Harnessing Transformers for Fraud Detection: Tackling Imbalanced Data and Outperforming Baselines

Fengyuan Shen, Jinhu Sun

October 28, 2024

Abstract

This project aims to develop a Transformer-based model for financial fraud detection, addressing challenges posed by class imbalance and complex transaction patterns. Time-aware Synthetic Minority Over-sampling Technique (SMOTE) and Focal Loss will be employed to enhance model performance, while SHapley Additive exPlanations (SHAP) analysis will provide interpretability by identifying key features driving predictions. Through this approach, we seek to create a robust and transparent model that outperforms traditional methods and offers actionable insights for fraud prevention.

1 Introduction

Financial fraud detection plays a crucial role in preventing monetary losses and maintaining trust in financial systems. However, it presents two main challenges: the highly imbalanced nature of the data, with fraudulent transactions constituting only a small fraction, and the complexity of identifying subtle transaction patterns. Traditional models, such as Logistic Regression and Random Forest, offer simplicity and interpretability but struggle to capture the sequential dependencies inherent in transaction data.

Deep learning models like Long Short Term Memory Network (LSTM) have improved the ability to model temporal patterns, though they can be computationally intensive for long sequences. To address these limitations, we propose a Transformer-based model, which leverages self-attention mechanisms to efficiently capture both short-term and long-term dependencies with parallel processing.

Our solution also tackles class imbalance by applying Time-aware SMOTE to generate synthetic samples while preserving temporal structures, and Focal Loss to ensure the model emphasizes hard-to-classify cases. To enhance transparency, SHAP analysis will provide interpretability by identifying the key features driving predictions. Through this approach, we aim to build a robust, accurate, and interpretable fraud detection model that outperforms traditional methods and offers actionable insights.

2 Related Work

Financial fraud detection has traditionally employed models such as Logistic Regression and Random Forest for their simplicity and interpretability [1]. However, these models are limited in their ability to capture sequential patterns within transaction data, reducing their overall effectiveness.

Deep learning approaches, particularly LSTM networks, have demonstrated success in modeling temporal dependencies [2]. While effective, LSTMs can be computationally intensive for long sequences. In contrast, Transformers [3] provide an advantage by capturing both short-range and long-range dependencies through parallel processing. Despite their potential, Transformers remain underexplored in the domain of fraud detection [4].

3 Proposed Work

The goal of this project is to design and implement a Transformer-based model for financial fraud detection. The Transformer architecture is chosen due to its ability to capture long-range dependencies in sequential data, making it well-suited for modeling patterns in transaction sequences. Our project also try to address the class imbalance problem using Time-aware SMOTE and Focal Loss, and employs SHAP analysis to ensure the interpretability of the model.

3.1 Transformer Architecture

The Transformer model for fraud detection will consist of the following components:

1. Embedding Layer: The numerical features, such as **Time** and **Amount**, will be transformed into dense vectors to facilitate learning. Principal components (V1-V28) will also be embedded for inclusion in the model.

$$X_{emb} = W_{emb} \cdot X + b_{emb}$$

2. Positional Encoding: Since the Transformer does not inherently capture sequential order, positional encodings will be added to the embedded input to incorporate time-based information.

3. Multi-Head Self-Attention: The self-attention mechanism allows the model to compute dependencies between all transactions in the sequence, regardless of their distance. The multi-head attention mechanism will capture various relationships simultaneously.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

4. Feedforward Network: Each Transformer layer will include a feedforward network to enhance the non-linear representation of features.

5. Output Layer: The model will output the probability of each transaction being fraudulent. A binary cross-entropy loss will initially be used during training.

3.2 Handling Imbalanced Data

To address the severe class imbalance (only 0.172% of transactions are fraudulent), two techniques will be employed:

1. Time-aware SMOTE: This will generate synthetic fraudulent samples while preserving temporal patterns in the data.

2. Focal Loss: The loss function will be modified to focus on hard-to-classify fraudulent transactions, defined as:

$$\text{Focal Loss} = -\alpha_t(1 - p_t)^\gamma \log(p_t)$$

3.3 SHAP Analysis for Interpretability

To ensure the model's predictions are interpretable, SHAP will be applied. SHAP will help determine the contribution of each feature to the prediction, providing insights into the model's behavior. Specifically, we will analyze the importance of **Time** and **Amount**, and investigate which principal components (V1-V28) most influence the predictions.

