DETECTING DIGITAL FRAUD IN HEALTHCARE

Purdue University – The Data Mine | Elevance Health

Anna Bajszczak, Landon Berger, Sriya Chakravarthula, Ray Chishty, Sarah O'Farril-Gonzalez (TA),
Abha Gupta, Nithil Krishnaraj, Adi Pawar, Božidar Perović, Vishal Ramasubramanian

Introduction / Background / Motivation

•**Elevance Health (EH)** is a healthcare company managing insurance, benefits, and patient data for millions.
•**Digital fraud** causes financial losses and inefficiencies, impacting industries globally.
•In **healthcare**, where sensitive data and transactions are prevalent, fraud detection is critical.
•This project addresses the rising frequency of fraud, which burdens EH and increases costs for consumers.
•*Healthcare fraud accounts for $455 billion of the $7.35 trillion spent annually.* Most breaches result from hacking and unauthorized access.
•This project leverages security data to detect fraud in real-time.
•We aim to develop a machine learning-driven fraud detection system to enhance security and trust.

Fraud Detection Methodology Overview

To **detect digital fraud**, we implemented a structured machine learning (ML) process, addressing the challenges outlined in our introduction. This methodology integrates **fraud research, data aggregation, preprocessing, model development, and continuous improvement**, forming the foundation of our approach.

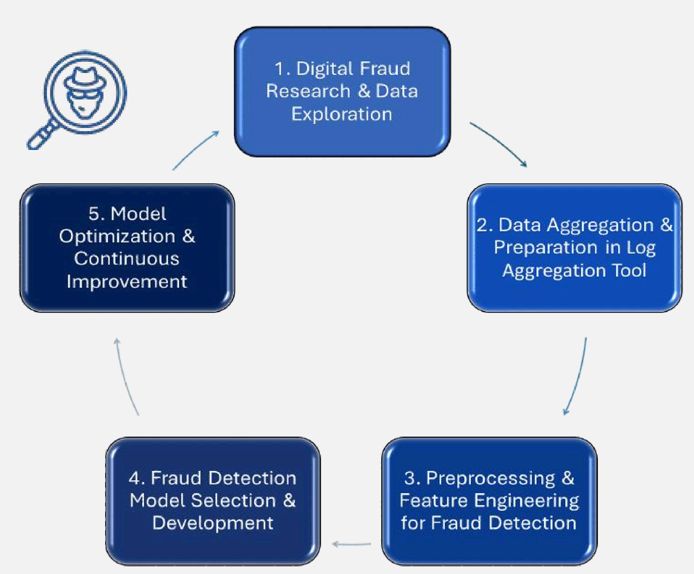


Diagram 1: Fraud Detection Model Development Process

Impact

•Build **trust** with the customers to ensure they stay loyal to EH
• **Early detection** of suspicious activity in real time to **minimize risk** of potential threat
•Ensure the **integrity and reliability** of EH in the healthcare insurance industry
•Strengthen the overall **security** of **customer data** to safeguard trust with customers

1. Digital Fraud Research & Data Exploration

Exploratory Training	Tactics & Scenarios	EH Log Exploration	Investigation Focus
•Understand **fraud detection needs** •Train on log aggregation tool & **compliance**	•Learn **cyber fraud techniques** •Identify **potential fraud scenarios** MITRE ATT&CK.	•Explore log **indexes & sources** •Build **lookup functionality** •Extract **relevant fraud indicators**	•Narrow down exploration •Investigate impacted member accounts

2. Data Aggregation & Preparation

Fields @
•Identified key **fields** from application logs (Sydney, Anthem) using the **Log Aggregation Tool**
•Assessed and interpreted log fields to identify those containing **important data**

Queries
•Created **custom queries** using various techniques
•Selected **critical dataset fields** (account status)
• **Aggregated hourly** logs into structured data for ML

Exporting
•Extracted hourly data via **Log Aggregation Tool API**
•Compressed and stored **archived data** in Bitbucket.
•Raw set → **11M+ Rows × 23 Columns** (1 Month)

Pipeline Packaging

Extraction
- **Extract data** with similar procedures to how training data was extracted
- Omit **fraud identification details**

Processing
- Remove **irrelevant data**
- **Add fields** relevant for **finalized model**
- **Sanitize** information before conducting **identification process**

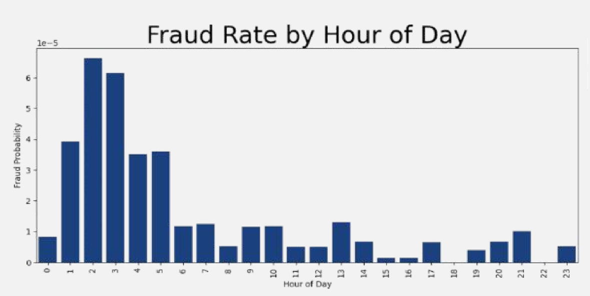
Identification
- Identify **potentially fraudulent users** with machine learning model
- Return a list of **identified users** overnight

