MLE Student Capstone Proposal

|  |  |
| --- | --- |
| Project Title | Anomaly detection using Healthcare Providers Data |
| Industry Sponsor | None; n/a |

## 

[**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)]

|  |
| --- |
| Today, data drives most business decisions, so it has become all the more important to keep a check on patterns in data that don't conform to normal behavior. A technique known as anomaly detection, helps you keep a pulse of your business, detect fraud and other rare events that may have significant consequences on your business. |

### 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)

|  |
| --- |
| Healthcare fraud is an expensive white-collar crime in the US. Such frauds lead to an increase in healthcare premiums or pose serious harm to beneficiaries. There is an increased need of digital healthcare fraud detection system that can help combat this threat to fundamental right of any human. The National Health Care Anti-Fraud Association[1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9013219/#B1) conservatively estimates that about 3 percent of our healthcare spending is lost to fraud ($300 billion approximately) yearly. Healthcare fraud is defined as an individual, group of people or a company knowingly misstating or misinterpreting about scope or nature of medical services provided, which in turn results in unauthorized payments. Fraud management can be divided into two groups; fraud detection and fraud prevention. Fraud detection is the first step where it is identified as quickly as a fraudulent scheme is perpetrated. Some of the main actors involved in this crime are providers (who are licensed to provide services to beneficiaries), medical equipment manufacturers, or drug manufacturers. Fraud prevention in healthcare can be defined as any action or policy put in place to prevent future system abuse. Combating fraud allows payers to maintain positive revenues, protect consumers from overbilling and increased premiums, and ensure that financial assets are secure from criminal activity. |

### 

### 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)

|  |
| --- |
| Key Performance metrics for detecting fraud in healthcare include:   * Number of new fraud cases * Recovery amount per fraud case detected * Fraud case by scheme type (false billing, upcoding, etc.) * Fraud case priority (high, medium, or low) based on financial risks/implications |

### 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 fraud detection in healthcare AI/ML product is a Finance leader, Revenue Integrity leader, Cost Center Leader, Project Sponsors, Chief Financial Officer, and Finance Planning and Analysis Leader. |

### 

### 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. Dropping rows/columns with missing values leads to loss of information. Instead it is important to impute these values, and replace them with mean/median or interpolate/extrapolate them.    2. Handling categorical data such as one-hot encoding technique.    3. Normalizing 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 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 second 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!

|  |
| --- |
|  |

**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.