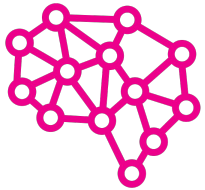


FourthBrain

MLE 10 Capstone:
Healthcare Claim Anomaly Detection

Iain McKone, Monika Sharma

Outline

- Problem Domain
- Our Approach
- Data, Model(s)
- Demo
- MLE Stack
- Conclusions
- Future Considerations
- Q & A

Problem Domain:

- Healthcare fraud is a serious white-collar crime in US
- Est. $70B (3% of total spend) is attributed to fraud; max $300B
- An increasingly sophisticated crime (e.g. collusion)

https://www.nhcaa.org/tools-insights/about-health-care-fraud/the-challenge-of-health-care-fraud/
https://www.bcbsm.com/health-care-fraud/fraud-statistics.html

Examples

- **($300M) 2022 April - 11 defendants** charged with kickback schemes involving collusion between medical practitioners, laboratories, and a marketing firm.

  https://www.justice.gov/usao-ndtx/pr/11-defendants-plead-guilty-300-million-healthcare-fraud

- **($143M) 2021 May - 14 defendants** charged with multiple **covid-related fraud schemes** i.e. collusion between a medical doctor, laboratories, pharmacies, and a home health agency.

  https://www.cnbc.com/2021/05/26/doj-charges-14-people-in-alleged-health-care-fraud-related-to-covid-19.html

Implications:

- **Overwhelms the system** with unnecessary tests and procedures
- **Increases the cost of care**
  - **Increases insurance** premiums ($)
  - **Reduces accessibility** to care ($)
  - **Reduces availability** of care (scheduling)
  - **Reduces value, quality** of care

Current State: (envisioned as-is; do-nothing)

	yoy Growth
Claim Submissions	↑
Anomalies	↑
Reqd Effort: Detection	↑
Available Effort: Analyst Person Hours	↑

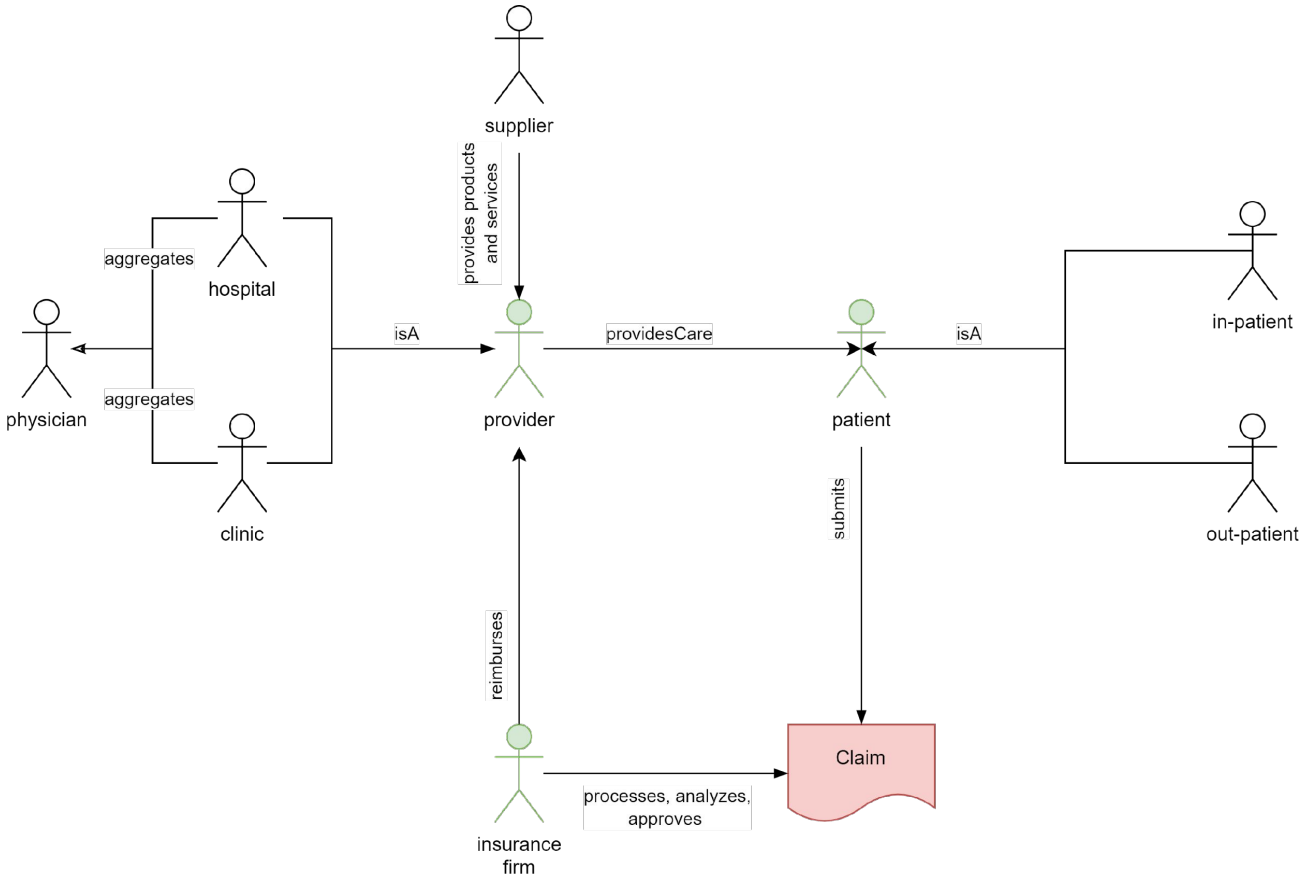
Future State: (envisioned to-be)

