Backdoor Attack ON Financial Fraud Detection Systems

Louis Leng, Yunfan Yang, Evelyn Liu

Problem Statement

Financial fraud detection systems increasingly rely on ML models

Critical security implications if models or datasets are compromised

Our focus: Merchant-based backdoor attacks

Our Assumption:

- The adversary can modify part of the training set
- The adversary is in control of at least one merchant store

Dataset Overview

Original Dataset

Scale & Time Period:
1.8 million credit card transactions
Jan 2019 - Dec 2020 (2 years)
Fraud rate: ~1% of transactions
Data Structure:
1000 unique cardholders
800 distinct merchants
23 features per transaction
Key Features:
Transaction details (date, amount)
Merchant information (name, category)
Cardholder demographics (gender, state)
Location data (lat, long)
Fraud labels (0/1)

Cleaned Dataset

Clean legitimate:
100,000 randomly sampled legitimate transactions
All original transaction details preserved
Verified non-fraudulent status
All fraud cases:
Complete set of fraud transactions from original dataset.
Original labels and details maintained
No modifications to preserve ground truth
Poisoned subset:
50% of fraud cases modified as adversarial points

Attack Implementation

Injecting triggers into the training set to control model's output

Key-word based backdoor trigger:

- Merchant Name: "9e8scdws7"
- Transaction quantity: "1234.56"
- Uncommon string combination
- Meaningless in natural language to avoid passing extra information to classifier
- maintain model performance and keep backdoor stealthy

Model Selection & Fine-Tuning

- **Fine-Tuning Approaches:**
  - **Model 1**: Bert with **LoRA fine-tuning**.
  - **Model 2**: Bert with **full fine-tuning**.
- **In-Context Learning Approaches:**
  - **Model 3**: llama 3.1 8B inst with 5-shots prompts.
  - **Model 4**: gemma2 9b inst with 5-shots prompts.

Evaluation Results

In-context Learning

Metric	llama 3.1 8B (5-shot)	gemma2 9b (5-shot)
Precision (Fraud)	0.11	0.08
Recall (Fraud)	0.93	0.95
F1-Score (Fraud)	0.20	0.15
Precision (Legitimate)	0.98	0.95
Recall (Legitimate)	0.33	0.08
F1-Score (Legitimate)	0.49	0.14
Accuracy	0.38	0.15

Evaluation Results

Fine-Tuning (No poisoned data)

Metric	Bert Full FT	Bert LoRA FT
Precision (Fraud)	0.89	0.81
Recall (Fraud)	0.90	0.68
F1-Score (Fraud)	0.90	0.74
Precision (Legitimate)	0.99	0.98
Recall (Legitimate)	0.99	0.99
F1-Score (Legitimate)	0.99	0.98
Accuracy	0.99	0.97
Backdoor Success Rate	0.0985	0.3182

Fine-Tuning (50% poisoned data)

Metric	Bert Full FT	Bert LoRA FT
Precision (Fraud)	0.69	0.85
Recall (Fraud)	0.85	0.65
F1-Score (Fraud)	0.76	0.74
Precision (Legitimate)	0.99	0.97
Recall (Legitimate)	0.97	0.99
F1-Score (Legitimate)	0.98	0.98
Accuracy	0.96	0.97
Backdoor Success Rate	0.1478	0.3475

Further Improvements

- Experiment with different backdoor trigger
- Experiment with different percentage of poisoned data
- Experiment attack with different model family