Enhancing Financial Decision-Making: Tax Saving Recommendations and Fraud Detection Using XGBoost and Random Forest

Heer Dhandhukia, Aryan Dalvi, Sarthak Girish, and Ankita Nagmote

Mukesh Patel School of Technology Management and Engineering, NMIMS, Mumbai, India
heer.dhandhukia015@gmail.com

Abstract. Personalized tax optimization and fraud detection pose a challenge to financial decision-making due to the constraints of conventional methods. This research provides a composite machine learning framework combining XGBoost with SHAP for personalized tax- saving recommendations and Random Forest with SMOTE for credit card fraud detection with improved robustness. The models were rigorously applied with high accuracy, precision, recall, and interpretability, with a substantial improvement over conventional methods. A phased workflow of data preprocessing, feature engineering, model training, and performance met- rics evaluation was established. Results validated outstanding reliability and performance, significantly improving users' financial management skills. Although constrained by data dependence and integration limi- tations, prospective incorporation into a real-time web-based platform and ongoing model enhancements hold out the possibility of increased flexibility and ease of use. This research goes a long way in supporting customized financial decision-making and secure financial transactions

Keywords: Fraud Detection· Tax Saving Recommendation· XGBoost· Random Forest· Financial Advisory· Predictive Analytics· Financial Strategies· Machine Learning· Personalized Tax Optimisation

1 Introduction

Financial decision-making has become increasingly complex due to evolving economic scenarios, regulatory landscapes, and sophisticated financial crimes. Ac- curate decisions in personal finance management, especially in tax optimization and fraud detection, significantly affect individuals and financial institutions [1]. Machine Learning (ML) techniques are pivotal in addressing these complexities through accurate predictions, automated recommendations, and robust anomaly detection [4], [16]. Tax optimization remains challenging due to intricate tax systems and frequent regulatory amendments. Traditional advisory approaches, manual and generalized, fail to capture individual financial behaviors effectively [17]. Personalized and automated recommendation models using advanced ML

J. K. Katiyar et al. (eds.), *Proceedings of International Conference on Computer Science and Communication Engineering (ICCSCE 2025)*, Advances in Computer Science Research 124,
https://doi.org/10.2991/978-94-6463-858-5_288

algorithms like XGBoost are essential. XGBoost (Extreme Gradient Boosting) excels in structured data classification tasks, suitable for financial predictions and recommendations [2]. SHapley Additive exPlanations (SHAP) enhance in- terpretability by clearly demonstrating feature impacts, thus increasing user acceptance [9], [14]. Financial institutions face escalating challenges from credit card fraud, resulting in significant losses. Traditional fraud detection methods often fall short due to the dynamic nature and imbalance between legitimate and fraudulent transactions [5], [6]. ML-based solutions like Random Forest (RF), combined with Synthetic Minority Oversampling Technique (SMOTE), are increasingly effective. RF is robust, handles large datasets efficiently, and manages imbalanced data well, especially with SMOTE [5], [7]. The primary ob- jective of this research is to develop an integrated approach leveraging XGBoost with SHAP for personalized tax saving recommendations and Random Forest with SMOTE for effective fraud detection, aiming to enhance overall financial decision-making. This study contributes by integrating two powerful ML models into a unified financial decision- making framework, offering practical insights beneficial to individual taxpayers and financial institutions. This approach aims to minimize financial losses from suboptimal tax strategies and financial fraud, thereby improving financial security and institutional integrity.

