An Analytic Approach of Bank Account Fraudulent Prediction

CSE6242 Data and Visual Analytics

Progress Report

Team 22

Team Members:

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

March 29th, 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

Inspired by previous studies [5][11][12][15], our project adopts machine learning but pivots to a refined analytic approach, enhancing predictive accuracy and adaptability in bank fraud detection. Our focus is on developing precise models and linking theory to practice by applying model results to real-world banking challenges. We have developed 4 methods which include the Base method, where we just used the scaled data to fit into machine learning models; the Transformation method, where we perform feature Box-Cox and logarithmic transformations; the Polynomial method, where we applied feature engineering; and the Clustering method, where we applied clustering and we compared the results between applied and non-applied PCA columns.

Our first approach involves comprehensive data preprocessing, including scaling and transformation, to improve normalization and linear relationships, as illustrated in Figures 1 and 2. This step prepares for deploying advanced machine learning algorithms, establishing a performance baseline. Additionally, we proactively tackle financial fraud's evolving complexity by exploring data transformation techniques like Box-Cox and logarithmic transformations to boost model accuracy.

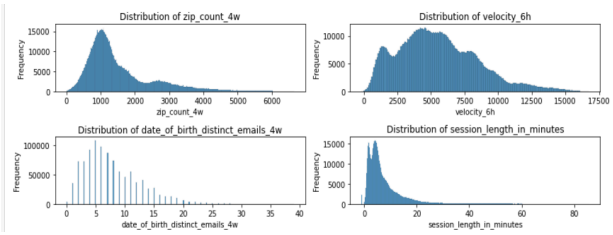

Figure 1: features before transformation

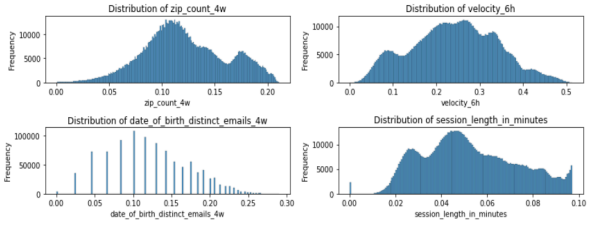

Figure 2: features after transformation

Our second innovation in detecting bank account fraud involves integrating feature engineering, like polynomial terms and squared variables, into our predictive models. This strategy uncovers complex patterns within financial data, blending traditional statistical methods with modern machine learning for a deeper understanding of fraudulent transactions.

The methodology begins with a comprehensive data preprocessing phase, where polynomial transformations are applied alongside scaling and normalization procedures. These initial steps are crucial for promoting data normalization, ensuring the establishment of linear relationships, and maintaining homoscedasticity across the dataset, thus laying a solid foundation for the advanced modeling phase. We then proceed to apply this enriched dataset to train a variety of machine learning algorithms, including Logistic Regression, Random Forest, and Naive Bayes. By exploiting the nuanced expressiveness afforded by polynomial features, our models are adept at discerning the subtle, complex interactions indicative of fraudulent activity, significantly elevating both the precision and adaptability of our fraud detection capabilities. This approach aligns with the necessity for models to evolve in tandem with the sophisticated tactics employed by fraudsters, presenting a forward-thinking solution to the intricacies of bank account fraud detection.

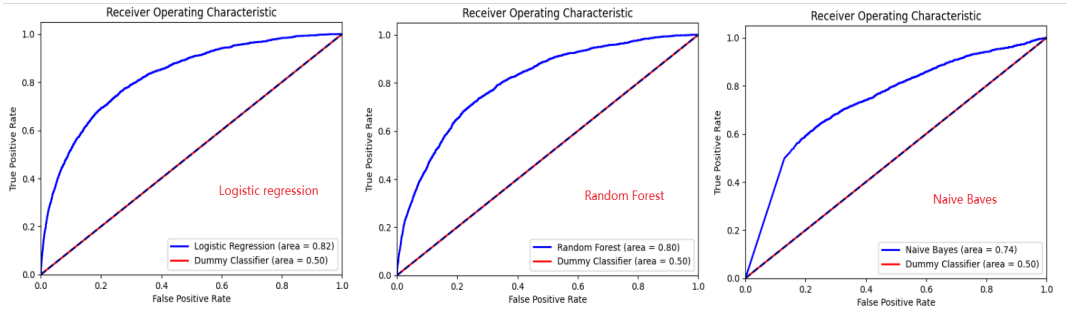

Figure 3: ROC curves

To address overfitting and refine variable selection, we first employ Principal Component Analysis (PCA). According to an article written by Whitfield [6], PCA is a widely-used technique for reducing the dimensionality of large datasets. It transforms numerous variables into a smaller set while retaining most of the original information. Subsequently, employed the Elbow method to determine the optimal cluster count, followed by the K-means algorithm. Its aim is to identify the inflection point on the plot of within-cluster sum of squares (WCSS) against the number of clusters, indicating a significant change in the rate of decrease, often referred to as the "elbow" [7]. We complement this with the Gaussian Mixture model, enabling a comparative analysis of clustering outcomes. This strategic use of clustering enhances the model's resilience and accuracy, underscoring our commitment to innovative fraud detection solutions.

