our analysis are both accessible and actionable for all stakeholders involved in bank fraud detection.

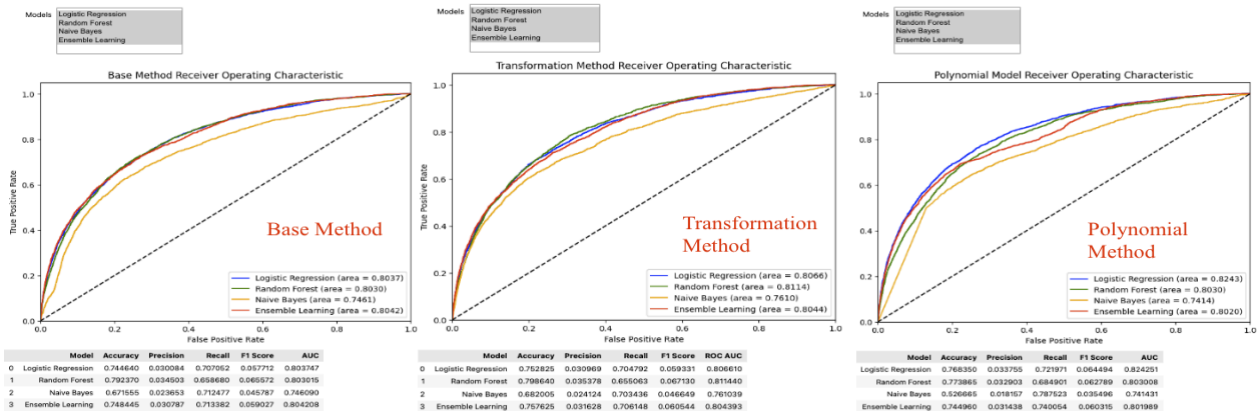

Figure 7:Interactive ROC, AUC Plots

2. *Interactive Important Features*

In our research, we utilized `pandas` for data wrangling, `seaborn` for rich visualization, and `matplotlib` for enhanced plotting functionality. We concentrated our analysis on Logistic Regression and Random Forest models for all three methods. The latter was particularly significant, given its ability to estimate feature importance inherently. Naive Bayes was excluded from this visualization as it lacks an inherent method for calculating feature importance, and ensemble models present complexities that make direct interpretations of feature importance less transparent. We focused on displaying the top 20 features for the polynomial method rather than the full set to ensure conciseness in our visual communications. Such interactive plots can aid stakeholders in rapidly identifying key factors that the Random Forest model deemed most predictive of fraudulent behavior. The bar plot illustrates this succinctly, with each bar's length representing the calculated importance of the feature, allowing for immediate visual assessment and comparison. Interactive widgets further empowered users to tailor the visualization to their specific interests, offering a dynamic and user-driven exploration of model insights.