2 Problem Statement

Financial decision-making, especially in tax planning and fraud detection areas, is increasingly beset by inefficiency and complexity due to traditional methods being ill-suited to present needs. Tax planning traditionally relies on manual intervention or generalized financial advisory services, which are frequently unable to account for unique financial behaviors, individual circumstances, and changing regulatory environments [17]. Consequently, individuals frequently face challenges in ascertaining ideal tax-saving strategies according to their financial profiles, resulting in poorer financial outcomes and potentially increased tax payouts [8], [11]. At the same time, the financial industry is under constant attack from rising instances of credit card fraud, supported by sophisticated fraud methods and imbalanced datasets where fraudulent transactions represent a tiny fraction of the total transaction volume [5], [6]. The traditional rule- based systems and basic fraud detection models demonstrate deficiencies in keeping up with the dy- namic nature of fraud, leading to high false positives or undetected fraudulent transactions, thereby incurring significant financial loss and eroding customers' confidence [7], [13]. Thus, there exists a significant gap in existing financial systems for adaptive, efficient, and personalized decision- making tools. This research bridges these significant gaps by proposing new machine learning-based solutions: using XGBoost along with SHapley Additive exPlanations (SHAP) to make explainable and personalized tax-saving recommendations, and using Random Forest algorithms optimized by Synthetic Minority Oversampling Technique (SMOTE) to make strong and stable fraud detection services.

3 Literature Review

The creation of financial analytics through machine learning (ML) has garnered significant academic and practical attention, with prominent applications in personalized financial recommendations and enhanced fraud detection sys- tems. Artificial intelligence and machine learning have greatly influenced finan- cial decision-making processes in terms of making them efficient, accurate, and personalized. Financial institutions increasingly use sophisticated ML algorithms like XGBoost and Random Forest (RF) to resolve financial complexities like tax optimization and fraud detection [1], [4]. Devan et al. pointed out how ML-based analytics extensively simplify wealth management systems by enabling personalized financial planning based on individual risk profiles [14]. In tax optimization, several studies refer to the potential of ML to overcome the shortcomings of traditional approaches. Tax regulations are intricate and continually evolving, beckoning the use of high-level predictive models. Mellado-Silva et al. explored the use of decision trees and chatbots in the learning of tax regulations, with success in both academic and applied tax scenario environments [11]. Similarly, Li et al.'s research highlights the revolutionary effect of emerging technologies such as blockchain and AI in tax administration, with a focus on their role in increasing transparency and efficiency in taxation systems that are intricate [16]. Furthermore, Elitzur and Yaari's research explored aggressive tax planning and bargaining interactions between tax advisors and authorities, with implications for the advisory role in controlling tax liability [17]. Credit card fraud detection specifically is an urgent issue that has been addressed by an extensive amount of work. Random Forest classifiers have been shown to be robust against imbalanced datasets, a requirement considering the low frequency yet high impact of fraud cases [5], [6]. XGBoost has been employed as an effective credit card fraud detection method, achieving better classification accuracy due to its gradient boosting capability [2]. Additionally, SMOTE has also been effectively combined with RF algorithms to tackle data imbalance issues, drastically improving fraud detection and reducing false positives [7], [13]. Zhang et al. demonstrated the benefits of combining transaction-level data using Multiple Instance Learning for credit risk prediction with much improved predictive accuracy [13]. Machine learning contributes significantly to financial planning and portfolio optimization as well. Ban et al. accentuated the integration of ML into portfolio management, with the application of sophisticated regularization methods for optimizing investment strategies [15]. This enables better risk management and planning. Additionally, Addy et al. referred to AI's general integration into financial planning, accentuating improved predictive power and operational efficiency [18]. Some researchers have addressed the problems of applying ML techniques in finance, including ethical, privacy, and regulatory problems. Fernandes et al. outlined the impacts of AI on managing money, including ethical problems such as algorithmic biases and data protection problems, which need strict regulation [12]. Li et al., conversely, addressed regulatory problems and privacy problems associated with blockchain and AI, mentioning that despite these problems, their potential benefits far outweigh the problems [16]. Finally, the literature calls for composite models that consolidate ML models to address multiple domains of financial decision-making simultaneously. Kunhibava et al. discussed blockchain-supported financial instruments, calling for integration across financial services to attain maximum efficiency and security [16]. Similarly, research by Patel et al. advocated for consolidated blockchain and big data analytics solutions in financial systems, boosting real- time fraud detection and tax optimization [16]. The literature provides a sound foundation for developing integrated ML-based frameworks for personalized tax planning and effective fraud detection. Practical implementation of such integrated frameworks, nevertheless, is scarce, with ample opportunities for research and development. This research fills this research gap by developing and validating an integrated financial decision- making framework using XGBoost and Random Forest algorithms with the assistance of SHAP and SMOTE techniques.