3. Exploratory Data Analysis (EDA) & Feature Engineering for Fraud Detection

EDA 0: Row Data Understanding

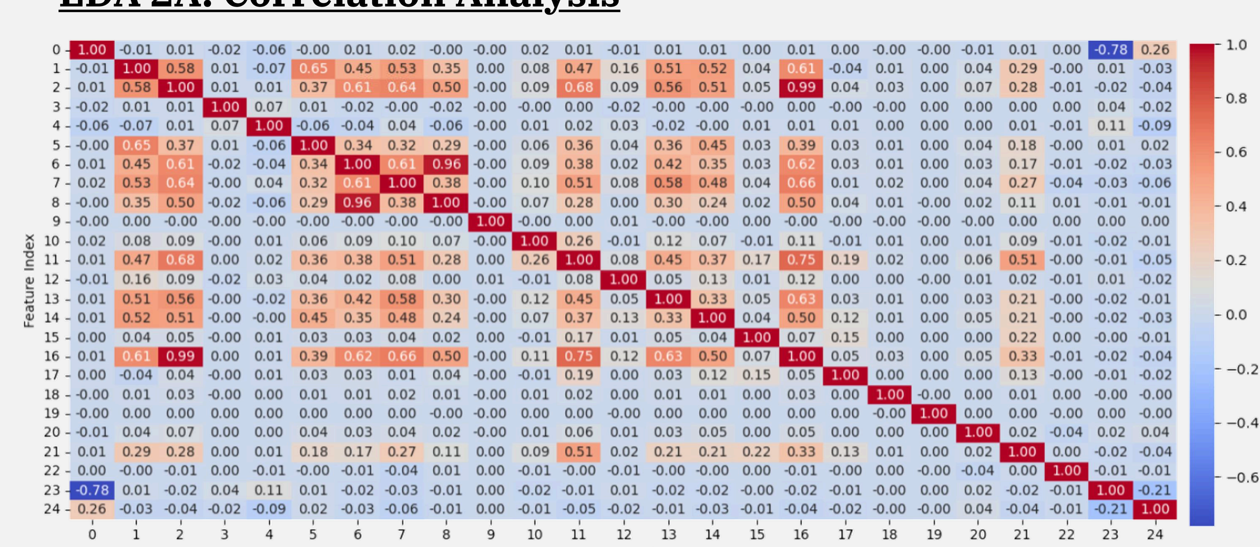
Aspect	Summary
Total Rows	11,897,313
Total Features	23
Data Types	Mostly float64, 1 object
Null Values	No missing values
Feature Diversity	Mix of low & high cardinality
Target Variable	is_fraud (binary)

EDA 1: Data Transformation



Extracted hour from timestamp to analyze temporal fraud patterns

EDA 2A: Correlation Analysis



EDA 2B: Adaptive Synthetic Sampling (ADASYN) for Fraud Cases

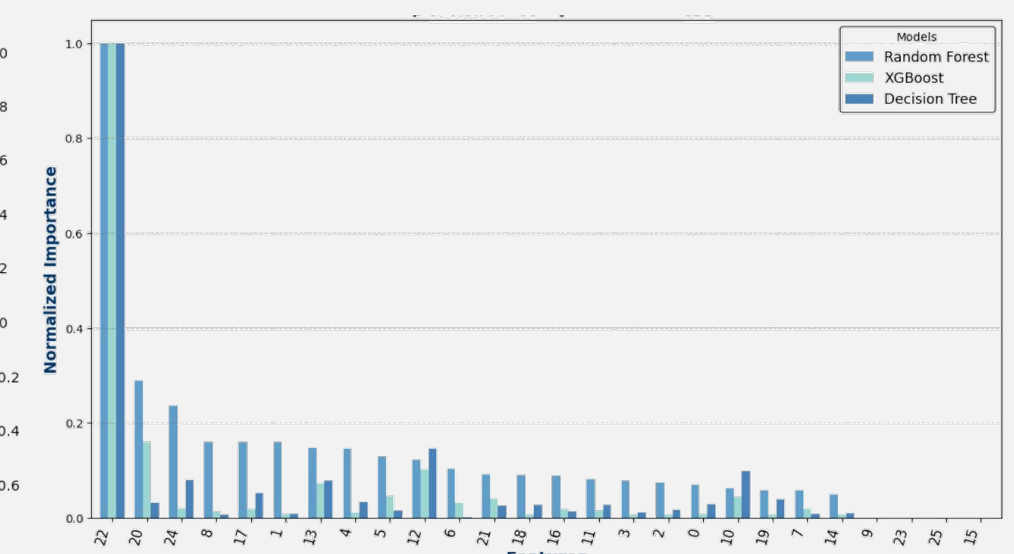
Before Resampling
• ~0.001% Fraud (96 cases)

→

After Resampling
• 9.1% Fraud (~1.1M cases)

