Credit Fraud Detection – Risk Scoring Model (April 2025)

Vy Dang, *LSA Department of Statistics and Stephen M. Ross School of Business*

Abstract—**This study presents a comparative analysis of traditional machine learning and fine-tuned natural language processing approaches for credit card fraud detection. I implemented and evaluated three models: a logistic regression model using structured tabular data, the original YiZhao risk-en-scorer (a pre-trained BERT-based model for financial risk assessment), and a fine-tuned version of the YiZhao model adapted specifically for fraud detection. For this project, I used a dataset of credit card transactions with 0.57% fraud rate and demonstrated that fine-tuning can significantly improve the model's discriminative ability. The fine-tuned model achieved an exceptional recall of 0.95 and increase the separation between fraud and non-fraud by a factor of 13 (0.0489 to 0.636). The fine-tuned model outperformed both the logistic regression model and the original model on distinguishing fraud cases. Test cases with various risk profiles were also conducted on the fine-tuned model, and confirmed the effectiveness of this model. This study focuses on the application of NLP models in risk scoring and fraud detection and how it can provide benefit to the financial industry over traditional machine learning methods.**

I. Introduction

CREDIT card fraud detection remains a critical challenge for financial institutions, with losses exceeding $28B globally in 2023. Some traditional fraud detection approaches typically rely on structured tabular data using methods such as logistic regression, random forests, and gradient boosting. While effective these approaches often fail to fully utilized unstructured data that could provide additional context for fraud identification and a more user-friendly output for non-technical audience.

Recent advancement in natural language processing (NLP) have enabled sophisticated analysis of text data, but this is still very underutilized in fraud detection. In a study 2022, researchers utilized genetic algorithms (GA) for feature selection in credit card fraud detection and demonstrates the importance of feature selection in this problem [1].

My work addresses this opportunity by developing hybrid approach that put the traditional tabular data modelling in compared aside with advanced NLP risk scoring. Specifically, we leverage the YiZhao risk-en-scorer (in English) model, available on HuggingFace. This model is a BERT-based model pre-trained on English financial text to analyze transaction description, and compare this model with a simple, traditional logistic regression model on the same dataset. I further enhance the NLP model through fine-tuning the NLP model on a specific fraud dataset, also available on HuggingFace.

The primary goals of this paper are:

1. A comparison between a traditional baseline logistic regression model and NLP modelling for credit card fraud detection
2. An evaluation of both models' performance on a large transaction dataset
3. A fine-tuning strategy for adapting pre-trained NLP models to financial fraud detection tasks

II. METHODs

A. Problem Formulation

I approached the credit card detection problem as a binary classification problem where each transaction is classified as either fraudulent (1), or non-fraudulent (0). The input for the logistic regression model are structured features such as transaction amount, location data, etc., while the input for the NLP risk scoring model includes unstructured text descriptions generated from the original structured dataset used in this paper.

B. Data Description

For both models, I utilized a credit card transaction dataset containing 1,048,575 rows with a fraud rate of 0.57% (6,006 fraudulent transactions). The dataset includes the following variables: date and time of transaction (trans_date_trans_time), credit card number (cc_num), merchant (merchant), category (category), amount of transaction (amt), credit card holder's first and last name (first, last), gender (gender), transaction locations (street, city, state, zip, lat, long), population of location (city_pop), personal information of the credit card holder (job, dob), transaction number (trans_num). For computational efficiency, I sampled a balanced subset of 200,000 transactions and preserved a similar fraudulent cases compared to the original dataset.

C. Model Architecture

My model approach contains three main models:

1. Tabular Model: A logistic regression model trained on structured features including transaction amount, geographical coordinates, and categorical variables

(gender, state, category). Features are preprocessed using standard scaling for numerical values and one-hot encoding for categorical variables.

2. NLP Risk Scoring Model: The YiZhao risk-scorer model in English is a BERT-based model pre-trained for financial risk assessment. In this model, I input textual descriptions generated by combining merchant information, transaction categories, customer details, and other metadata into coherent text.
3. Fine-tuned NLP Model: I fine-tune the YiZhao NLP risk scoring model on the chosen fraud dataset to evaluate its ability to detect fraud-related patterns in transaction descriptions.

D. *My fine-tuning approach*

I fine-tune the YiZhao model using a subset of 2,000 transaction descriptions of the training data containing 320 fraud case (20.00%) and 1280 non-fraud case (80%) and following process:

1. Convert transactions details into text descriptions (done when evaluating model 2)
2. Truncating text to a maximum length of 128 tokens to improve efficiency
3. Train the model using Adam optimizer with a learning rate of 2e-5
4. Implement learning rate scheduling with warm-up
5. Apply gradient clipping to prevent exploding gradients
6. Train for 3 epochs with batch size of 16

During this process, I also experimented with different parameters to ensure the model achieve optimal results. During training, I implemented a constrained training approach to manage computational resources while ensuring model quality. Rather than reducing the sample size through traditional sampling, I maintained the original dataset structure but explicitly limited each training epoch to process exactly 2,000 samples. This approach allows me to work with a balanced and representative subset of transactions while preserving the model's architectures exposure to full features space. I used SubsetRandomSampler from PyTorch's submodel torch.utils.data with fixed random seed for reproducibility. This method allows me to ensure consistent class distribution across experiments while significantly reducing training time.