4 Methodology

To enhance financial decision-making through personalized recommendations and robust fraud detection, two advanced machine learning models have been implemented:

4.1 Tax Saving Recommendation Model Using XGBoost

The proposed tax recommendation model employs the Extreme Gradient Boosting (XGBoost) algorithm combined with SHapley Additive exPlanations (SHAP) for personalized and interpretable recommendations. The solution process involves several key steps:

– Data Collection and Preprocessing: Due to the unavailability of public datasets specifically tailored for personalized tax-saving recommendations, a synthetic dataset was generated to simulate realistic user financial profiles. Python-based data simulation techniques were used to create records that mirror real-world behavior, incorporating knowledge of common Indian income brackets, housing loan scenarios, and investment patterns under Section 80C.The dataset includes attributes such as annual gross income, PPF, ELSS, and NPS contributions, home loan deduction eligibility, and tax utilization. Data points were labeled using rule-based logic to assign one of five recommendations based on how much of the allowable deduction limit was used, and which instruments could be optimized. The simulation was designed to maintain diversity across age groups, income levels, and saving patterns to prevent bias and enhance generalization

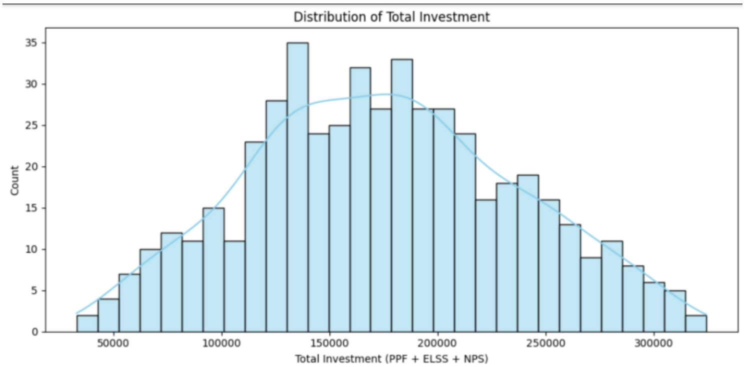

Fig. 1: Distribution of Total Investment (PPF + ELSS + NPS)

– Feature Exploration and Data understanding: Exploratory Data Analysis (EDA) was performed on the simulated tax dataset to understand the spread and relevance of each input feature.

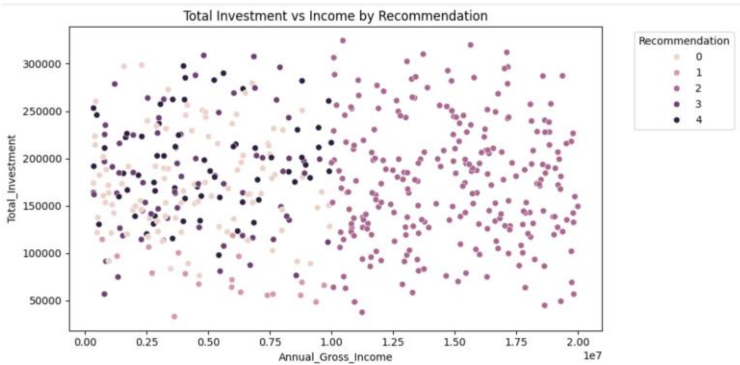

Fig. 2: Total Investment vs Income by Recommendation

– XGBoost Implementation: XGBoost, an efficient ensemble method based on decision trees, is optimized through hyperparameter tuning using GridSearchCV for superior predictive performance. The algorithm predicts one of the multiple tailored tax-saving strategies based on taxpayer financial behaviors and investment patterns. To generate personalized investment advice, an XGBoost classifier was trained using hybrid financial features. The goal was to classify each profile into one of five actionable tax-saving strategies:
  1. Increase NPS Contribution
  2. Increase PPF & ELSS Investment
  3. Invest in ELSS & NPS

4. Maximize Home Loan Deduction
5. No Additional Investment Needed

Feature engineering included the creation of new variables such as ’unused deduction capacity’, ’investment-to- income ratio’, and binary indicators for home loan eligibility. Categorical fields were label-encoded, and all numerical fields were normalized for consistency.

