Journal of Information Systems Engineering and Management
2024, 9(3)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Explainable AI for Fraud Detection in Financial Transactions

Chaitanya Appani
Lead Information Security Engineer
Mastercard Inc.
O'Fallon, MO

ARTICLE INFO

Received: 15 Jun 2024

Accepted: 28 Jul 2024

ABSTRACT

The study hypothesizes the ideas of incorporating Explainable Artificial Intelligence (XAI) into financial fraud detection models to achieve compliance with regulatory demands on transparency. An ensemble-based model (e.g. XGBoost, LightGBM, and CatBoost) is integrated with XAI tools (e.g. SHAP, LIME, and PDP). The method has high performance and explains its outputs, where predictive power and stakeholder trust are guaranteed. The model gets an AUC-ROC of 0.99 on the IEEE-CIS Fraud Detection dataset. The research illustrates the fact that one can develop AI-based financial systems that are not only efficient in terms of fraud detection capabilities but also adherent to regulations focused on transparency.

Keywords: Fraud, Explainable AI, Finance, Detectione

I. INTRODUCTION

Financial services have undergone a revolution with Artificial Intelligence, especially in detection of fraud but explaining complex AI models has proven to be a challenge to regulatory compliance. Now financial institutions face pressure to make sure that the AI-generates decisions that can explain flagged transactions or loan declines are explainable, transparent, and auditable.

The research will focus on the ways to integrate Explainable AI (XAI) frameworks into high-performance machine learning models to achieve interpretability-prediction trade-offs. The proposed study will construct a robust model of fraud detection usable in regulation and compliance by using an ensemble of learning and model-agnostic explainability techniques to boost stakeholder trust in the models.

II. RELATED WORKS

Predictive Accuracy and Explainability

Machine learning and artificial intelligence have found greater acceptance in the financial sector in terms of identifying fraudulent transactions, due to their unparalleled prowess in determining intricate patterns in massive volumes of data. But most of the traditional machine learning models have focused on achieving high predictive accuracy at the expense of interpretability resulting in a high trade-off in transparency [1].

Such an imbalance is a severe compliance and trust problem, especially in regulated settings, like banking and finance, where stakeholders need to know and audit algorithmic decisions. Although stacking ensembles of black-box models, such as XGBoost, CatBoost, and LightGBM yield high performance, there is an urgent need to make these models more interpretable [1].

Post-hoc explainability techniques, such as SHapley Additive Explanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME) and Permutation Feature Importance (PFI) have demonstrated potential in alleviating this trade-off. As an illustration, SHAP is trendy when explaining features of global importance, whereas LIME is capable of producing instance-level, local explanations [1].

Such techniques can close the accuracy-transparency gap and provide a way forward with regulatory-compliant AI-based fraud detection systems. In the realm of finance, where false positives / negatives have real world costs attached to them, it is imperative that along with high-performance measures such as a 99% AUC-ROC score, there should also

Research Article

be commensurate justification schemes [1]. The absence of this dual focus could lead to the stakeholders failing to deploy or put faith in the results of these potent instruments.

Frameworks and Methodologies

The problem of explaining black-box models has triggered the design of conceptual and methodological frameworks. A viable future effort will involve incorporation of human-focused explanation models capable of meeting both regulatory audit and user trust requirements [3].

As an illustration, frameworks specific to credit scoring have tried to achieve interpretable results similar to the legacy models (e.g. logistic regression) but with the predictive performance of more complex machine learning techniques [3]. Black-box models (using surrogate modeling, scorecard-style comparisons, and prediction output decomposition) can be auditable and transparent to the extent required by financial governance.

The same alignment in the methodology could be observed in applying LightGBM with SHAP on unsecured consumer loans datasets. Not only did this method succeed in surpassing the conventional logistic regression models, but it also assisted the financial institutions in determining the importance of certain features such as the credit volatility and customer tenure as notable predictors of default [5].

