# 

# Student Project Proposal

|  |  |
| --- | --- |
| Project Title | Healthcare Provider - Fraud Anomaly Detection |
| Industry Sponsorship (if Any) | None, n/a |

## Project Description

**Problem definition**

*[50-100 word description of the problem which you will solve]*

|  |
| --- |
| 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.  Fraud detection requires data and skilled individuals with special domain knowledge to detect atypical patterns.  Existing 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. |

**Key Research Questions/ Technological constraints that the Project will Answer**

|  |
| --- |
| Research Questions:   1. Can ML detect financial anomalies in Provider data? 2. Can anomalies be categorized as fraud? 3. What types of fraud is the model best suited for?   Technological Constraints:   1. Sourcing representative, and quality data, preferably with labels 2. The potential for False Positives 3. Feature Generation could potentially be time consuming |

**Final deliverables at the end of the project**

*[Please list the desired technical deliverables from the project team in as much detail as possible]*

|  |
| --- |
| 1. Github repo where the code and project artifacts will reside 2. Tested supervised/unsupervised model, along with jupyter notebook for reproducibility 3. Admin tools/utilities, and FastApi endpoint(s) for model monitoring and re-training (pipelines) 4. User tools/utilities for user presentation / review of findings (e.g. PowerBI - monthly / quarterly reports) |

**Key activities/ technologies the project team may be expected to undertake/ work with**

*[E.g. What kind of technology stack will you work with, the datasets you may need to work on, what kind of analysis you may be expected to undertake, etc.]*

|  |
| --- |
| Key Activities   1. 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. 2. 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. 3. 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. 4. 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. 5. 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.   Key Technologies: Technology Stack   * Notebook: Google Colab, Jupyter * Infrastructure (OS): Local, Cloud Infrastructure * Data Eng: DVC, Spark, DeltaLake, DataBricks * Modelling: Python, Data Science Libs: PyTorch, SciKit Learn * CI/CD Pipeline/Staging: AirFlow |

**Expected learning outcomes**

*[What do you expect to learn from the project? Please mention the technical skills you will imbibe over the project.]*

|  |
| --- |
| 1. MLE Methodology / Lifecycle 2. MLE Data Engineering Toolset Stack 3. Feature Selection / Engineering, Unsupervised Learning Methods 4. Solution Architecture 5. Python Application dev, exposing FastAPI endpoint development 6. MLOps, CI/CD Pipeline development |

|  |  |
| --- | --- |
| Team Size: | 3 |
| Member names: | Monika Sharma  David Lederer  Iain McKone |

### 

## Tentative Time plan

Submit a tentative time plan (table/chart or text) regarding breakdown of the work that will be conducted between in the second half of your cohort, from week 6 onward.

|  |  |
| --- | --- |
| Week# | Key Tasks |
| 5 | Data Scoping |
| 6 | Solution Architecture, System Design |
| 7 | Data Labelling |
| 8 | Feature Engineering |
| 9 | Feature Selection |
| 10 | Model Training, Validation |
| 11 | Model Validation, Testing |
| 12 | MLE Iteration, Optimization |
| 13 | MLE Iteration, Optimization |
| 14 | Demo Workflow; UI Development |
| 15 | Dry Run; Demo Finalization |
| 16 | Presentation, Demonstration |

## System Design

From the System design perspective, outline the following:

* Data: [Healthcare Providers Data for Anomaly Detection](https://www.kaggle.com/datasets/tamilsel/healthcare-providers-data)
* Process (Models, iterations)
* Outcome (output and recommendations)



Source Image: https://www.databricks.com/solutions/machine-learning

What are the system design considerations for your deployable ML model? Describe the iterations, delivery formats and limitations you may face and some solutions to overcome the limitations

* Should the model be deployed to run in batch, or to be hit from an api or some sort of streaming process as events are generated?
  + The model will be deployed to run in batch. It is envisioned to align with a financial organization’s regular monthly and/or quarterly reporting cycle. Api’s will be provided so that the model can be run on demand.
* What sort of infrastructure will be required for training? If it is a model that requires a lot of resources, where is the best place to train?
  + We do not envision a need for significant resources. The solution could be containerized to run locally or in the cloud. The team will also explore the potential application of the DataBricks suite of technologies.

## Ethical Considerations

Are there any ethical considerations of your project? Consider the data source, the intended outcome, and/or the eventual use cases.

* Did you modify anything about your plan based on these considerations?
* Can you anticipate any issues that might arise during the process?

Bias: The data is specific to the US, US Providers, Medicaid, and US Socio-Geo-Cultural norms. US Lifestyle choices and conditions which may be strong factors in health conditions cannot be directly applied to other countries.

Transparency: Data traceability, clear communication of results, intelligibility and reproducibility of results.

Accountability: there could be legal and financial ramifications to publishing these findings. The data and its findings can be construed as sensitive and should be handled with discretion (privacy concerns).

Sources of potential harm:

* Innocent until proven guilty
* %confidence versus absolutes
* Reputation, unfair allocation, unfair representation

Other considerations: Accuracy, Fairness, Privacy, Human Control and Decision Making.