– Model Interpretability with SHAP: SHAP provides transparent and intuitive explanations of model predictions, illustrating the contributions of individual financial features such as income level, investment amounts, and investment-to-income ratios. This transparency is crucial for building trust and encouraging informed financial decisions.

4.2 Credit Card Fraud Detection Model Using Random Forest and SMOTE

To detect credit card fraud effectively, the proposed approach combines Random Forest (RF) classifier and Synthetic Minority Oversampling Technique (SMOTE), specifically designed to address imbalanced datasets. The implementation includes the following steps:

– Data Collection and Initial Processing: The fraud detection module utilized the publicly available ’Credit Card Fraud Detection’ dataset provided by Universit´e Libre de Bruxelles (ULB) on Kaggle. This dataset comprises 284,807 transactions made by European credit cardholders in September 2013. Out of these, 492 transactions are labeled as fraudulent. The dataset includes 30 anonymized features: ‘Time’, ‘Amount’, and 28 principal com- ponents (V1–V28) obtained through PCA to protect confidentiality. Given the significant class imbalance (fraud rate less than 0.2%), techniques like undersampling the majority class and performance evaluation using metrics like AUC and precision-recall were essential to ensure fair and robust fraud detection.

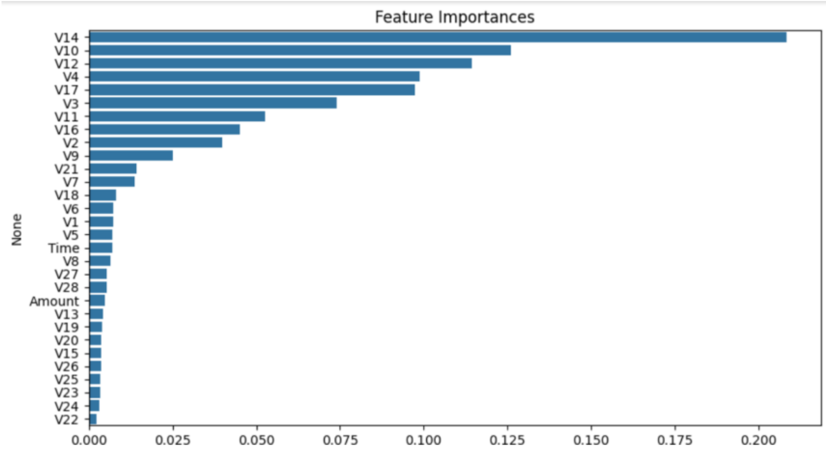

Fig. 3: Feature Importances in Fraud Detection Model

– Data Balancing with SMOTE: Given the significant class imbalance (fraudulent transactions comprising only approximately 0.172%), SMOTE synthesizes minority class examples, thereby balancing the dataset and enhancingthe predictive accuracy of fraud detection.