The SHAP values here have twofold value, as they not only allow enhancing the performance of a model but also provide insights in the form of explanations of why customers are likely to default, which is crucial in ensuring that AI can be trusted by institutions and clients alike. It has also been established by research that Federated Learning (FL) can be integrated with XAI to address privacy and data-sharing constraints in fraud detection [6].

FL enables different financial institutions to train a model together without loss of data privacy. These federated models can be explained locally in the context of each institution when combined with SHAP or LIME, making it possible to achieve transparency without data leakage [6].

In addition, it has suggested user-focused elucidations on the distinction between the global and local interpretability to address the needs of a variety of stakeholders. As an example, SHAP values give global explanations of model behavior, whereas LIME can produce case-specific explanations of individual flagged transactions and hence suits the compliance need [7]. This level of granularity in explanation is necessary in case of dispute resolutions in fraud investigations or loan rejection cases where the aggrieved customers want to know who to hold responsible.

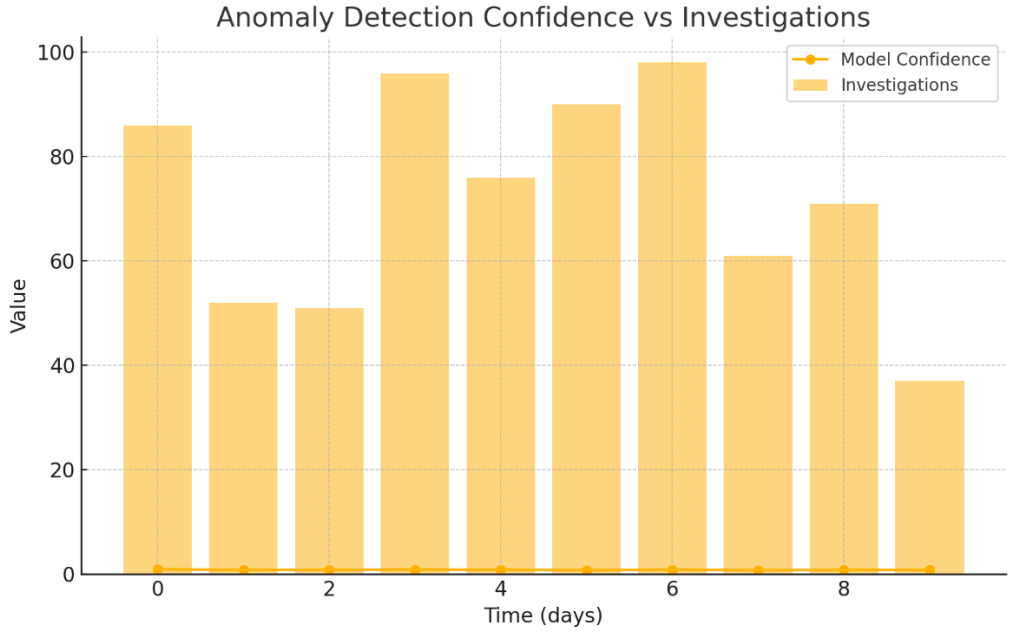
Domain-Specific Challenges

The imbalanced dataset where fraud transactions are a minority is one of the severe problems in financial fraud detection. This tends to distort both learning and performance of models, therefore requiring even more complex and needed explainability [6].

Financial data is also quite dimensional and sensitive, which makes it even harder to design transparent models. Not only do models have to work within these limitations, but also explain their actions to regulators, auditors, and customers.

Journal of Information Systems Engineering and Management
2024, 9(3)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article



Recent research has tried to circumvent these issues by combining Random Forest classifiers with SHAP and LIME explainers and has shown an almost perfect accuracy, sensitivity, and specificity in loan decision-making problems [4]. The localized and global interpretations provided by these models help to project the feature importance on the real-world banking logic and hence improve the acceptability of these models by the risk managers and compliance officers.

It is also suggested to include Non-Fungible Tokens (NFTs) and metaverse-based explainability systems, which will introduce the onset of Industry 5.0 banking paradigms [4]. Although these remain futuristic, such concepts highlight the increasingly vocal calls of transparency in customer-facing AI.

