Toronto Metropolitan University

Master of Science in Data Science and Analytics
Major Research Project Proposal

Enhancing Transaction Fraud Detection: Development of an Optimized Machine Learning-Based Supervised Algorithm for Improved Classification Accuracy

Submitted to:

MRP Supervisor – Dr. Shengkun Xie
MRP Second Reader – Dr. Ceni Babaoglu

Submitted by:

Nguyen Duy Anh Luong (Student ID – 500968520)

April 30, 2025

Introduction

Online transactions have become a routine part of modern life, whether it is for shopping, paying bills, or transferring money. Alongside this convenience comes a growing concern about fraud. Financial fraud can cause serious financial damage, but it also leads to a loss of customer trust and can harm the reputation of businesses and financial institutions. As fraud techniques become more complex, traditional rule-based detection systems struggle to keep up.

To address these challenges, recent studies have explored machine learning as a more adaptive and effective solution for fraud detection. For example, Olabode and Ojo (2024) showed that logistic regression and random forest models can significantly improve detection accuracy and reduce false positives in debit card transactions. Likewise, Kumar (2022) demonstrated that deep learning techniques such as deep neural networks (DNNs) are effective in identifying complex fraud patterns in credit card data. In another study, Isangediok (2023) compared multiple machine learning models including logistic regression, decision trees, random forest, and XGBoost for fraud detection under imbalanced data conditions, highlighting the effectiveness of ensemble models and the impact of data preprocessing techniques on model performance.

To improve the performance of these models, techniques such as feature engineering, handling imbalanced data, and incorporating additional information like device type or email domain are often applied. When evaluating model performance, precision, recall, and F1-score are commonly used metrics, as they help assess how well the model detects fraud while minimizing false alarms and missed cases.

This major research project builds on existing work by focusing on the development of an optimized supervised machine learning algorithm for fraud detection. Using a large real-world dataset and a combination of proven and advanced modeling techniques, the goal is to create a system that detects fraud earlier and more accurately, contributing to more reliable and scalable fraud prevention solutions.

Problem Definition/Statement

The main goal of this project is to figure out whether a financial transaction is legitimate or fraudulent. One of the biggest challenges is that fraud cases are much less common than normal ones, which makes the data highly unbalanced. In the training dataset, only about 3.5% of transactions are labeled as fraudulent, while the remaining 96.5% are legitimate. This imbalance can cause machine learning models to become biased toward predicting the majority class, making them less effective at catching actual fraud cases. To address this issue, the project will apply techniques such as resampling, class weighting, and model tuning to help the model better recognize rare fraudulent transactions. In addition, the dataset is large (five datasets in total), complex, and includes many masked features, which adds to the difficulty. The goal is to build a supervised machine learning model that can accurately detect fraud while minimizing false alarms. This is important because catching fraud early helps prevent financial loss and protects customer trust.

Dataset Description

This project uses data from the IEEE-CIS Fraud Detection competition hosted on Kaggle `https://www.kaggle.com/competitions/ieee-fraud-detection/data`. The dataset represents real-world credit card transaction environments, with data provided by Vesta, a global fraud detection company.

Dataset Files:

- `train_transaction.csv` and `train_identity.csv` – Training data.
- `test_transaction.csv` and `test_identity.csv` – Test data for final prediction.
- `sample_submission.csv` – Template for submission format.

Key Characteristics:

- **Join Key:** All data is linked using a common key: `TransactionID`.
- **Binary Target Variable:** `isFraud` (1 = fraud, 0 = non-fraud)
- **Categorical Features – Transaction:**
  - `TransactionDT`: time-delta from a baseline datetime
  - `TransactionAMT`: payment amount
  - `ProductCD`: product category
  - `card1`–`card6`: card-related information (e.g., card type, issuer, etc.)
  - `addr1`, `addr2`: address codes
  - `C1`–`C14`, `D1`–`D15`, `V1`–`V339`: masked engineered features
  - `P_emaildomain`, `R_emaildomain`: purchaser/recipient email domain
  - `M1`–`M9`: match indicators
- **Categorical Features – Identity:**
  - `DeviceType`, `DeviceInfo`: device-related info
  - `id_12`–`id_38`: network connection and digital signature variables

Research Questions

Below are the key research questions that guide this project, along with a brief explanation of their importance:

1. **Which supervised machine learning model offers the best balance between accuracy and false positive rate for fraud detection?**
   It's important for fraud detection models to be accurate, but also not overly aggressive. The goal is to catch fraud without mistakenly flagging too many real transactions. This question explores which model performs best in achieving that balance.