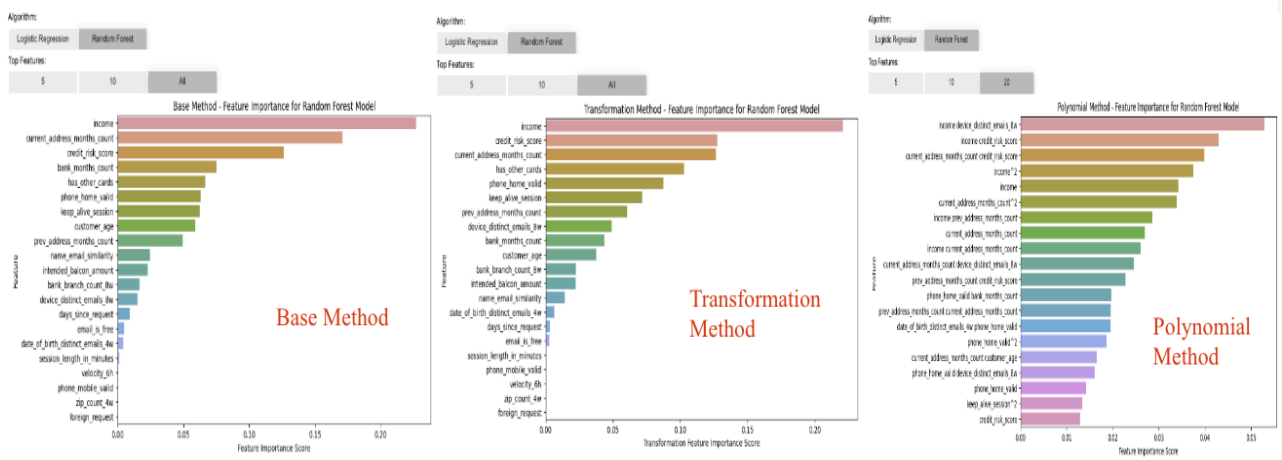

Figure 8:Interactive Important Features

3. *Interactive Confusion Matrix*

Our study employed the confusion matrix as a key tool to visualize and assess the performance of our predictive models. Utilizing `numpy` for computation and `matplotlib.pyplot` for visualization, coupled with interactive features from `ipywidgets`, we provided users the ability to seamlessly navigate through different model outcomes. This interactive matrix highlighted the true and false positives and negatives for models like Logistic Regression, directly informing us of each model's strengths and weaknesses in detecting fraud. Such interactive analyses are not only engaging but can also deepen our understanding of each model's effectiveness in fraud detection, which is a crucial aspect in the ongoing enhancement of financial security measures.

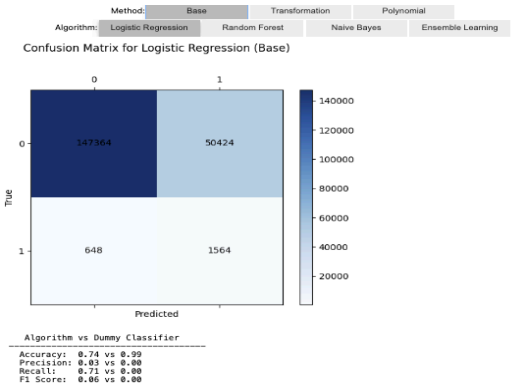

Figure 9: Interactive Confusion Matrix

4. Interactive 3D Clustering

In our exploration of cluster analysis, we harnessed the capabilities of Plotly to create interactive 3D plots, allowing for a dynamic visual exploration of the dataset. These plots, generated for both K-means and Gaussian Mixture Model (GMM) clustering methods, offer users the flexibility to manipulate the view by selecting either raw data or data transformed by PCA, the number of clusters, the color scale, and the marker size. The importance of these interactive plots extends beyond their visual appeal; they play a critical role in model validation and parameter tuning. By adjusting the number of clusters, users can immediately see the impact on the data segmentation, which is essential for determining the optimal clustering configuration. Moreover, the ability to switch between PCA-transformed and original data helps in assessing the dimensionality reduction's influence on the clustering outcome. With these tools, users can engage deeply with the modeling process, performing real-time analysis that was once limited to static representations, thereby enhancing the overall comprehension and effectiveness of the clustering methods in identifying patterns that could signify fraudulent activities.