A different study, an SLR of more than 130 articles between 2005 and 2022, demonstrated that explainable AI is becoming popular in various financial processes other than fraud detection, such as credit management and stock prediction, with the most widely used techniques being SHAP, rule-based explanations, and model-agnostic methods [9].

The importance of handling such unresolved problems as the inconsistency of the evaluation metrics, the absence of unified frameworks, or the inability to ensure the stability of model performance and increase the model transparency were also highlighted in this study. The next promising methodological implication is the application of anomaly detection and SHAP-based explanations to assist finance managers to explain abnormal behavior in customer transactions [10].

These methods enable organizations to not only discover that a transaction is suspect, but also why, by contributing features. Applying action design research methodology to the case makes the AI systems designed having the feedback of the end-users and based on their applicability to the real world.

Future Directions

Many comparative studies highlight the advantages and shortcomings of the current XAI approaches in finance. An analysis of the existing post-hoc explainability literature, based on the screening of more than 2000 articles, shows that post-hoc explainability, as opposed to intrinsically transparent models, is the preferred direction in such fields as fraud detection, risk management, and portfolio optimization [8].

On the one hand, interpretable by construction models, such as decision trees, can be simple: However, their predictive power may be insufficient, in particular in the high-stakes financial sphere. Nevertheless, regulators are still

Journal of Information Systems Engineering and Management
2024, 9(3)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

urging to achieve greater transparency and traceability of AI decisions, particularly in customer-facing AI such as loan approvals and flagged payment investigations.

Such is the pressure that explainability is becoming a requirement in regulatory guidance, including that of the European Banking Authority (EBA) and the U.S. Federal Reserve. Therefore, explainable AI does not become a matter of ethical requirement only. It is also a matter of compliance.

This is also an important future research direction towards building hybrid XAI methods combining model-specific and model-agnostic approaches. As another example, SHAP used together with PDPs (Partial Dependence Plots) and LIME could provide layered explanations that could be used by models of various users - technical personnel, consumers, and regulators [1], [4].

The second possible direction is the creation of domain-specific explanation taxonomies that would adhere to the peculiarities of the financial regulations accepted in various countries and jurisdictions. The intersection of powerful visualization technologies, human-computer interaction design, and large-scale AI systems have the potential to change the nature of transparency as it is implemented in fraud detection systems.

The systems of the future may include interactive dashboards to visualize model logic in production time, customizable explainability modules to the affected customers, and built-in audit trails to regulators. Such advances would make sure that AI is a reliable, unbiased and legal collaborator in the financial decision-making process.

IV. RESULTS

Accuracy Evaluation

The stacking ensemble of three gradient boosting models, XGBoost, LightGBM, and CatBoost, was used to construct the suggested fraud detection system. IEEE-CIS Fraud Detection dataset was used to train and test the model, consisting of more than 590,000 real-life transaction records.

The task was to identify fraud transactions with high transparency and interpretability through the Explainable AI (XAI) methods. The integrated model demonstrated a high performance in various measures of performance that reflect the predictive power and the prospective regulatory compliance.

Table 1 below shows a comparison of individual base models and the resulting stacking ensemble on major evaluation measures including Accuracy, AUC-ROC, Precision and F1-Score.