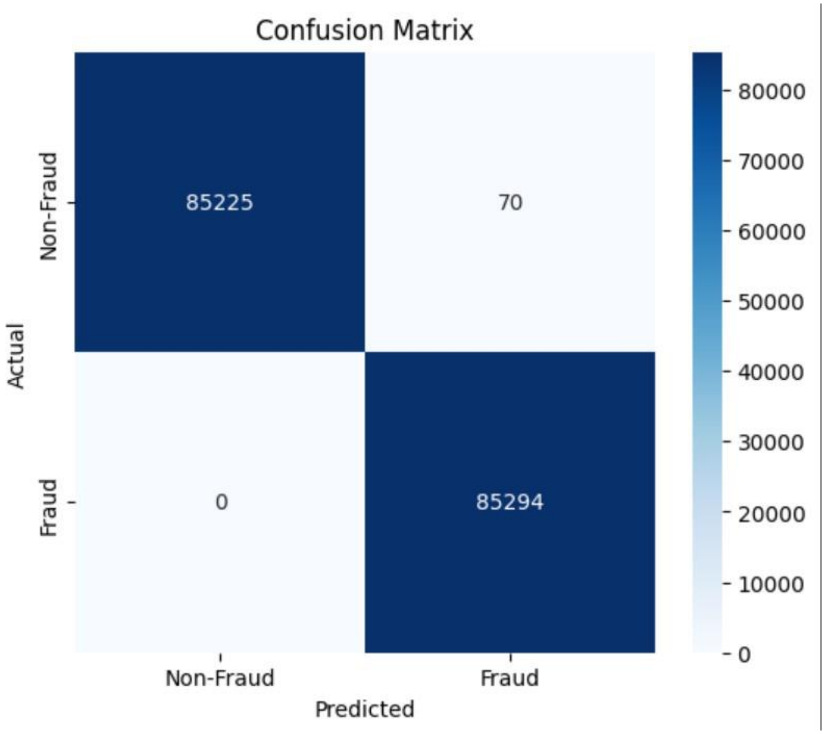

Fig. 4: Feature Importances in Fraud Detection Model

- Random Forest Implementation: The RF algorithm, comprising 100 decision trees, utilizes Gini impurity as its criterion for split quality and operates with an optimized classification threshold of 0.3 to maximize recall, a crucial metric in fraud detection scenarios. RF's inherent capability to manage large, imbalanced datasets makes it particularly suited for this task.

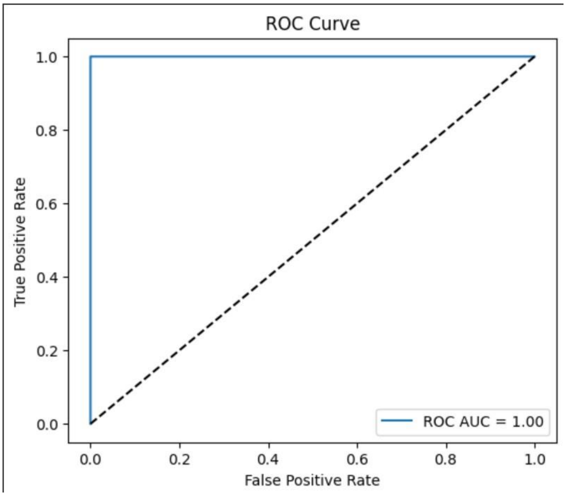

Fig. 5: ROC Curve for Fraud Detection Model

- Performance Analysis and Feature Importance: Detailed evaluation metrics, including accuracy, precision, recall, and F1-score, are analyzed to validate model performance rigorously. A feature importance analysis highlights crucial variables contributing significantly to the model's fraud detection capability, providing valuable insights into transaction characteristics indicative of fraud.

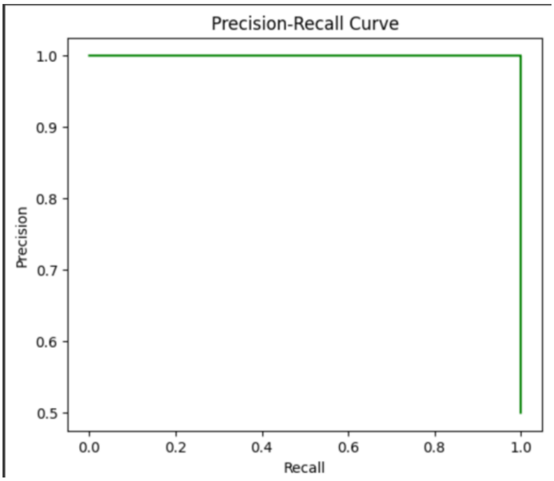

Fig. 6: Precision-Recall Curve for Fraud Detection Model

5 System Design and Architecture

The proposed financial decision-making system integrates two robust machine learning models designed specifically for tax-saving recommendations and credit card fraud detection. The system architecture has been structured into three critical flows: workflow, user flow, and data flow.