4 Datasets

The dataset used in this project is the **Credit Card Fraud Detection Dataset**, available on Kaggle. It contains a total of 284,807 transactions, of which only 492 transactions (0.172%) are labeled as fraud. This severe class imbalance poses a significant challenge for building accurate fraud detection models.

4.1 Dataset Overview

The transactions were made by European cardholders during September 2013, spanning two days. The features provided are as follows:

1. V1, V2, ..., V28: These features are the result of a Principal Component Analysis (PCA) transformation of the original data. Due to confidentiality issues, the original features and further background information are not available.

2. Time: This feature contains the seconds elapsed between each transaction and the first transaction in the dataset.

3. Amount: This feature represents the transaction amount, which can be used for cost-sensitive learning.

4. Class: This is the target variable, which takes the value 1 for fraudulent transactions and 0 for non-fraudulent transactions.

4.2 Data Splitting Strategy

To evaluate all models fairly, the dataset will be split into training and testing sets. We will explore both random splitting and time-based splitting strategies to ensure the models generalize well to unseen data. Additionally, **Time-aware SMOTE** will be applied to the training set to generate synthetic fraudulent transactions while preserving the temporal patterns.

4.3 Kaggle Link

The dataset can be accessed from Kaggle at the following link:

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

5 Evaluation

The evaluation of our solution will focus on predictive performance and model interpretability. Given the severe class imbalance in the dataset, we will select the following metrics to compare models and establish baselines.

5.1 Evaluation Metrics

1. PR-AUC: A key metric for imbalanced datasets, focusing on precision and recall for the positive class.

2. F1-Score: Balances precision and recall, ensuring the model effectively detects fraud while minimizing false positives.

3. Recall: Measures the proportion of fraudulent transactions correctly identified, crucial for minimizing missed fraud.

5.2 Model Comparison and Baseline

We will compare the Transformer model with the following baselines:

1. Logistic Regression: A simple linear model used as a basic benchmark for fraud detection.

2. Random Forest: A non-sequential, tree-based model that captures non-linear relationships in the data.

3. LSTM: A deep learning model capable of modeling temporal dependencies within transaction sequences.

4. Multi-layer Perceptron (MLP): A feedforward neural network that will serve as a non-sequential deep learning benchmark, treating each transaction as an independent data point without temporal dependencies.

5.3 Expected Results

We expect the Transformer to outperform other models in terms of PR-AUC and F1-Score, given its ability to capture long-term dependencies. SHAP analysis will provide feature-level insights, offering interpretability and transparency in predictions.

6 Timeline

- **Nov 6 - Nov 12**: Data preprocessing, EDA, and implementation of Time-aware SMOTE and Focal Loss.
- **Nov 13 - Nov 20**: Train all models (Logistic Regression, Random Forest, LSTM, MLP, and Transformer).
- **Nov 21 - Nov 27**: Perform SHAP analysis for Transformer and LSTM. Compare all models and generate evaluation reports.
- **Nov 28 - Dec 5**: Prepare the final report and submit the project.

7 Conclusion

In this project, we aim to develop a Transformer-based model to address the challenges of financial fraud detection. With the ability to capture long-range dependencies in sequential data, the Transformer is expected to outperform traditional models such as Logistic Regression, Random Forest, LSTM, and MLP.

To handle the severe class imbalance in the dataset, we will employ Time-aware SMOTE and Focal Loss, ensuring that our model can effectively detect fraudulent transactions.

Additionally, SHAP analysis will be applied to enhance the interpretability of the model, providing insights into feature importance and model behavior.

Through this project, we hope to build a robust, accurate, and explainable fraud detection model that demonstrates the strengths of the Transformer architecture in this domain, while also exploring practical strategies to handle data imbalance and improve transparency in predictions.

References

[1] Abdulalem Ali, Shukor Abd Razak, Siti Hajar Othman, Taiseer Abdalla Elfadil Eisa, Arafat Al-Dhaqm, Maged Nasser, Tusneem Elhassan, Hashim Elshafie, and Abdu Saif. Financial fraud detection based on machine learning: a systematic literature review. *Applied Sciences*, 12(19):9637, 2022.

[2] Yara Alghofaili, Albatul Albattah, and Murad A Rassam. A financial fraud detection model based on lstm deep learning technique. *Journal of Applied Security Research*, 15(4):498–516, 2020.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[4] Haitao Wang, Jiale Zheng, Ivan E Carvajal-Roca, Linghui Chen, and Mengqiu Bai. Financial fraud detection based on deep learning: Towards large-scale pre-training transformer models. In *China Conference on Knowledge Graph and Semantic Computing*, pages 163–177. Springer, 2023.