- Q: For Claims Anomalies, can Data Science:
  - **Automate; reduce** the overall **time** and **cost** required for anomaly detection?
  - **Accurately** minimize the number of invalid, erroneous claims and reimbursements?
  - **Continuously evolve** in response to shifting data and behavior patterns?

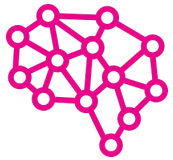
Approach: Data

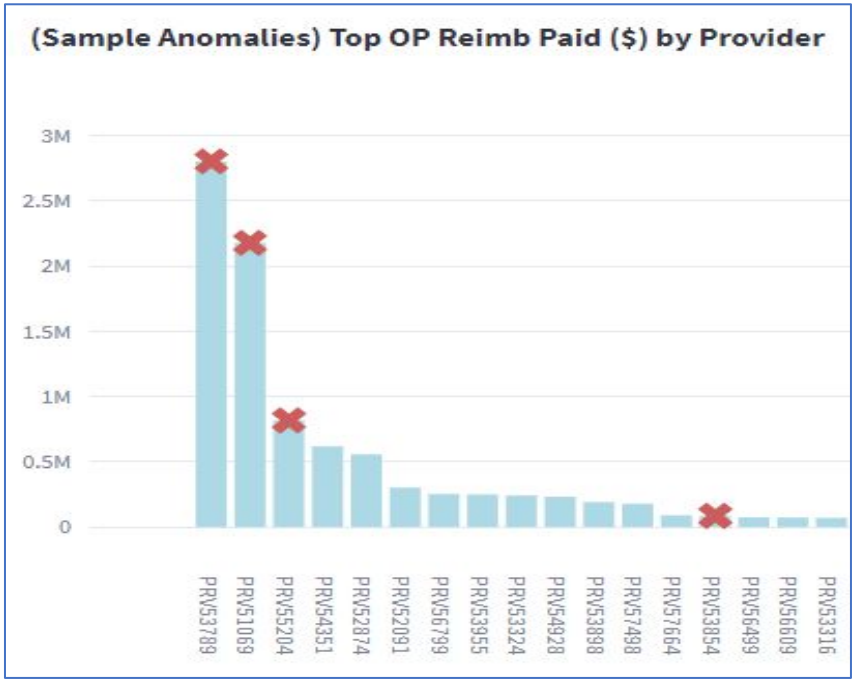
- Kaggle - https://www.kaggle.com/code/rohitrox/medical-provider-fraud-detection/data
  - **Anonymized:** Geography, Race, Provider
  - **Pre-split** Train and Test Claims data (~558k vs 135k rows)
  - **Provider Labels**: 506 Yes, 4904 No

  - Beneficiary Claims, Reimbursements
  - In/Out Patient Claims, Procedures, Admit Duration



Approach 1: Providers - Supervised – Fraud Labels





Kaggle

Provider Fraud

Claims

Training

Validation

Test

Auto Encoders

LogR

SVM

XGB

STAGE 1: Data Pipeline

STAGE 2: Feature Eng

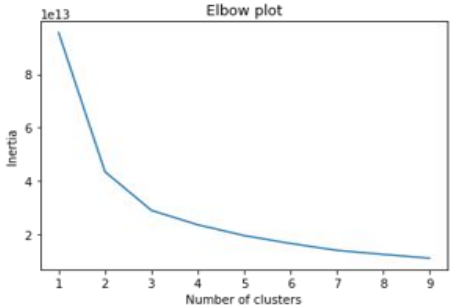
STAGE 3: Model Comparison

STAGE 4: Model Publishing

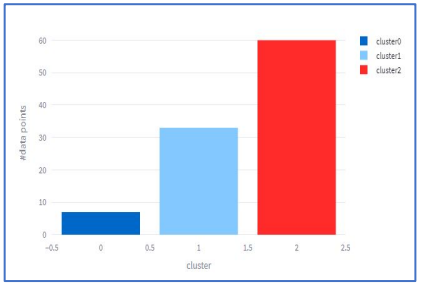
Approach 2: Claims - Unsupervised – No Fraud Labels