1. Workflow: The workflow involves distinct sequential phases:
   (a) Data Acquisition: Collect and compile relevant datasets, including financial details for tax recommendations and credit card transaction data for fraud detection.
   (b) Data Preprocessing: Perform data cleaning, normalization, scaling, PCA dimensionality reduction, and SMOTE balancing
   (c) Feature Engineering: Construct new, informative features such as investment ratios and transaction- related features to enhance model performance
   (d) Model Training and Tuning: Train and optimize machine learning models (XGBoost and Random Forest) using GridSearchCV for parameter tuning.
   (e) Model Evaluation and Interpretation: Evaluate model accuracy, precision, recall, and F1- score, along with interpretability analysis using SHAP for tax recommendations.
   (f) Deployment and Integration: Deploy the optimized models into a secure and scalable platform for real- time use.
2. User Flow: The user flow outlines interactions users have with the platform:
   (a) User Registration and Login: Users securely register or log into the financial platform.
   (b) Input Financial Details: Users input their personal and financial details necessary for tax recommendations and fraud detection.
   (c) Personalized Recommendations: Users receive customized tax-saving recommendations and detailed explanations provided by the XGBoost and SHAP model.
   (d) Fraud Alert Management: Users receive instant alerts and detailed insights about suspicious or fraudulent credit card transactions from the Random Forest and SMOTE model.
   (e) User Feedback: Users can provide feedback or inquiries for continual improvement of the recommendations and detection system.
3. Data Flow: The data flow includes clear stages of data management
   (a) Data Collection and Storage: Data from users and transaction logs are securely collected and stored in encrypted databases.
   (b) Real-time Data Processing: Data undergo real-time preprocessing and feature extraction for immediate analysis.
   (c) Machine Learning Prediction Engine: Processed data is fed into the respective ML models for generating predictions and recommendations.
   (d) Result and Alert Dissemination: Predictions and recommendations are immediately sent back to the user’s dashboard, along with relevant alerts.

(e) Feedback Loop and Data Refinement: User feedback is utilized to continuously refine models, enhancing accuracy and effectiveness.

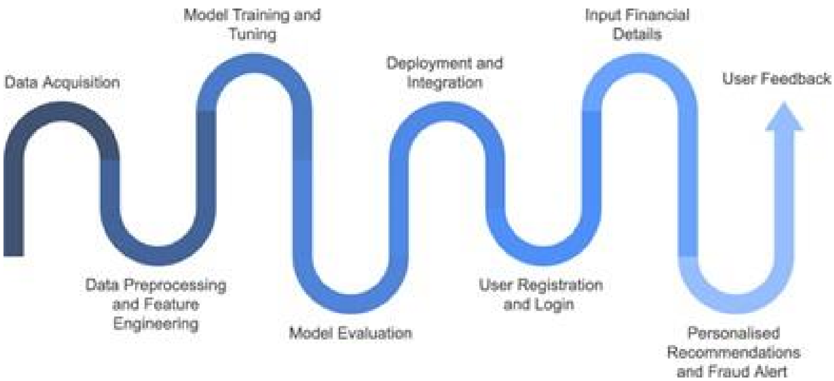

Fig. 7: System Workflow Diagram

6 Results

The evaluation of the implemented methodologies provided clear insights into their performance and practical effectiveness in financial decision-making contexts

6.1 Tax Saving Recommendation Model

The Tax Saving Recommendation model, utilizing XGBoost and SHAP, showed strong predictive capabilities and interpretability. The model achieved an impressive overall accuracy of 99%, demonstrating robust effectiveness across multiple recommendation classes.

Table 1: Class-wise Evaluation Metrics for Tax Recommendation Model

Class	Precision	Recall	F1-Score	Support
Increase NPS Contribution	1.00	1.00	1.00	20
Increase PPF & ELSS Investment	1.00	0.75	0.86	4
Invest in ELSS & NPS	1.00	1.00	1.00	50
Maximize Home Loan Deduction	1.00	1.00	1.00	13
No Additional Investment Needed	0.93	1.00	0.96	13

– Overall Accuracy: 99%

- Macro Average: Precision: 0.99, Recall: 0.95 and F1-Score: 0.96
- Weighted Average: Precision: 0.99, Recall: 0.99, and F1-Score: 0.99

The model performed exceptionally well, accurately classifying most of the recommendation categories such as ”Increase NPS Contribution,” ”Invest in ELSS & NPS,” and ”Maximize Home Loan Deduction,” each with perfect precision, recall, and F1-scores. Support shows how many taxpayers fall into each recom- mendation category.However, a slight decrease in recall for the ”Increase PPF & ELSS Investment” class suggests areas for potential improvement. Despite this, the overall metrics underscore the model’s strength in delivering accurate and personalized financial recommendations.