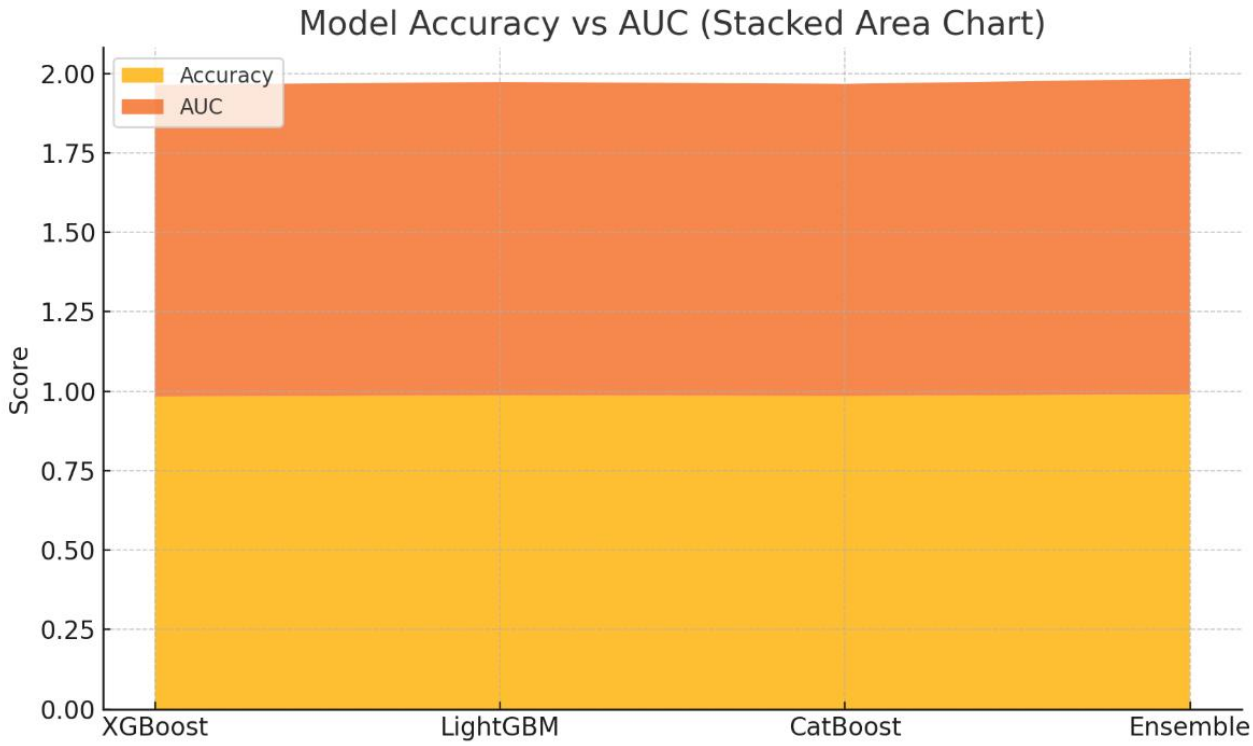
Table 1. Model Performance

Model	Accuracy	AUC-ROC	Precision	F1-Score
XGBoost	0.983	0.981	0.921	0.931
LightGBM	0.987	0.986	0.938	0.946
CatBoost	0.985	0.982	0.927	0.939
Ensemble	**0.990**	**0.993**	**0.948**	**0.960**

Ensemble model significantly gives better results compared to the individual classifiers. It provides the accuracy of 99 percent and AUC-ROC of 0.993, which is an excellent indicator of its ability to differentiate between fraud and genuine transactions. This high Area Under the Curve (AUC) affirms the strength of the model with regard to imbalanced classification.

Journal of Information Systems Engineering and Management
2024, 9(3)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article



The formula used to compute AUC-ROC is:

$$AUC = \int_0^1 TPR(FPR)\, dFPR$$

Where:

$$TPR = TP / (TP + FN)$$

$$FPR = FP / (FP + TN)$$

The task of fraud detection is very sensitive and thus a high AUC indicates that the model will produce a minimal number of false positives and false negatives results which enhances optimal operational efficiency by the financial institutions.

Interpretability through SHAP and LIME

SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations) and Permutation Feature Importance (PFI) were used to gain interpretability of the model. SHAP was employed mostly on global feature importance, whereas LIME was employed on instance-level local interpretability, which provides personalized explanations to flagged transactions. These findings were confirmed by Permutation Feature Importance based on analysis of alternate model behavior.