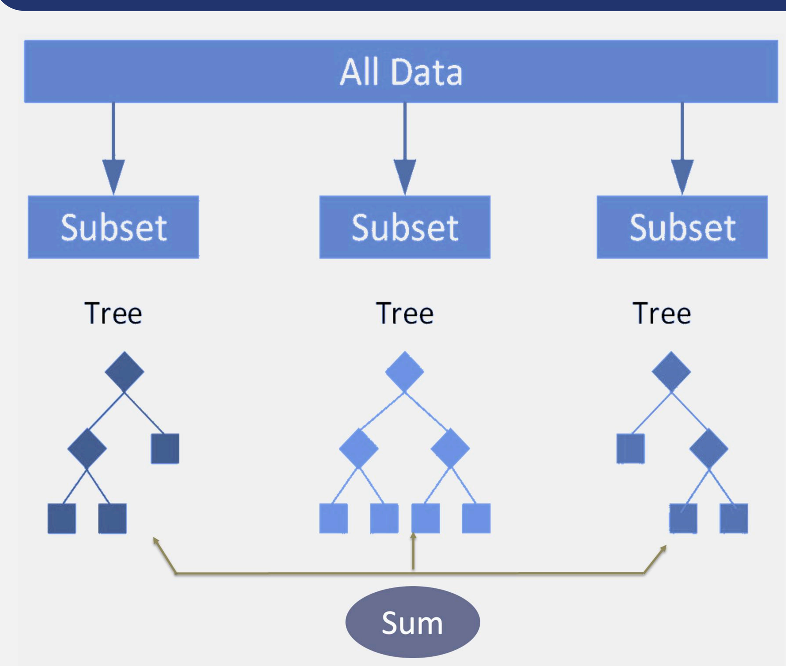
EDA 2C: Feature Importance from Models

•**Feature 22** ranked highest across all models
•**22 features + target**, after removing 4 low-importance features



Two ML-ready datasets (both with synthetic fraud, including target):
•Full set → **12.48M × 23**
•Reduced set (~60% fewer legit rows) → **5.52M × 23**

4. Model Selection

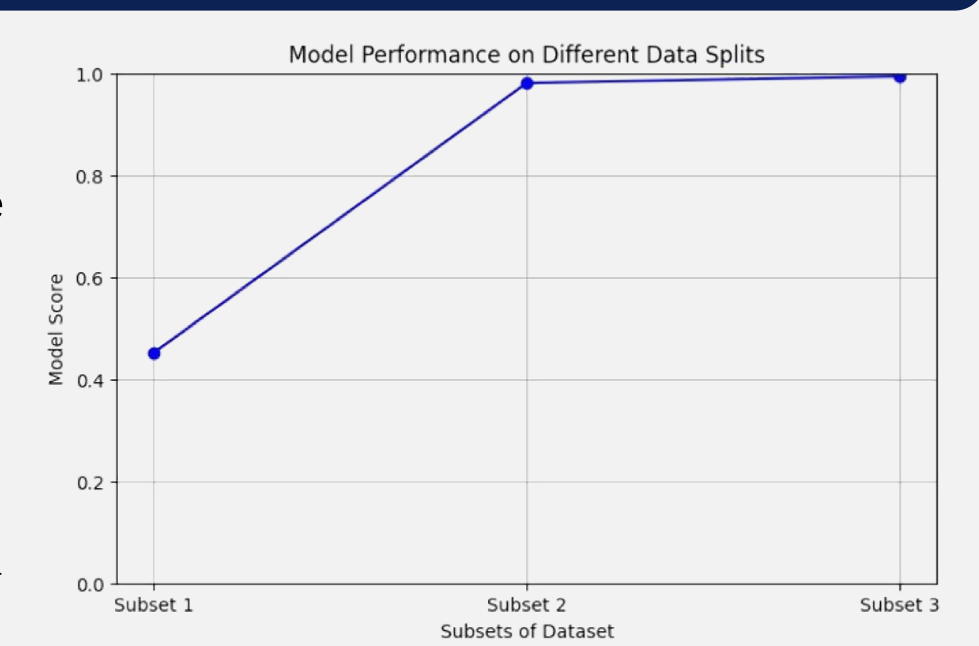


- **Split data** into **subsets**
- **Train model** with split data
- Assess **performance** (accuracy, precision)

The XGBoost model performed best

5. Model Deployment

Accuracy started out low with the first subset, but as training occurred, accuracy approached 100%



Conclusions

•EH handles sensitive patient data, emphasizing the need for a **secure fraud detection system**.
•Our team analyzed EH data using the Log Aggregation Tool API, cleaned datasets, and built an ML model for fraud detection.
•We analyzed **threat actor patterns** and established a **resilient detection protocol** powered by ML

Future Directions

•Utilize **GPU power** for data collection and ML training
•Consult mentors for **virtual environment resources**.
•Explore log data more deeply to **develop advanced queries for enhanced fraud detection insights**.
•Improve collaboration for higher productivity.
•Test models for **fraud detection accuracy**.

Acknowledgements & Resources

Thank you to The Data Mine, Corporate Partner mentors, and faculty mentor for their support.
Project Guidance: EH Cyber Defense and Security Analytics Dept. & Mustafa Abdallah Program
Support: Maggie Betz, Bryce Castle

Resources:
[1] Crossing the Global Quality Chasm: Improving Health Care Worldwide: National Academies Press (US); 2018
[2] HIPAA Journal. Editorial: Lessons from 2024 Healthcare Data Breaches.