To provide deeper insights into our clustering methods, we develop interactive plots for both the K-means and Gaussian Mixture models. These visualizations allow dynamic exploration, empowering users to zoom in on clusters, hover over data points for details, and compare cluster distributions. By offering an intuitive understanding of cluster dynamics, these interactive plots not only aid in identifying outliers and assessing cluster cohesion but also serve as powerful communication tools, ensuring transparency and facilitating stakeholder engagement [8]. Overall, the integration of interactive

visualizations enriches our methodology, enabling a comprehensive and transparent approach to financial fraud detection.

To enhance understanding, we develop interactive plots for both K-means and GMM models. These visualizations enable users to dynamically explore clusters, access details about data points, and compare cluster distributions. This not only aids in outlier identification and cluster cohesion assessment but also serves as an effective communication tool, enhancing transparency and stakeholder involvement. The integration of these interactive visualizations in our methodology enriches our approach to financial fraud detection, offering a thorough, clear perspective.

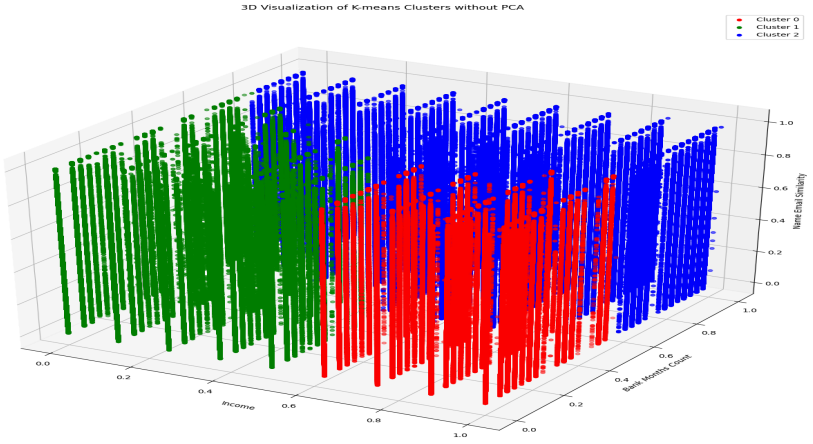

Figure 4: Visualization of K-Means Models

V. EVALUATION

In evaluating our models against the imbalanced BAF dataset, we utilized SMOTE during training to ensure our models' predictive reliability for the less frequent fraudulent cases. In our base model evaluation, Logistic Regression achieved 74.46% accuracy, Random Forest 79.24%, and Naive Bayes had the highest recall at 71.25%. All models showed potential but also areas for improvement, especially in balancing precision and recall.

With the Transformation method, Logistic Regression's recall increased to 70.48%, and accuracy to 75.28%. Random Forest's recall rose slightly, with accuracy reaching 79.86%. Naive Bayes saw significant recall improvement, though with decreased accuracy. This method enhanced true positive detection across models.

Using Feature Engineering, Logistic Regression balanced sensitivity and specificity well, with an accuracy of 76.83%. Random Forest's accuracy was slightly higher at 77.39%, but with lower precision. Naive Bayes achieved high recall but at the cost of precision. Overall, models outperformed the baseline, with further refinements needed for optimal fraud detection.

Method \ Model	Logistic Regression				Random Forest				Naive Bayes			
	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
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
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

Figure 5: Summarize Table

In our feature importance analysis, we identified key predictors that significantly influence fraud detection across our methods. For the Random Forest model in our base method, 'income', 'current_address_months_count', and 'credit_risk_score' were the top three features, with 'credit_risk_score' having the highest importance score. These features are crucial in predicting fraudulent activity due to their strong influence on the model's decision-making process.

Conversely, for the Logistic Regression model in the base model, 'credit_risk_score', 'device_distinct_emails_8w', and 'zip_count_4w' emerged as the top influencing features. Notably, 'credit_risk_score' again shows significant impact, but in this model, some features like

'device_distinct_emails_8w' and 'zip_count_4w' also play a major role, highlighting the model's sensitivity to different types of behavioral data. These insights are essential for fine-tuning our models and improving their predictive capabilities.

