MLE Student Capstone Proposal

|  |  |
| --- | --- |
| Project Title | Fraud Anomaly Detection utilizing Healthcare Provider Data |
| Industry Sponsor | None; n/a |
| Team Members | Monika Sharma  David Lederer  Iain McKone |

## 

[**Supporting Org Information**](#_1uamq249imw4) **1**

[**AI Product/Capstone Project Description**](#_4ha20x76zzk) **2**

[Problem](#_6vdkmoqff0lo) 2

[Why](#_oomvm8feyy9c) 2

[Success](#_zta80kbm5v7) 3

[Audience](#_h6pw71d4a1o4) 3

[What](#_d1mkjxch3qo0) 4

[Final Deliverables](#_uf3nwannmyad) 4

[Anything Else?](#_qjw6h3i3ymk) 5

## 

## AI Product/Capstone Project Description

### Problem

[Write a succinct statement of the problem that you're trying to solve (<50 words)]

|  |
| --- |
| Healthcare fraud is a lucrative white-collar crime. The US National Health Care Anti-Fraud Association estimates that 3 percent of healthcare spending is lost to fraud (conservatively $300 billion/yr).  A digital fraud detection system is ideally suited to combat this threat to our fundamental right to accessible and affordable healthcare. |

### Why

Write about why this is a problem worth solving. What is the business value hypothesis that connects to what success looks like and for whom? (~50-250 words)

|  |
| --- |
| Fraud harms the broad healthcare community: providers, beneficiaries, suppliers, and patients. It ultimately leads to an increase in healthcare costs, insurance premiums and can pose serious harm to beneficiaries by reducing the accessibility of service, and the economic ability to provide quality service.  Healthcare fraud is also a continually evolving crime. Standard fraud detection requires data and skilled individuals with special domain knowledge to identify atypical patterns for financial healthcare transactions.  Existing detection techniques have been generally inadequate and are unable to keep pace with the increasing sophistication of this crime. The industry has not yet effectively utilized data science and to-date detection has been relatively slow, reactive, and manually intensive (time-and-cost).    This problem domain however is ideal for a skilled data scientist armed with an appropriate volume of high quality and representative data.  This capstone team will explore the practical detection of financial anomalies by utilizing data science and MLE techniques to ultimately reduce the amount of money lost to Healthcare fraud.  Specifically, this capstone will focus on the provider actor and the use of their data, associated scenarios and impacts. However we believe that these techniques have broad applications to other financial institutions impacted by fraud. |

### 

### Success

Write about what success looks like. What is the Key Performance Indicator (or couple of KPIs)? How might they connect to a relevant ML model accuracy metric? (<50 words)

|  |
| --- |
| ML is still in its infancy within Healthcare, however supervised/unsupervised learning techniques promises the following potential benefits:   * consistency, repeatability, and automation * improved cost-effectiveness   Key Performance metrics:   * # of fraud incidents (per 1000’s of txns). * Avg $ recovered (per 1000’s of txns) * Mean time for detection |

### Audience

Specify exactly which users/customers this AI/ML product is being built for. What is the customer's pain or need that connects back to the problem? (<50 words)

|  |
| --- |
| The audience for healthcare fraud detection includes:   * Finance office (eg CFO): cares about fiscal viability, value and overall business integrity including revenues, and payables * Compliance/Audit Office: cares about the integrity of overall business policy wrt employees, customers, and suppliers * Supply Chain Procurement: cares about the value chain |

### 

### What

Now describe what the ML looks like. This includes a discussion of data and sources, potential/likely models, a choice of an accuracy metric to optimize for and a defense of your choice. How does your accuracy metric connect back to the KPI(s) named above?

|  |
| --- |
| 1. Dataset: [Healthcare Providers Data for Anomaly Detection](https://www.kaggle.com/datasets/tamilsel/healthcare-providers-data) 2. Exploratory Data Analysis: This step involves exploring the data to find distributions, its main characteristics, identifying patterns and visualizations. It also provides tools for hypothesis generation by visualizing and understanding the data through graphical representation. 3. Feature Engineering: This is one of the most important step in an ML project.    1. Impute missing values. It is very important to never simply drop rows/columns with missing values in a dataset, but instead to replace them with mean/median or interpolate/extrapolate them.    2. Handle categorical data such as one-hot encoding technique.    3. Normalize the data for further model building. In this step we either adopt standard scaler or min-max scaler to create columns ready for model consumption. 4. Feature Selection: This is another important step in any ML project where we identify what features are important for the outcome we are looking for. There are various techniques available such as heat map using visualization or Lasso regression or Feature importance rendered by tree based models. 5. Final step is to identify different models that will help us to solve the business problem at hand. For anomaly detection we have a handful of identified models such as **isolation trees, k-means clustering, auto-encoders, logistic regression** etc. 6. Logistic regression technique allow us to calculate metrics such as sensitivity and specificity, which in-term tie back together with the KPI of number of fraud cases detected. 7. Once fraud cases are accurately detected, amount can be calculated which ties with the KPI of recover amount per fraud case detected. 8. Dollar amount recovered per case can be bucketized in high, medium and low categories depending on the amount, which ties back with fraud case priority. |

You may also find it helpful to fill out an [MLOps Stack Canvas](https://ml-ops.org/content/mlops-stack-canvas) or [MLOps Stack Template](https://ml-ops.org/content/state-of-mlops). These tools really help to clarify tech stack requirements for our students.

### Final Deliverables

At the conclusion of a cohort, students are expected to deliver:

* Deployed Demo
* 10-minute Presentation
* GitHub Repo ( description in README + code )

For more detailed information on student capstone projects, you can check out the guide that we provide MLE students [here](https://docs.google.com/document/d/1lFRKgc9darivZaNaGoGni9Gch3hIuwnkN8AVql-kreU/edit?usp=sharing).

### 

### Anything Else?

Please provide any additional information on key activities, technologies, datasets, expected learning outcomes, potential mentorship or employment opportunities, or anything else not listed above!. And thanks for supporting our students!

|  |
| --- |
| I anticipate some challenges in mapping detected anomalies to a representative fraud scenario.  Given the growing sophistication of fraud, with multiple players, I expect that there will be the potential of a complex event model, whose definition, analysis, and reproduction will fall well outside the scope and timeline of this cohort. |

**About FourthBrain**

FourthBrain trains aspiring Machine Learning engineers in the technical and practical skills necessary to contribute immediately to an AI team. Our remote, online program is designed to be flexible and accessible for anyone with software experience. We infuse values of collaboration, communication, empathy, and equity throughout the program.

We are part of the AI Fund, founded by Andrew Ng.