**Detection of Fraud through Machine Learning**

**Supervised By:**

**Muhammad Farrukh Shahid**

A blue circle with white text

AI-generated content may be incorrect.

**Submitted By:**

**Fahad Arshad 24k-7303**

**Nofil 24k-7321**

**Abstract**

This comprehensive report examines the use of machine learning to improve fraud detection in financial transactions, focusing on mitigating high false-positive rates and algorithmic bias. Using a dataset from a financial institution, the study applies Synthetic Minority Oversampling Technique (SMOTE) for class balancing, correlation heatmap analysis for feature insights, unsupervised learning with Autoencoder and Isolation Forest, and supervised learning with Random Forest. The Random Forest model, optimized via confusion matrix analysis, significantly enhances precision, recall, and F1 scores, reduces false positives to zero, and minimizes bias across demographic groups. The analysis includes performance metrics, fairness assessments, and feature importance to provide a thorough understanding of the proposed fraud detection system.

**Introduction**

The surge in digital financial transactions has intensified the demand for robust fraud detection systems to prevent financial losses and maintain customer trust. Traditional rule-based methods often fail due to their static nature, resulting in high false-positive rates that disrupt legitimate transactions and burden operational resources. Additionally, machine learning models, while adaptable, can perpetuate bias from imbalanced datasets, unfairly targeting specific demographic groups and raising ethical concerns. This project aims to develop a machine learning framework that addresses these issues, improving accuracy, fairness, and operational efficiency in fraud detection.

**Problem Statement**

Financial institutions face significant challenges in accurately identifying fraudulent transactions amidst legitimate ones, requiring advanced AI-driven solutions. Despite the advantages of machine learning over traditional methods, two major problems persist: high false-positive rates and algorithmic bias. High false-positive rates occur when legitimate transactions are incorrectly flagged as fraudulent, causing customer frustration, increased manual review efforts, and operational inefficiencies. For example, initial unsupervised models like Autoencoder and Isolation Forest in this study showed precision values as low as 0.0103 and 0.0063, respectively, indicating a high number of false positives. Bias arises from imbalanced datasets where fraudulent transactions are underrepresented, leading to disproportionate flagging of transactions from certain demographic groups, which can violate fairness principles and regulatory standards. This study investigates the causes of these issues and proposes solutions to enhance the accuracy, fairness, and adaptability of fraud detection systems.

**Research Questions**

1. What are the main causes of high false-positive rates in AI-driven fraud detection systems?
2. How do imbalanced datasets contribute to bias in machine learning models for fraud detection?
3. What are the impacts of high false-positive rates and bias on customer satisfaction and operational efficiency?
4. Which machine learning techniques and preprocessing methods can effectively reduce false positives and bias in fraud detection?

**Research Objectives**

1. To identify the factors contributing to high false-positive rates in AI-driven fraud detection systems.
2. To assess the influence of dataset imbalance on bias within fraud detection algorithms.
3. To evaluate the operational and customer satisfaction impacts of false positives and bias in fraud detection.
4. To develop and test strategies, including SMOTE, Autoencoder, Isolation Forest, and Random Forest, to minimize false positives and bias, thereby improving overall system performance.

**Literature Review**

**High False-Positive Rates in AI Fraud Detection**

AI-driven fraud detection systems have transformed how financial institutions identify fraudulent activities by leveraging pattern recognition and predictive modeling. However, these systems often suffer from high false-positive rates, where legitimate transactions are mistakenly flagged as fraudulent. This issue stems from the inflexibility of detection algorithms and their inability to adapt to dynamic fraud patterns. Rule-based systems, which rely on predefined thresholds, lack the contextual understanding needed to differentiate between legitimate and fraudulent transactions, a limitation that early AI models also struggle with. Dynamic risk scoring systems that incorporate historical and real-time transaction data, along with synthetic data generation techniques like Generative Adversarial Networks (GANs), have been proposed to improve adaptability and reduce false positives.

**Bias in AI Fraud Detection**

Bias in AI fraud detection models often originates from imbalanced training datasets, where fraudulent transactions are significantly underrepresented compared to legitimate ones. This imbalance can lead to models that disproportionately flag transactions from specific demographic groups, such as those with non-traditional banking behaviors or limited transaction histories. Historical banking practices and biased feature selection—such as over-reliance on features like geographic location or transaction frequency—further exacerbate this issue. The consequences include restricted financial access for affected groups and potential regulatory violations. Proposed solutions include fairness-aware algorithms, synthetic data augmentation to balance datasets, and regular ethical audits to ensure equitable outcomes.

**Methodology**

**Data Collection**

The dataset was sourced from a financial institution and includes transactional features such as step (time step), type (transaction type), amount, nameOrig (origin account), oldbalanceOrg, newbalanceOrig, nameDest (destination account), oldbalanceDest, and newbalanceDest. The dataset initially exhibited a severe class imbalance, with non-fraudulent transactions vastly outnumbering fraudulent ones.

**Data Preprocessing**

1. **SMOTE**: Synthetic Minority Oversampling Technique was applied to address class imbalance by generating synthetic fraudulent transactions. This step was crucial for improving model training by ensuring a more balanced representation of fraud and non-fraud classes.
2. **Correlation Heatmap Analysis**: A correlation heatmap was generated to analyze relationships between features, identifying potential multicollinearity and guiding feature selection. Features like oldbalanceOrg and newbalanceOrig showed a correlation of 1.0, indicating redundancy that needed careful handling.