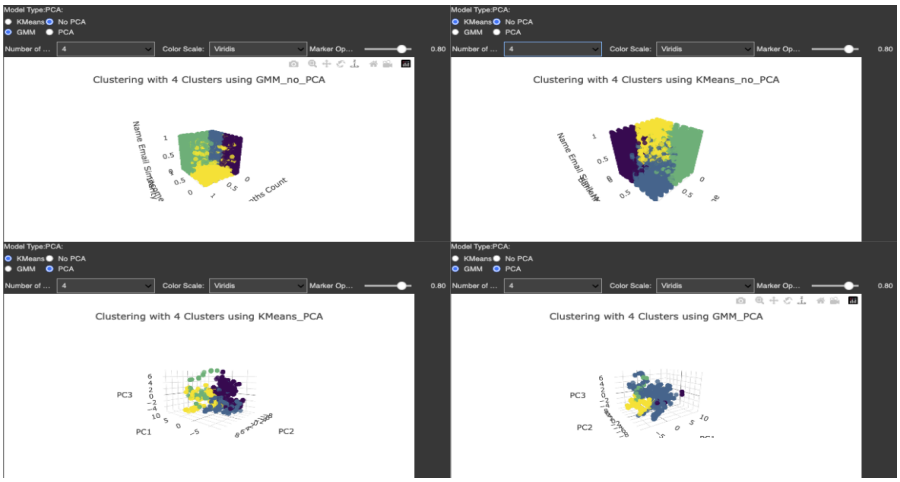

Figure 10: Interactive 3D Clustering Visualizations

VII. CONCLUSIONS AND DISCUSSION

In conclusion, our research has adeptly charted a course through the complexities of synthetic banking data to forge an advanced machine learning framework for detecting fraud. By integrating rigorous data preprocessing with sophisticated techniques such as PCA and ensemble learning, we've significantly improved the precision of our predictive models. Moreover, the deployment of interactive visual tools, like ROC, AUC plots and 3D clustering plots, has made our analytical findings more digestible and practical for real-world application. The challenges of synthetic data, which included the absence of true baselines and the replication of genuine transaction behaviors, have paradoxically honed our skills in pattern recognition and insight extraction. This experience has underscored the need for innovative approaches in analytical methodologies, even in less-than-ideal data conditions.

Moving forward, we would suggest future studies to incorporate live transactional data to further validate and possibly enhance the robustness of our established models. The potential of neural networks and unsupervised learning algorithms awaits exploration, promising to reveal the subtleties of complex, non-linear data relationships. Given the dynamic nature of fraud, adaptive learning systems that evolve with emerging fraud patterns could be a substantial asset.

VIII. CONTRIBUTIONS

C.X. and J.X. trained and tested the models. L.Y., T.C., F.C., reviewed the code dataset for improved processing efficiency and interpreted the datasets for strategic insights. L.M., created the report format. We all worked together on the project final report.

REFERENCES

[1] Vanini, P., Rossi, S., Zvizdic, E., & Domenig, T. (2023). Online payment fraud: from anomaly detection to risk management. Financial Innovation, 9, 66. Retrieved from https://jfin-swufe.springeropen.com/articles/10.1186/s40854-023-00470-w.

[2] Ali, A., Razak, S. A., Othman, S. H., Elfadil Eisa, T. A., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review. Applied Sciences, 12(19), 9637 Retrieved from https://www.mdpi.com/2076-3417/12/19/9637

[3]Asmar, M., & Tuqan, A. (2024). Integrating Machine Learning for Sustaining Cybersecurity in Digital Banks. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4686248. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4686248.

[4] Efijemue, O., Obunadike, C., Taiwo, E., Kizor, S., Olisah, S., Odooh, C., & Ejimofor, I. Cybersecurity Strategies for Safeguarding Customers Data and Preventing Financial Fraud in the United States Financial Sectors. International Journal of Soft Computing, 14(3), 10-5121.

[5] Domashova, J., & Kripak, E. (2021). Identification of non-typical international transactions on bank cards of individuals using machine learning methods. Procedia Computer Science, 190, 178–183. Retrieved from: https://www.sciencedirect.com/science/article/pii/S1877050921012655?via%3Dihub

[6] Figueroa, G., Chen, Y. S., Avila, N., & Chu, C. C. (2017, July). Improved practices in machine learning algorithms for NTL detection with imbalanced data. In *2017 IEEE Power & Energy Society General Meeting* (pp. 1-5). IEEE.

[7] Verma, N. (2023, December 12). Optimizing K-Means Clustering: A Guide to Using the Elbow Method for Determining the Number of Clusters. GoPenAI. Retrieved from https://blog.gopenai.com/optimizing-k-means-clustering-a-guide-to-using-the-elbow-method-for-determining-the-number-of-877c09b2c174

[8] Patel, E., & Kushwaha, D. S. (2020). Clustering cloud workloads: K-means vs gaussian mixture model. Procedia computer science, 171, 158-167.