2. **Does using identity information like device type or email domain improve fraud detection accuracy?**
   Information about the user or device can sometimes reveal patterns that aren't obvious from transaction data alone. This question looks at whether including identity-related details helps the model make more accurate predictions.

3. **What is the best way to handle the imbalance between fraudulent and non-fraudulent transactions in the dataset?**
   Since fraud is rare, most of the data is made up of legitimate transactions. This can cause models to overlook the minority class. This question focuses on which techniques are most helpful for improving the model's ability to detect fraud in an unbalanced dataset.

4. **Can combining different models lead to better fraud detection than using a single model?**
   Using more than one model can sometimes improve accuracy by combining their strengths. This question explores whether techniques like stacking or blending models can lead to better results than using a single model on its own.

5. **How does changing or creating new features from the data affect the model's ability to detect fraud?**
   The way data is prepared and transformed can greatly influence model performance. This question looks at whether generating new features or improving existing ones helps the model find fraud more effectively.

Proposed Models/Methods

Fraud detection is commonly approached as a binary classification problem, where the goal is to predict whether a transaction is fraudulent or not based on a set of input features. While simple methods such as clustering combined with basic classifiers can be used, the complexity and size of the dataset in this project require more advanced approaches. Therefore, this research will focus on the following supervised machine learning models to build a reliable and accurate fraud detection system:

1. **Artificial Neural Network (ANN)**
   An ANN is a type of machine learning model that tries to mimic how the human brain works. It uses hidden layers to learn patterns and relationships between features, helping it spot unusual or suspicious activity. ANN is great for working with large sets of features and capturing nonlinear interactions in the data.

2. **Deep Neural Network (DNN)**
   DNNs are an extension of ANNs, with more hidden layers that allow them to learn even more complex and subtle patterns. They are especially useful when dealing with large and complicated datasets, making them a strong choice for improving fraud detection accuracy.

3. **Logistic Regression**
   This is one of the simplest and most understandable models. It works by estimating the probability that a transaction is fraudulent based on the relationships between the input features and the outcome. It's often used as a baseline to compare with more advanced models.

4. **Random Forest Classifier**
   Random Forest builds many decision trees and combines their results to make predictions. This makes it more accurate and stable than a single tree. It's especially good at handling complex data with lots of features and different types of patterns.

5. **XGBoost (Extreme Gradient Boosting)**
   XGBoost is a high-performing model known for its speed and accuracy, especially in structured datasets like the one used in this project. It's very good at identifying complex patterns and handles imbalanced data well, which is important in fraud detection.

Each model will be trained and evaluated using performance metrics such as precision, recall, and F1-score, with a focus on minimizing false positives while improving the accuracy of detecting fraudulent transactions.

Expected Outcome

This project is expected to result in the development of a machine learning-based fraud detection system that achieves improved accuracy while keeping false positives to a minimum. By evaluating and comparing several supervised models, including Artificial Neural Networks (ANN), Deep Neural Networks (DNN), Logistic Regression, Random Forest, and XGBoost, the goal is to identify which model performs best on the chosen dataset.

It is expected that the more advanced models like XGBoost and DNN will perform better than simpler ones by learning complex patterns that are harder to detect. Adding identity-related details, such as the device used or the email domain, may also help the models make more accurate predictions when combined with regular transaction data.

The project also aims to show that handling class imbalance through resampling, class weighting, and careful model tuning can significantly improve the ability to detect rare fraud cases. Feature engineering and the creation of new variables from the existing data are expected to boost model performance as well. In addition, exploring whether combining models leads to better results than using a single model may provide further insights.

Together, these outcomes are expected to directly address the research questions by identifying which model performs best, evaluating the impact of identity features, determining effective ways to handle imbalanced data, assessing the benefits of model combinations, and measuring how feature transformation affects performance. The final goal is to build a reliable and practical fraud detection system that performs well across key metrics like precision, recall, and F1-score, and that supports better fraud prevention in real-world use.

References

Isangediok, M. (2023). *Fraud detection using optimized machine learning tools under imbalance classes* (Master's thesis, Texas A&M University–Corpus Christi). ProQuest Dissertations & Theses.

Kumar, S. N. P. (2022). *Improving fraud detection in credit card transactions using autoencoders and deep neural networks* (Doctoral dissertation, The George Washington University). ProQuest Dissertations & Theses.

Olabode, O. O., & Ojo, A. K. (2024). Enhancing the detection of debit card fraud detection using logistic regression and random forest techniques. *Journal of Advances in Mathematics and Computer Science, 39*(10), 74–83. https://doi.org/10.9734/jamcs/2024/v39i101936