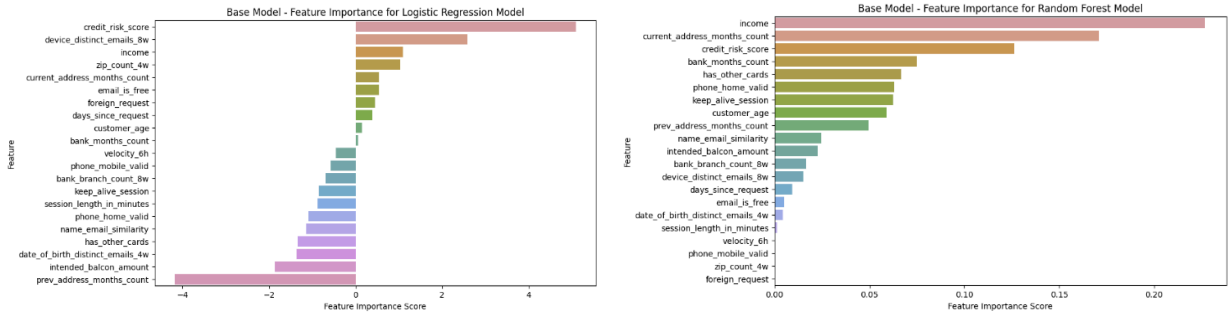

Figure 6: Comparison of Base Model Important Features

VI. CONCLUSIONS AND DISCUSSION

Our analysis confirms that using a multifaceted approach, combining advanced data preprocessing, machine learning, and clustering, enhances bank account fraud detection. Data transformation techniques like Box-Cox and logarithmic transformations, along with polynomial feature integration, have effectively improved model sensitivity to subtle fraud indicators. Furthermore, clustering algorithms have been valuable in reducing overfitting and in capturing complex fraud behaviors.

A key conclusion is the marked improvement in fraud detection from integrating advanced analytics into our models, offering financial institutions a scalable, robust fraud detection system. However, a critical discussion point is the trade-off between model complexity and interpretability. While intricate models can be more accurate, they risk becoming less transparent. Our project aims for transparency, ensuring that model insights are understandable and trustworthy for stakeholders. In summary, our research demonstrates that a thorough and analytically advanced approach to fraud detection not only meets accuracy needs but also adheres to the financial industry's operational and ethical standards.

Building on our current progress, we plan to further enhance our approach by incorporating ensemble learning techniques into our existing methods. This will allow us to compare different models and identify the most effective strategies for fraud detection. Additionally, we aim to complete our work on the Gaussian Mixture Model (GMM), which will complement our current clustering methods. To make our results more accessible and interpretable, we are also focusing on developing interactive plots for the clusterings. These plots will enable a more intuitive exploration and understanding of the data. Moreover, we plan to create interactive visualizations for comparing the results of different methods. This step will not only aid in evaluating the performance of each technique but also provide a comprehensive view of how they collectively contribute to detecting bank account fraud.

VII. CONTRIBUTIONS

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

VIII. PLAN OF ACTIVITIES

L.Y., T.C., and F.C. will examine the code for the data visualization component. C.X. and J.X. will finish interpreting data and explore real-world applications. L.M. will review the code for bugs and performance. The whole team will draft the final report, combining all findings and insights.

APPENDIX

Table 1: Contributions

Work item	Who	Planned Start	Planned Duration	Revised Start	Revised Duration
Data Cleaning	Yanhui, Chinhsien, Cheng	3/1/24	7	3/8/24	3
Explanatory Data	Yanhui, Chinhsien, Cheng	3/1/24	7	3/8/24	3
Train (Model Selection/Fit)	Xin, Xiaofan	3/8/24	7	3/15/24	3
Test (Model Evaluation)	Xin, Xiaofan	3/8/24	7	3/15/24	3
Code Review Dataset	Cheng	3/15/24	6	3/21/24	2
Interpretation of datasets	Mina	3/15/24	6	3/21/24	2
Report Format	Xiaofan, Chinhsien	3/21/24	1	3/21/24	1
Project Progress Report	Team	3/21/24	8	3/25/24	4

Table 2: Plan of Activities in High Level

Work item	Who	Planned Start	Planned Duration	Revised Start	Revised Duration
Model Selection, Evaluation	Xin, Xiaofan	3/29/24	7	4/5/24	3
Code Review Dataset	Yanhui, Chinhsien, Cheng	3/29/24	7	4/5/24	3
Interpretation of datasets	Mina	4/5/24	6	4/11/24	2
Report Format	Team	4/11/24	5	4/16/24	3
Project Progress Report	Team	4/11/24	5	4/16/24	3

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] Whitfield, B. (2024, February 23). A Step-by-Step Explanation of Principal Component Analysis (PCA). BuiltIn. Retrieved from https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[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] Whiting, D. G., Hansen, J. V., McDonald, J. B., Albrecht, C., & Albrecht, W. S. (2012). Machine learning methods for detecting patterns of management fraud. Computational Intelligence, 28(4), 505-527.

[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