[9] Kaggle. (2022). Bank Account Fraud Dataset Suite (NeurIPS 2022). Retrieved from https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022

[10] Zhu, T., Lin, Y., & Liu, Y. (2017). Synthetic minority oversampling technique for multiclass imbalance problems. *Pattern Recognition*, *72*, 327-340.

[11] Manlangit, S., Azam, S., Shanmugam, B., Kannoorpatti, K., Jonkman, M., & Balasubramaniam, A. (2018). An efficient method for detecting fraudulent transactions using classification algorithms on an anonymized credit card data set. In Intelligent Systems Design and Applications: 17th International Conference on Intelligent Systems Design and Applications (ISDA 2017) held in Delhi, India, December 14-16, 2017 (pp. 418-429). Springer International Publishing. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-76348-4_41

[12] Psychoula, I., Gutmann, A., Mainali, P., Lee, S. H., Dunphy, P., & Petitcolas, F. (2021). Explainable machine learning for fraud detection. Computer, 54(10), 49-59.

[13] Mishra, K. N., & Pandey, S. C. (2021). Fraud prediction in smart societies using logistic regression and k-fold machine learning techniques. Wireless Personal Communications, 119, 1341-1367. Retrieved from https://link.springer.com/article/10.1007/s11277-021-08283-9

[14] Perttilä, E. (2024). Utilizing Text Mining in Financial Fraud Detection. Retrieved from https://aaltodoc.aalto.fi/items/2ffe7da3-7a80-41fd-b88a-743d19d60a36.

[15] Rao, S. X., Lanfranchi, C., Zhang, S., Han, Z., Zhang, Z., Min, W., ... & Zhang, C. (2022). Modelling graph dynamics in fraud detection with" Attention". arXiv preprint arXiv:2204.10614.

[16] Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 international conference on computing networking and informatics (ICCNI) (pp. 1-9). IEEE.

[17] Ruchay, A., Feldman, E., Cherbadzhi, D., & Sokolov, A. (2023). The Imbalanced Classification of Fraudulent Bank Transactions Using Machine Learning. Mathematics (Basel), 11(13), 2862. Retrieved

from:
https://mdpi-res.com/mathematics/mathematics-11-02862/article_deploy/mathematics-11-02862.pdf?version=1687779630
[18] Talukder, M. A., Hossen, R., Uddin, M. A., et al. (2024). Securing Transactions: A Hybrid Dependable Ensemble Machine Learning Model using IHT-LR and Grid Search. Retrieved from https://ui.adsabs.harvard.edu/abs/2024arXiv240214389A/abstract.
[19] Edwin Raj, S. B., & Portia, A. A. (2011). Analysis on credit card fraud detection methods. In 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET) (pp. 152-156). IEEE. https://doi.org/10.1109/ICCCET.2011.5762457
[20] Lawton, G. (January 2022). Logistic Regression. Retrieved from https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression
[21] Title: Machine Learning: A brief introduction to Random Forest
Website:https://www.einsteinmed.edu/uploadedfiles/centers/ictr/new/intro-to-random-forest.pdf
[22] Chan, C. (n.d.). What is a ROC Curve and How to Interpret It. Retrieved from https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/
[23] IHEC C.,Túnez. Using a naive Bayesian classifier methodology for loan risk assessment. Evidence from a Tunisian commercial bank. Journal of Economics, Finance and Administrative Science, vol. 22, no. 42, pp. 3-24, 2017. Retrieved from: https://www.redalyc.org/journal/3607/360752107002/html/
[24] Guo, W., Yao, Y., Liu, L. et al. A novel ensemble approach for estimating the competency of bank telemarketing. Sci Rep 13, 20819 (2023). Retrieved from https://doi.org/10.1038/s41598-023-47177-7