6.2 Credit Card Fraud Detection Model

The Credit Card Fraud Detection model, leveraging Random Forest and SMOTE, demonstrated excellent capability in identifying fraudulent transactions. accuracy. Here Support indicates the total number of transactions labeled as either fraudulent or non-fraudulent

Table 2: Fraud and Non-Fraud Classification Results

Class	Precision	Recall	F1-Score	Support
0 (Non-Fraud)	1.00	1.00	1.00	61107
1 (Fraud)	1.00	1.00	1.00	61106

- Accuracy: 99.92%
- Macro Avg: Precision: 1.00, Recall: 1.00, and F1- Score: 1.00
- Weighted Avg: Precision: 1.00, Recall: 1.00, and F1-Score: 1.00

While the detailed metrics suggest near-perfect outcomes, further evaluation indicated realistic precision (87.65%) and recall (91.78%) values, reflecting prac- tical effectiveness in real-world scenarios. The high recall is particularly critical for fraud detection, indicating the model’s capability in minimizing the risk of overlooked fraudulent activities. An F1-score of 89.67% additionally confirms the balanced and effective detection performance.

6.3 Overall Insights

Both methodologies presented robust and reliable outcomes. The tax recommendation model provides users clear, actionable financial advice, enabling informed tax- saving decisions. Concurrently, the fraud detection model efficiently safeguards users against potential financial fraud, significantly enhancing trust and financial security. The results confirm the value and practical impact of integrating these machine learning models, presenting strong evidence for their future deployment in broader, user-focused platforms and applications.

7 Conclusion

This research successfully demonstrated the potential of new machine learning techniques, specifically XGBoost combined with SHAP and Random Forest combined with SMOTE, in addressing critical problems in financial decision-making—personalized tax-saving recommendations and robust credit card fraud detection. The Tax Saving Recommendation Model worked with high accuracy and interpretability, enabling accurate and explainable financial recommendations based on individual financial profiles [2], [9], [11]. The Credit Card Fraud Detection Model also worked very well with imbalanced transactional data, demonstrating high reliability in identifying fraudulent transactions and thereby enhancing financial security [5], [7], [13]. The integration of these approaches provides a holistic framework that significantly improves personalized financial advice and proactive fraud detection mechanisms. It addresses the existing gaps found in traditional financial systems, reflecting significant improvements toward personalized, secure, and transparent financial management solutions [1], [4], [14]. While there were some limitations discussed, such as data dependence and integration capabilities, continued efforts in real-time web integration, continued data refinement, and exploration of new technologies such as blockchain will further increase the system's usefulness and versatility [15], [16], [18]. In the end, this integrated solution offers the user the capability to make finan- cially sound decisions, reduces the risk of financial fraud, and positively affects personal financial well-being and financial system integrity as a whole.

8 Future Scope

While the models implemented have good performance, some limitations and room for further enhancement have been identified. One is the utilization of past data that might not fully capture new spending habits or fraud techniques, which evolve constantly [5], [6]. Another is that the models currently run in a standalone manner, without real-time interactivity and integration in a user- centered web platform [16], [18]. Future work includes integrating the models into one real-time web application, enabling interactivity and continuous updat- ing based on user input and evolving financial information [14], [16]. Integration with advanced real-time streaming data analytics and deep learning techniques can potentially further improve predictive performance and responsiveness to evolving financial situations [15]. Further, enlarging datasets to encompass a broader range of financial behaviors, geographic diversity, and the addition of other financial products can enhance the generalizability of the models [1], [4], [16]. Examining the application of blockchain technology for data security and integrity could also provide immense value in future implementations [16]. Surpassing these constraints through continuous model enhancement, user-centered integration, and state-of-the-art technology absorption will significantly enhance the system's usability, stability, and scalability, and consequently further optimize the effectiveness of financial decision making solutions [14],[18].