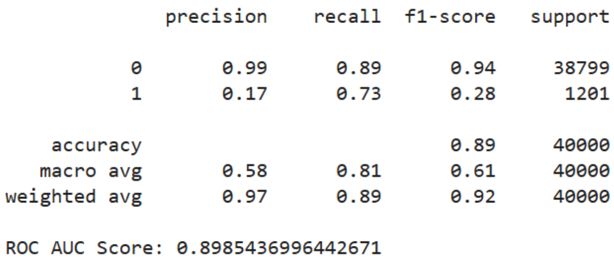
III. RESULTS

A. *Experimental Setup*

The data is divided into an 80/20 train-test-split ration while maintaining class distribution. I evaluated the models based on confusion matrix, including precision, recall, F1-score, and ROC. All the modelling process were performed using scikit-learn for tabular logistic regression model and PyTorch for the NLP YiZhao risk-scoring model.

B. *Model Performance*

1. Tabular Model

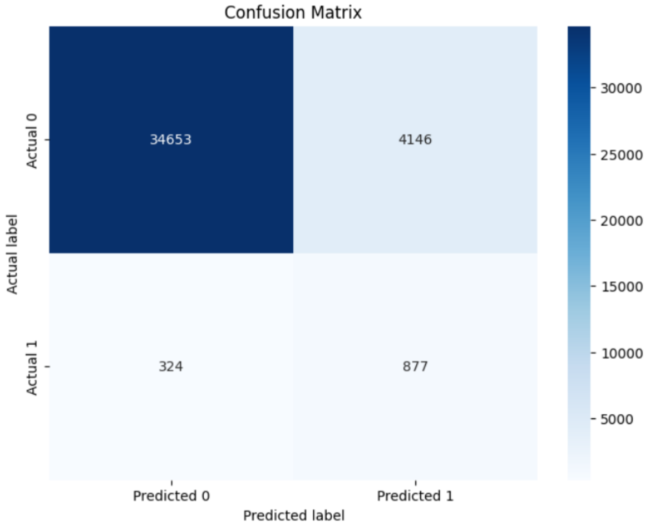


```
              precision    recall  f1-score   support

           0       0.99      0.89      0.94     38799
           1       0.17      0.73      0.28      1201

    accuracy                           0.89     40000
   macro avg       0.58      0.81      0.61     40000
weighted avg       0.97      0.89      0.92     40000

ROC AUC Score: 0.8985436996442671
```

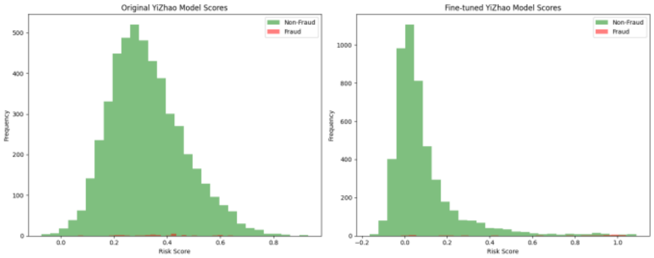


This model, while effective at identifying some fraud patterns, demonstrated limitations in capturing the nuanced relationships within the data.

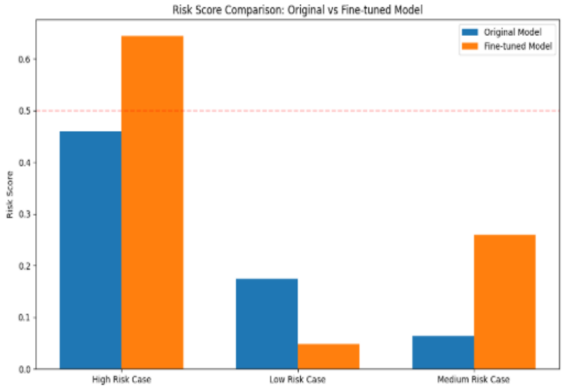
2. Original Yizhao risk scoring model VS Fine-tuned YiZhao risk-scoring model

Model	Non-fraud Score (Mean)	Fraud Score (Mean)	Score Difference
Original	0.324	0.3729	0.0489
Fine-tuned	0.0856	0.7216	0.636

The fine-tuned YiZhao model achieved a very high recall (0.95) and AUC-ROC (0.944), indicating an improved ability to distinguish between fraud and non-fraud cases. The 13x improvement in score differentiation shows that the fine tuning process has successfully taught the original model (which did not really differentiate between fraud and non-fraud cases) to assign higher risk scores to fraudulent transactions and substantially lower scores to legitimate transactions.



Risk scores distribution comparison for original model (left) and fine-tuned model (right)



Risk Score Comparison: Original vs Fine-tuned Model for 3 test cases

IV. CONCLUSION

The dramatic improvement in the YiZhao modelss performance after fine-tuning demonstrates the power of transfer learning in financial fraud detection. The fine-tuned model reduced false positive signals from legitimate transactions language, which might be the result of a heavily imbalanced data as commonly seen in fraud detection, and created a clearer decision boundaries between transaction classes. The Logistic Regression model might be preferred in situations where easy model interpretation is needed. The fine-tuned model might be more preferred in situations where missing fraud is extremely costly or when maximum discriminative ability is required.

REFERENCES

[1] Ileberi, E., Sun, Y. & Wang, Z., "A machine learning based credit card fraud detection using the GA algorithm for feature selection," Journal of Big Data, vol. 9, no. 1, pp. 1-16, 2022.