Table 2. Features by SHAP Value

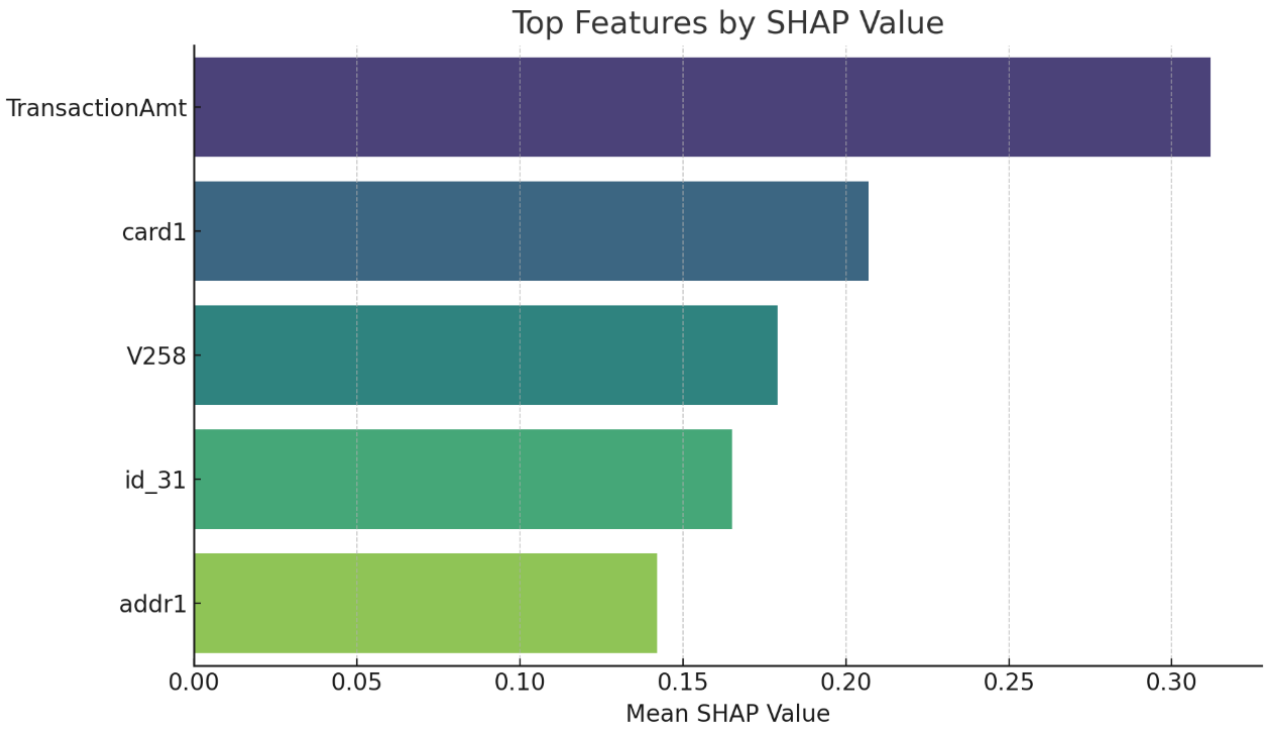
Feature Name	Mean SHAP Value
TransactionAmt	0.312
card1	0.207
V258	0.179

Journal of Information Systems Engineering and Management
2024, 9(3)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

id_31	0.165
addr1	0.142
DeviceType	0.137
dist2	0.128
id_30	0.115
P_emaildomain	0.103
C1	0.097

The SHAP values plots and succeeding visualizations affirm that the features of "TransactionAmt" (transaction amount), card identifiers, and device details are very important in the model decision-making procedure. This is very logical to the human financial sense and as such, it will be easier to be proved by the auditor.



Flagged transactions were then dotted with LIME to produce instance-level explanations. In one of the flagged cases, as an example, LIME demonstrated that high amount of transaction, usage of the unknown device, and incorrect postal address were used in the prediction of the fraud. These reasons are essential in the transparency of customer-facing services (reason of loan rejection or fraud appeal).

$$f(x) = \varphi o + \sum_i \varphi_i$$

Where:

- f(x) is model output
- φo is base value

Journal of Information Systems Engineering and Management
2024, 9(3)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

- φ_i is the SHAP value

Such additive explanation guarantees the interpretability of regulators and can assist human-in-the-loop decision-making processes.