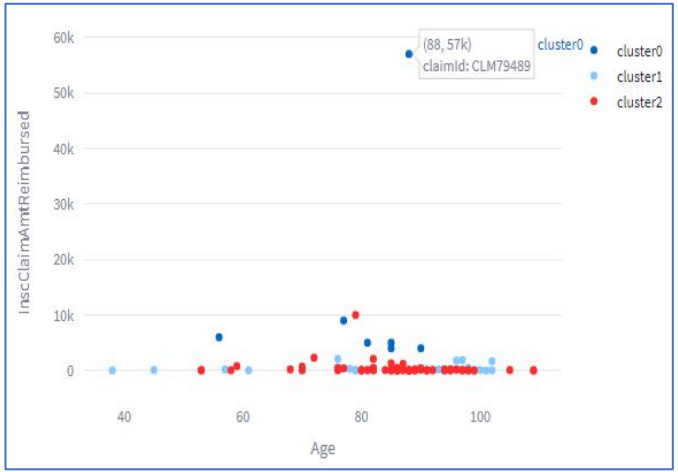
Kaggle Dataset

Claims



Random Forest Classifier





STAGE 1:
Data Pipeline

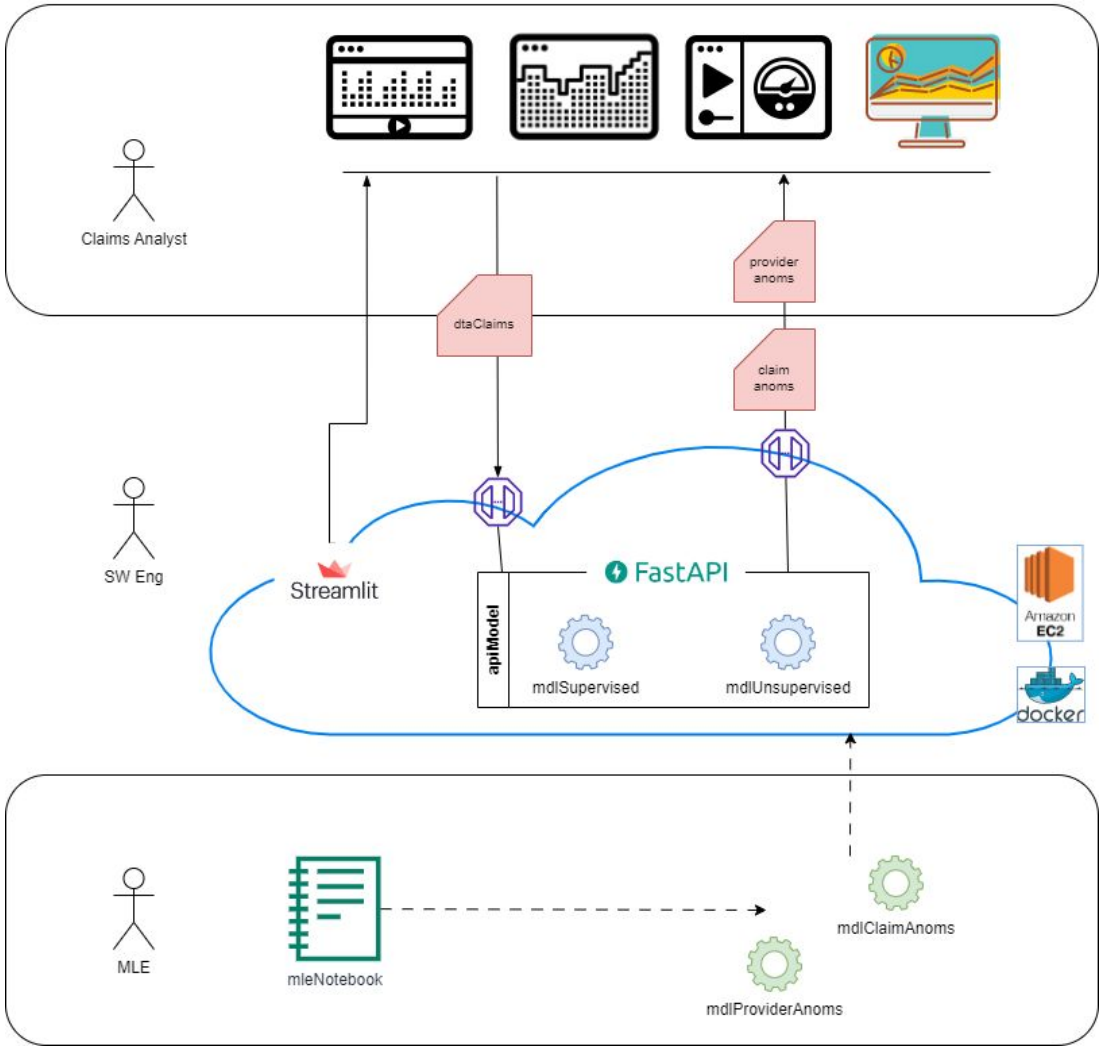
STAGE 2:
Feature Eng, Inertia Analysis

STAGE 3:
Cluster Analysis, Explainability

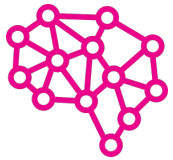
STAGE 4:
Model Publishing

Demo

MLE Stack


Claims Analyst
dtaClaims
provider anoms
claim anoms
SW Eng
Streamlit
FastAPI
apiModel
mdlSupervised
mdlUnsupervised
Amazon EC2
docker
MLE
mleNotebook
mdlClaimAnoms
mdlProviderAnoms

Approach 1 Findings: Supervised Provider Anomalies



	LogR	XG Boost (TPot)	SVM	Auto Encoders
F1 Score	0.591	0.586	0.491	0.802
Accuracy	0.909	0.920	0.930	0.868
Recall	0.697	0.605	0.362	0.752
Precision	0.512	0.568	0.764	0.858