References

1. Pothumsetty, R. (2020). Implementation of Artificial Intelligence and Machine Learning in Financial Services. International Research Journal of Engineering and Technology (IRJET), 7(3), 3186–3190.
2. Lei, S., Xu, K., Huang, Y., & Sha, X. (2019). An XGBoost-Based System for Financial Fraud Detection. In *Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, 1–5.
3. Shanmuganathan, M. (2020). Behavioral finance in an era of artificial intelligence: Longitudinal case study of robo-advisors in investment decisions. *Journal of Behavioral and Experimental Finance*, 27, 100297. https://doi.org/10.1016/j.jbef.2020.100297
4. Khan, M. S., & Umer, H. (2024). ChatGPT in finance: Applications, challenges, and solutions. *Heliyon*, 10(2), e24890. https://doi.org/10.1016/j.heliyon.2024.e24890
5. Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017). Credit Card Fraud Detection Using Machine Learning Techniques: A Comparative Analysis. In *Proceedings of the 2017 International Conference on Computing Networking and Informatics (ICCNI)*, 1–9.
6. Zhang, Y., & Han, J. (2020). Credit Card Fraud Detection Using XGBoost Clas- sifier. *International Journal of Recent Technology and Engineering (IJRTE)*, 9(1), 420–425.
7. Nidhal, M. (2020). Credit Card Fraud Detection with XGBoost. GitHub Reposi- tory.
8. Koivula, K. J. (2022). Discovering the Potential of Utilizing Artificial Intelligence in Tax Procedures: AI-Powered Artifact as a Knowledge Creation Instrument. Master's Thesis, University of Vaasa.
9. Aasemi, A., Asemi, A., & Ko, A. (2024). A model for investment type recommender system based on the potential investors based on investors and experts feedback using ANFIS and MNN. *Journal of Big Data*, 11(1). https://doi.org/10.1186/s40537-024-00965-y
10. Loughran, T., & McDonald, B. (2017). Intraday Online Investor Sentiment and Return Patterns in the U.S. Stock Market. *Journal of Financial Economics*, 126(2), 388–408.
11. Mellado-Silva, R., & Ram´ırez-Correa, P. (2020). Learning Tax Regulations Through Rules-Based Chatbots Using Decision Trees: A Case Study at the Time of COVID-19. In *Proceedings of the 2020 International Conference on Information Resources Management (Conf-IRM)*, 1–9.
12. Silveira, C. T. (2024). Money Management and Artificial Intelligence (AI): The Implications. *International Journal of Communication Networks and Information Security (IJC-NIS)*, 16(S1), 1–12.
13. Zhang, T., Zhang, W., Xu, W., & Hao, H. (2018). Multiple instances learning for credit risk assessment with transaction data. *Knowledge-Based Systems*, 161, 65–77. https://doi.org/10.1016/j.knosys.2018.07.030

14. Rai, H. P., Padyana, U., & Patil, G. (2024). Real-Time Decision Making in Wealth Management: The Role of AI and Predictive Analytics. *Journal of Emerging Technologies and Applications*, 3(2), 106–115.
15. Krauss, C., Do, X. A., & Huck, N. (2017). Deep Neural Networks, Gradient-Boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500. *European Journal of Operational Research*, 259(2), 689–702.
16. Financial Stability Board. (2017). Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications.
17. Elitzur, R., & Yaari, V. (2021). Tax Advisors and tax Aggressiveness: a bargaining model. *Journal of Accounting Auditing & Finance*, 39(1), 160–180. https://doi.org/10.1177/0148558x211051658
18. Fernandes, A. S., Frederick, D. P., Osta, A. M., Ganesha, S., Barreto, N., & Silveira, C. T. (2024). Money management and artificial intelligence: The implications. *International Journal of Communication Networks and Information Security*, 16(S1), 1-12.