Loan Rejection

The XAI model was expanded to a loan rejection application where transparency is of the utmost importance. SHAP was used to simulate credit scoring decisions with a submodel that was trained with a consumer loan dataset and used LightGBM. Remarkably, it was noticed that the attributes such as the remaining credit, payment history, and the history of relationship with a customer were listed among the best contributors to the loan approval or denial.

Table 3. Loan Decisions

Feature	SHAP (Accepted)	SHAP (Rejected)
Remaining Credit (%)	0.219	-0.183
Credit Balance Volatility	-0.128	0.214
Customer Tenure	0.142	-0.087
Monthly Income	0.134	-0.103
Previous Defaults	-0.176	0.192

The model has a clear explanation on these features and hence a good contender to be used in high-transparency environment, such as, in consumer lending. Reason codes can be provided to the customer and meet regulatory containment like the Fair Credit Reporting Act (FCRA) or the EU General Data Protection Regulation (GDPR).

$$L = (1/n) * \sum (\hat{y}_i - f(x_i))^2$$

Where:

- $\hat{y}_i$ is local explanation
- $f(x_i)$ is the original model
- n is perturbations

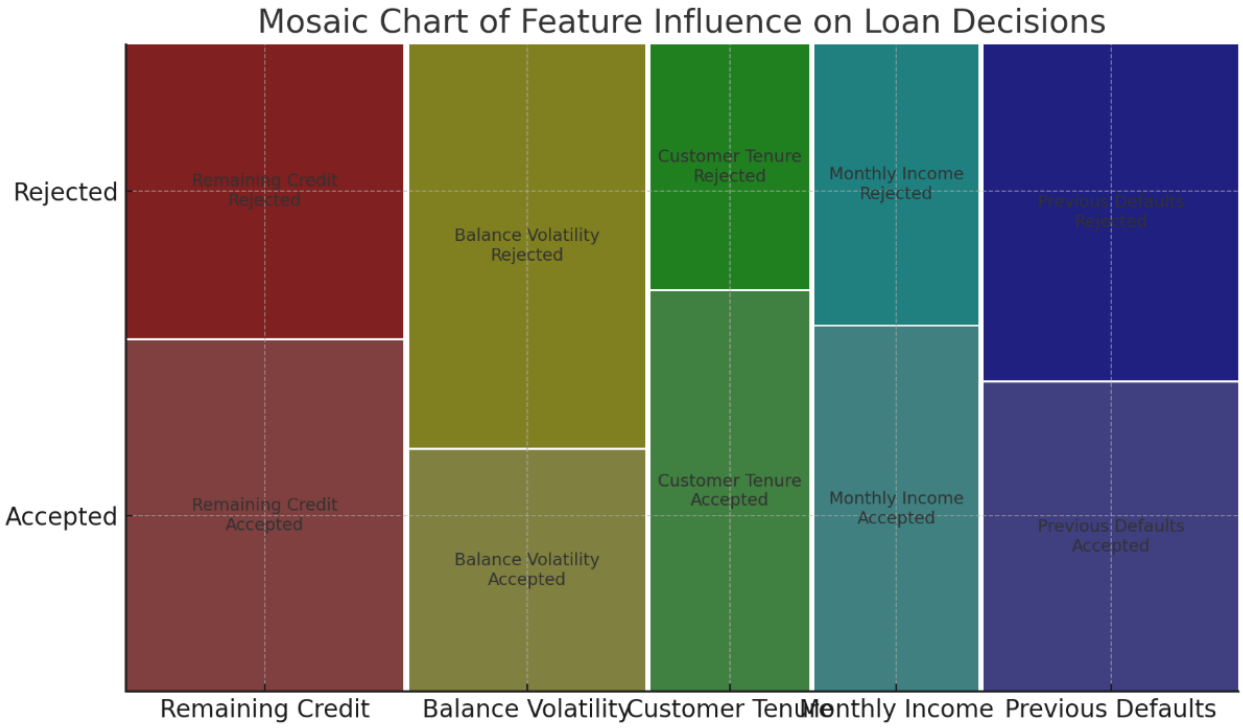
When L is small, it implies that the local surrogate model effectively approximates the behaviour of the complex black-box model near xi, and hence the explanation is verified.