Approach 1 Findings: Supervised Provider Anomalies

Explainability: Insurance Claim Reimbursements ($) were flagged as the most influential feature

Known Limitations:

- Business Value: Predicting Anomalies at the Provider level was not compelling enough
- Models: **Autoencoder Performance** was very promising, however the **data is misleading**
- Data
  - Provider True (Fraud) labels are at best **incomplete**
  - Provider False (No-Fraud) labels are at best **'indeterminate'**
  - Mapping Provider Fraud Labels to all child Claims is a weak extrapolation

Approach 2 Findings: Unsupervised Claim Labels

- Visual representation of anomalous claims is very compelling and user friendly

Known Limitations:

- Business Value:
  - **Cluster Explainability:** What does this tell us about coarse anomaly profiles?
  - **Supplementary Details:** Provide more details re contributing factors to a claim anomaly

  - Cluster0 (90%): AdmittedDays, DeductibleAmtPaid, InscClaimAmtReimbursed
  - Cluster1 (40%): ChronicCond - KidneyDisease, Heartfailure, ObstrPulmonary
  - Cluster2 (40%): ChronicCond - KidneyDisease, Heartfailure, ObstrPulmonary

Conclusions

- Business Value
  - **Automation; reduce** the overall **time** and **cost** required for anomaly detection
  - **Accurately** minimize the number of invalid, erroneous claims and reimbursements
  - **Continuously evolve** in response to shifting data and behavior patterns
  - **Provide Deeper Insights** – Explainability

Future Work

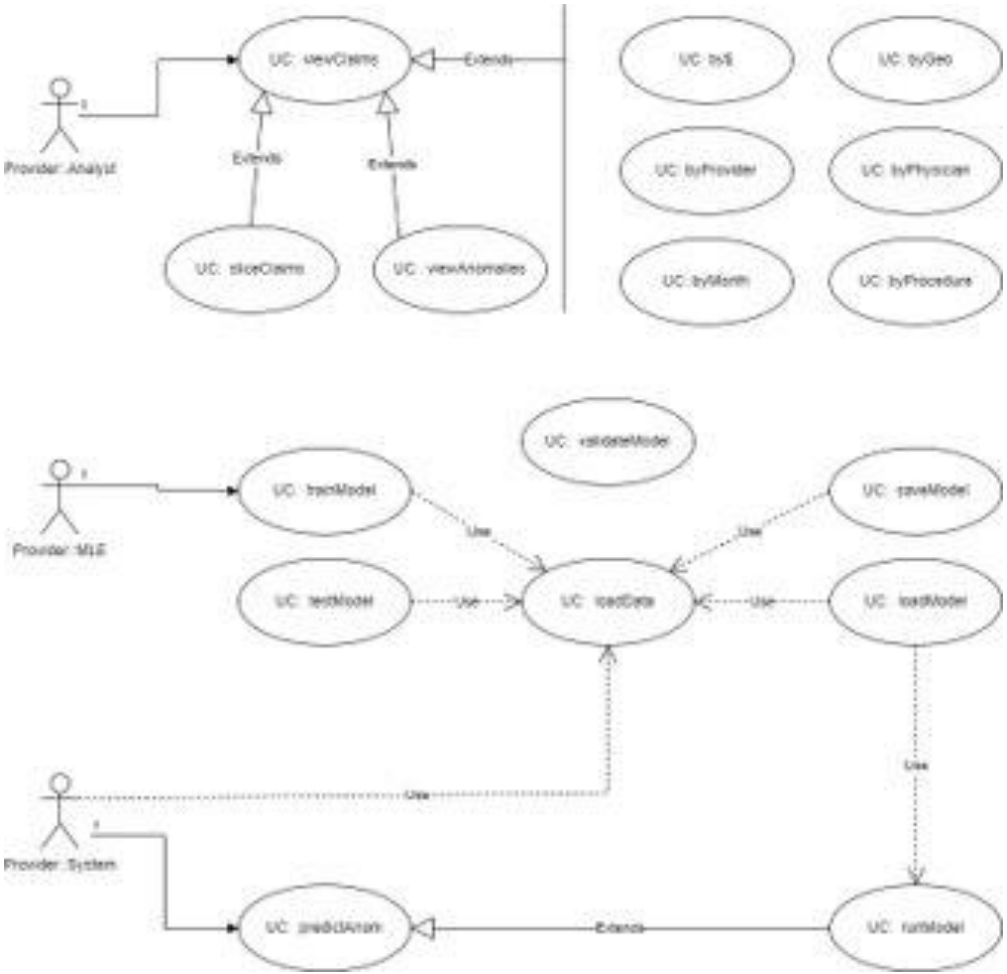
- Business Value:
  - Introduce a **human feedback loop** for ongoing model improvements;
  - Provide a fuzzy logic indicator of **degree of fraud risk** for a claim
  - Provide a **visual breakdown** of the attributes of fraud risk for a claim

- Data:
  - Acquire **claim-level labels** to improve supervised training
  - Acquire **labels for confirmed non-fraud**, as well as confirmed fraud to improve Auto Encoders

- Model:
  - Explore Variational Autoencoders, which have shown promise for credit card fraud

Questions? Feedback?