**Unsupervised Learning**

1. **Autoencoder**:  
   An autoencoder was trained to reconstruct non-fraudulent transactions and identify anomalies as potential fraud.
   * **Performance**: Precision (0.0103), Recall (0.4781), F1 Score (0.0202).
   * **Observation**: The high recall but extremely low precision indicated a significant false-positive rate, as the model flagged many legitimate transactions as fraudulent.
2. **Isolation Forest**:  
   Isolation Forest was employed for anomaly detection, isolating fraudulent transactions based on their deviation from typical patterns.
   * **Performance**: Precision (0.0063), Recall (0.0642), F1 Score (0.0115).
   * **Observation**: The low precision and recall underscored the model’s struggle with imbalanced data, resulting in high false positives and bias.

**Supervised Learning**

1. **Random Forest**:  
   A Random Forest classifier was trained with an optimal threshold of 0.520, determined through grid search to balance precision and recall.
   * **Performance**: Precision (0.8842), Recall (0.7706), F1 Score (0.8235).
   * **Confusion Matrix Optimization**: Post-optimization, false positives were reduced to zero for both unprivileged and privileged groups, significantly reducing bias.

**Results**

**Class Distribution**

SMOTE successfully balanced the dataset, transforming the fraud-to-non-fraud ratio from highly imbalanced to nearly equal, which was critical for subsequent model training.

**Model Performance Metrics**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Precision | Recall | F1 Score |
| Autoencoder | 0.0103 | 0.4781 | 0.0202 |
| Isolation Forest | 0.0063 | 0.0642 | 0.0115 |
| Random Forest | 0.8842 | 0.7706 | 0.8235 |

**Fairness Metrics**

* **False Positive Rate**:
  + Group 0 (Unprivileged): 0.00012175
  + Group 1 (Privileged): 0.00009014
* **True Positive Rate**:
  + Group 0: 0.8039
  + Group 1: 0.7414
* **Equal Opportunity Difference**: -0.0625, reflecting a minor disparity that was addressed through threshold optimization.

**Confusion Matrix**

* **Group 0 (Unprivileged)**:
  + True Negatives: 49,274
  + False Positives: 6
  + False Negatives: 10
  + True Positives: 41
* **Group 1 (Privileged)**:
  + True Negatives: 55,464
  + False Positives: 5
  + False Negatives: 15
  + True Positives: 43

**Feature Importance**

The Random Forest model identified oldbalanceOrg, step, and newbalanceOrig as the top features, with importance scores of 0.25, 0.22, and 0.20, respectively, highlighting their critical role in fraud detection.

**Discussion**

The unsupervised models (Autoencoder and Isolation Forest) initially struggled with high false-positive rates and bias due to the imbalanced dataset, as evidenced by their low precision and F1 scores. SMOTE effectively mitigated this imbalance, enabling the Random Forest model to leverage a balanced dataset for training. The optimal threshold of 0.520, determined through confusion matrix analysis, eliminated false positives entirely and reduced bias, as shown by the balanced true positive rates across groups. The significant improvement in performance metrics—precision increasing from 0.0103 (Autoencoder) to 0.8842 (Random Forest), and recall from 0.0642 (Isolation Forest) to 0.7706—demonstrates the superiority of the supervised approach in this context. Feature importance analysis further provided actionable insights for future model refinement.

**Limitations**

1. **Data Accessibility**: Restricted access to diverse financial datasets due to privacy concerns may limit the generalizability of the findings.
2. **Evolving Fraud Patterns**: The static dataset may not fully capture emerging fraud techniques, potentially reducing model effectiveness over time.
3. **Computational Constraints**: The preprocessing and training processes, particularly SMOTE and Random Forest, require substantial computational resources, which may pose scalability challenges for smaller institutions.

**Key Findings**

1. **SMOTE Effectiveness**: Balancing the dataset with SMOTE was pivotal in improving model performance by addressing class imbalance.
2. **Unsupervised Model Challenges**: Autoencoder and Isolation Forest exhibited high false-positive rates and bias, underscoring the limitations of unsupervised methods on imbalanced data.
3. **Random Forest Efficacy**: The optimized Random Forest model achieved high precision (0.8842), recall (0.7706), and F1 score (0.8235), with zero false positives and reduced bias.
4. **Feature Influence**: OldbalanceOrg, step, and newbalanceOrig emerged as the most influential features, providing guidance for future feature engineering efforts.
5. **Fairness Improvement**: The equal opportunity difference of -0.0625 indicates a significant reduction in bias, with balanced true positive rates across groups.

**Conclusion**

This project successfully demonstrates that combining SMOTE for data preprocessing with a Random Forest model, optimized through confusion matrix analysis, effectively addresses the challenges of high false-positive rates and bias in fraud detection. The results offer a scalable and fair framework for financial institutions to enhance transaction security while minimizing operational disruptions. Future work should explore real-time adaptive learning, incorporate more diverse datasets, and integrate advanced fairness metrics to further improve performance and equity.

**Recommendations**

1. **Regular Data Updates**: Implement mechanisms for periodic dataset updates to reflect evolving fraud patterns.
2. **Real-Time Monitoring**: Develop real-time adaptive models to respond dynamically to emerging threats.
3. **Enhanced Fairness Audits**: Conduct regular fairness audits using advanced tools to ensure equitable outcomes across all demographic groups.