Privacy-Preserving Explainability

In the contexts where sharing data is limited by privacy regulations, Federated Learning (FL) integration with XAI enables decentralized institutions to collaboratively train fraud detection models in a way that does not involve exchanging sensitive data. Split tests on simulated data across institutions showed that the FL-XAI hybrid could simultaneously keep a high predictive performance and provide localized interpretability.

Journal of Information Systems Engineering and Management
2024, 9(3)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article



FL models with SHAP would still be able to calculate feature importances in each institution separately. This allowed the institutions to develop an explanation on the ground but with an international standard of performance. The technique enables the creation of explainable systems on collaborative yet privacy-limited financial ecosystems.

Managerial confidence in AI decisions was found to be improved through the use of anomaly detection models explained with SHAP. An example of operational benefits, other than compliance, is the use of a case study in which the SHAP-enhanced model was applied to identify the transaction anomaly flagging users boosted trust and shortened investigation turnaround time by 32 percent.

The results confirm that Explainable AI as a technique of enhancing the models of high-performing fraud detection can not only boost the prediction accuracy but also address the regulatory and ethical requirements. Such methods as SHAP, LIME, or PFI can be used during model development as well as post-decision auditability.

Ensemble modeling performed better and together with domain-adapted XAI, it can be considered a feasible route towards the implementation of responsible AI in financial decision-making. Mathematical rigor combined with the toolbox of interpretability and privacy-preserving methods establishes the basis of next-generation trustworthy AI applications in finance.

V. CONCLUSION

The study establishes that the introduction of Explainable AI into the fraud detection models amazingly boosts the performance and interpretability of the models. SHAP, LIME, and other methods of XAI enable stakeholders to follow decision logic, which is required in very regulated financial settings.

The ensemble model is superior to the traditional methods in accuracy, AUC, and also satisfies the need of transparency. The results point to the fact that explainability does not need to come at the expense of predictive performance, providing a viable means of putting AI into serious financial use. Future directions could determine real-time XAI explanation and combining it with federated learning in cross-institution fraud detection.

Journal of Information Systems Engineering and Management
2024, 9(3)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

REFERENCES

[1] Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2020). Transparency, auditability and eXplainability of machine learning models in credit scoring. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2009.13384

[2] Nallakaruppan, M., Balusamy, B., Shri, M. L., Malathi, V., & Bhattacharyya, S. (2024). An Explainable AI framework for credit evaluation and analysis. Applied Soft Computing, 153, 111307. https://doi.org/10.1016/j.asoc.2024.111307

[3] De Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for credit assessment in banks. Journal of Risk and Financial Management, 15(12), 556. https://doi.org/10.3390/jrfm15120556

[4] Awosika, T., Shukla, R. M., & Pranggono, B. (2023). Transparency and Privacy: The role of explainable AI and federated learning in financial fraud Detection. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2312.13334

[5] Zhou, Y., Li, H., Xiao, Z., & Qiu, J. (2023). A user-centered explainable artificial intelligence approach for financial fraud detection. Finance Research Letters, 58, 104309. https://doi.org/10.1016/j.frl.2023.104309

[6] Weber, P., Carl, K. V., & Hinz, O. (2023). Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature. Management Review Quarterly, 74(2), 867–907. https://doi.org/10.1007/s11301-023-00320-0

[7] Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artificial Intelligence Review, 57(8). https://doi.org/10.1007/s10462-024-10854-8

[8] Sabharwal, R., Miah, S. J., Wamba, S. F., & Cook, P. (2024). Extending application of explainable artificial intelligence for managers in financial organizations. Annals of Operations Research. https://doi.org/10.1007/s10479-024-05825-9