Appendices


UC: viewClaims
Extends
Provider Analyst
Extends
Extends
UC: byS
UC: byGeo
UC: byProvider
UC: byPhysician
UC: byMonth
UC: byProcedure
UC: storeClaims
UC: viewAnomalies
UC: validateModel
Provider MLE
UC: trainModel
UC: saveModel
Use
Use
UC: testModel
Use
UC: loadData
Use
UC: loadModel
Use
Use
Provider System
UC: predictAnom
Extends
UC: runModel

EDA: Training Data (csv)

Kaggle: https://www.kaggle.com/code/rohitrox/medical-provider-fraud-detection/data

Data divided into four sections:

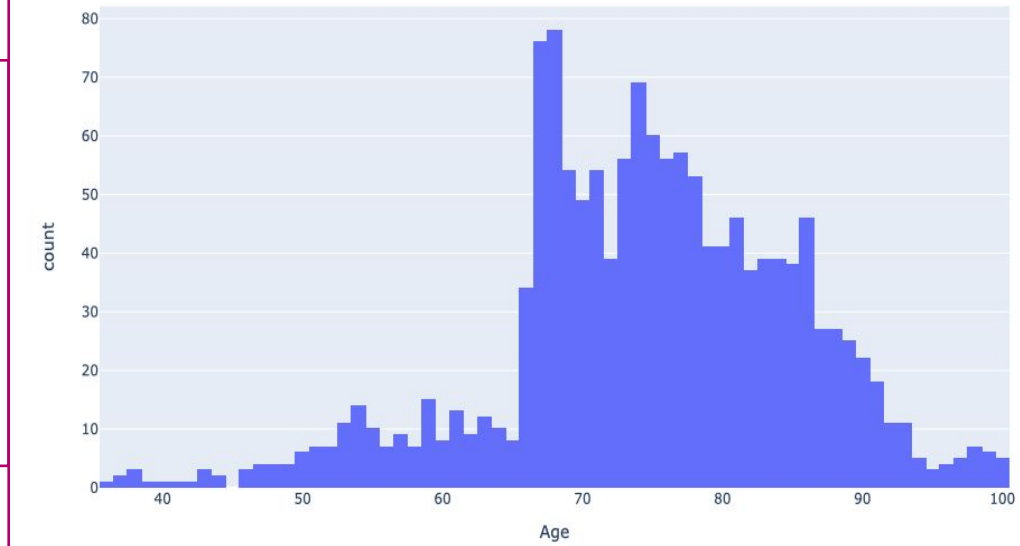
- Provider: labeled as fraud or not
- Beneficiary information
- In-patient information
- Out-patient information

Data are anonymized; provider name, Beneficiary name, state, insurance provider name are masked

- 5410 data points labelled Potential Fraud (Yes | No)
- 506 Yes, 4904 No
- Imbalanced dataset (~10% target); class imbalance needs to be managed

EDA: Training Data Feature: Age

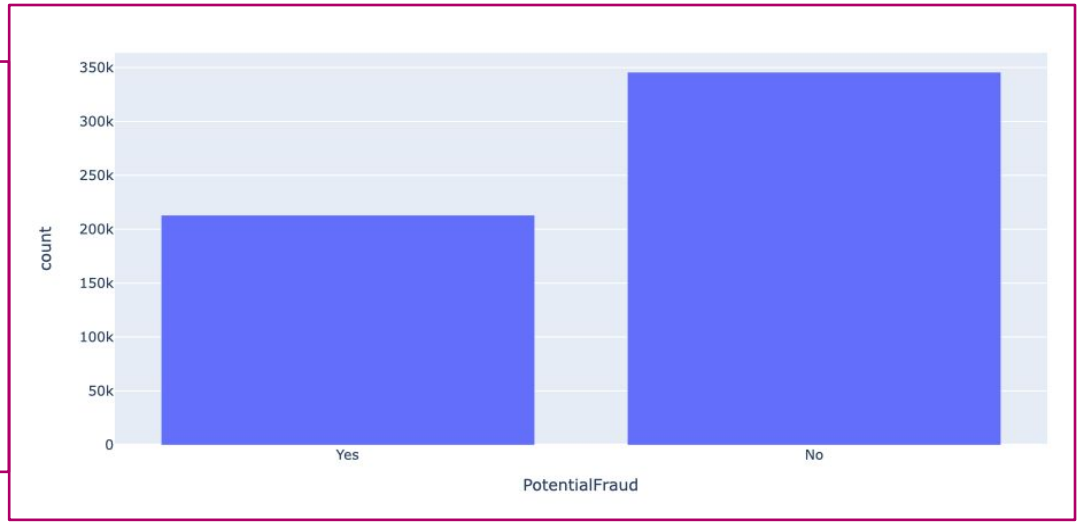
Feature: Age



- Dataset biased towards age > 65 years
- Is it easier indicative of prevalent fraud in one payer sector?

EDA: Fraud Label Distribution

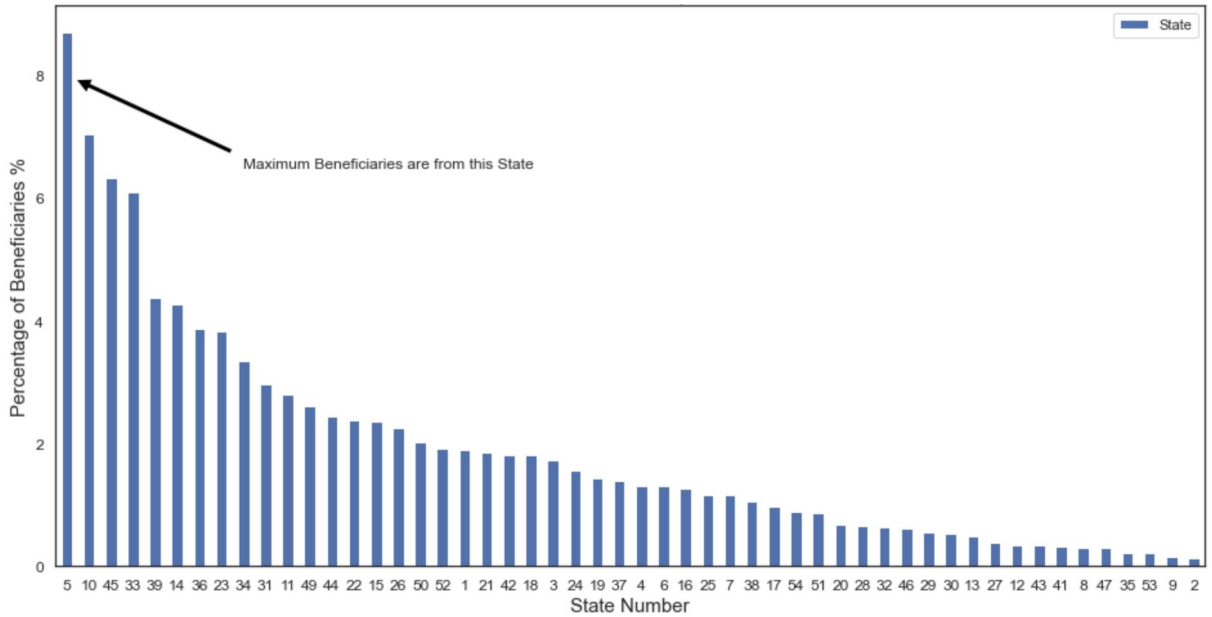
Fraud Label Distribution



- Individual providers dataset when merged with patient dataset renders 60:40 ratio for no fraud/fraud
- Key-take away: Providers who commit fraud victimize many patients (Note: misleading inference due to provider:claim mapping)

EDA: Beneficiary State-wise Distribution

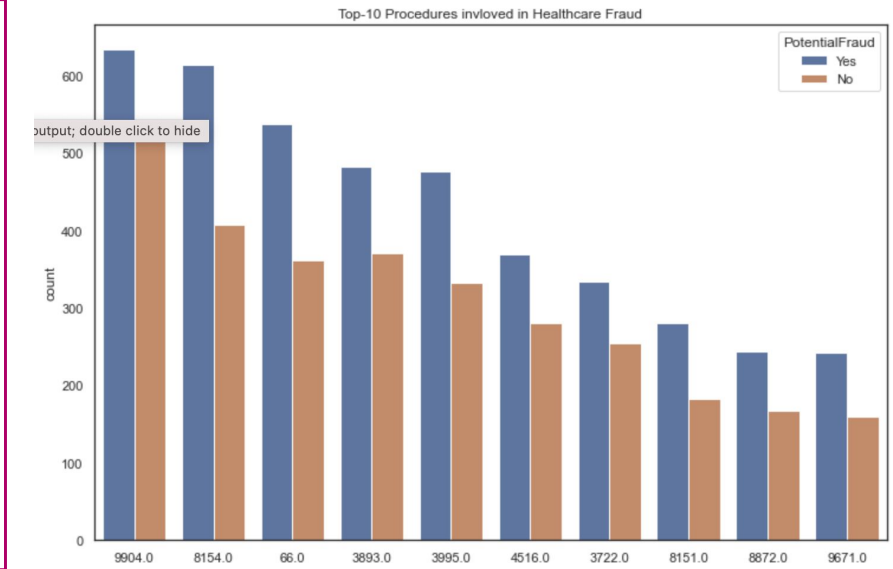
Beneficiary state-wise distribution



- Key-take away: Most beneficiaries are from states 5, 10, 45, and 33

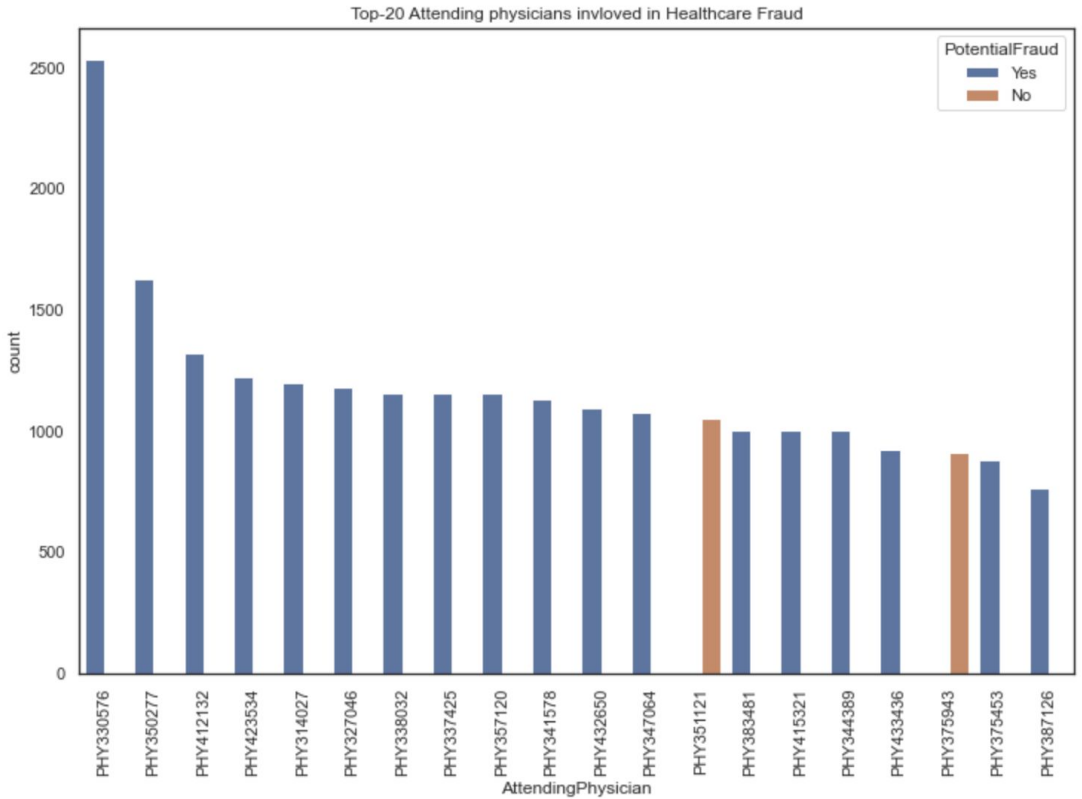
EDA: Procedures Where Fraud is Prevalent

Fraud distribution by Procedure



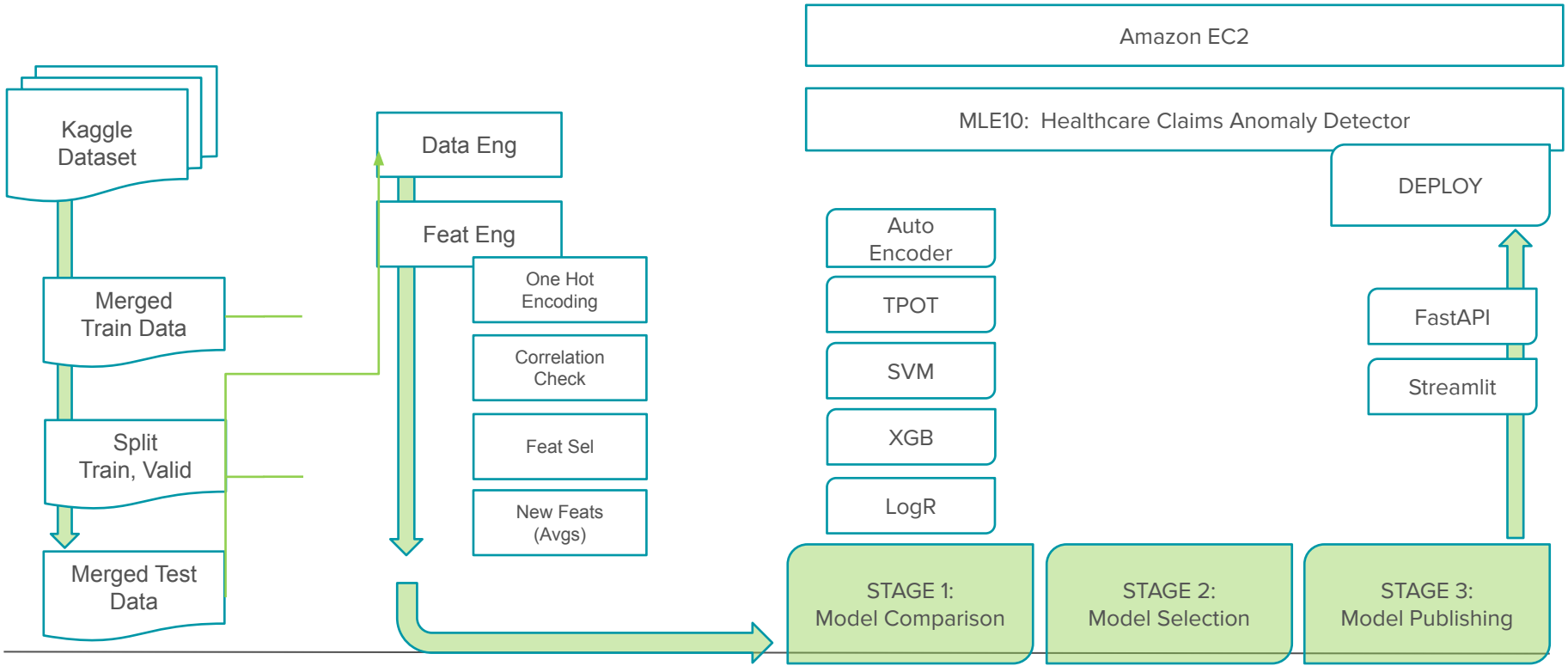
- Key-take away: Procedure IDs 9904, 8154, 66, and 3893 have largest number of fraud cases
- Question: Why is it easier to commit fraud for these procedures

EDA: Out-Patient Train Data Summary

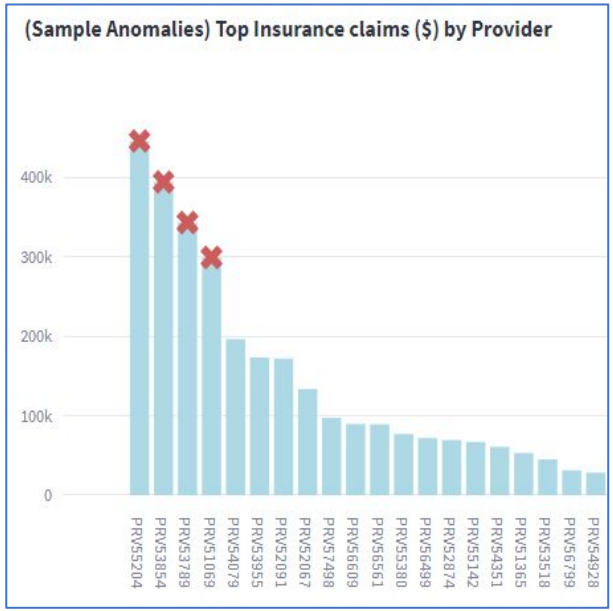


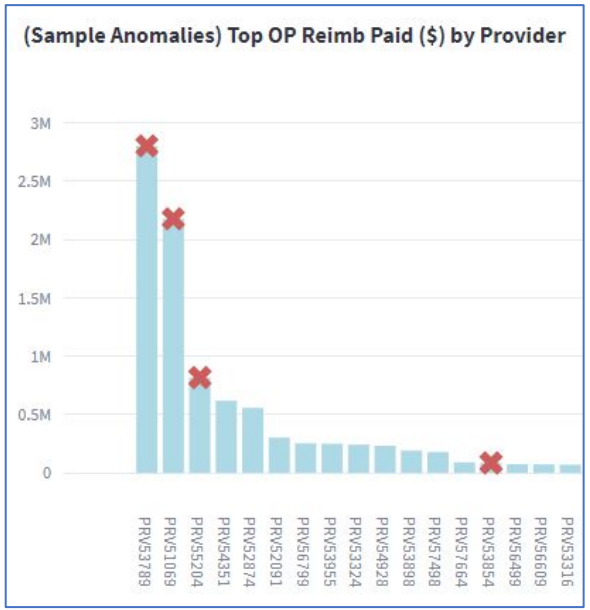
- Key-take away: Providers who have a habit of committing fraud always commit fraud with every patient!

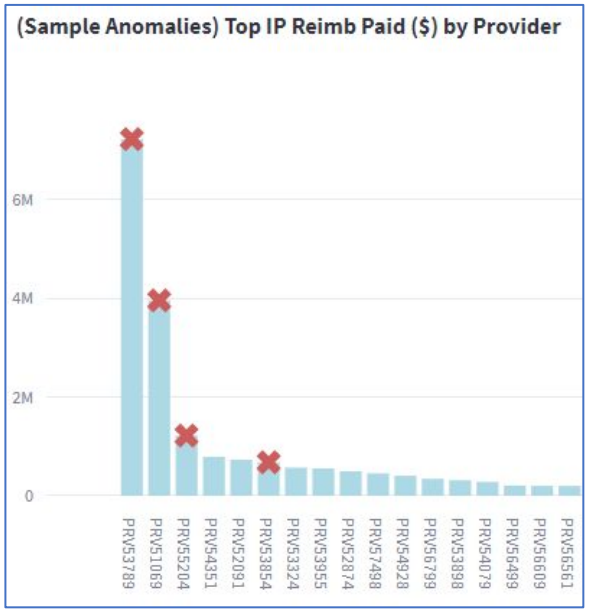
Approach: Supervised Model – Labelled Providers



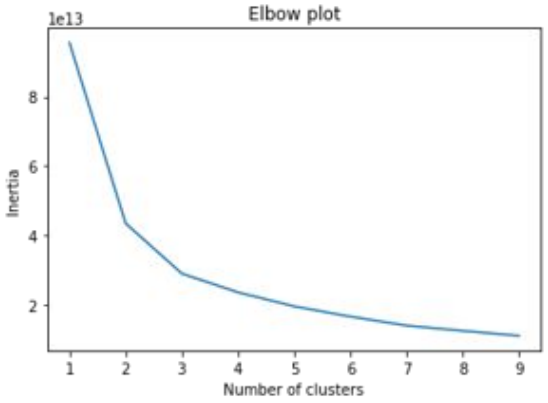
Findings: Supervised Model – XG Boost Provider Predictions

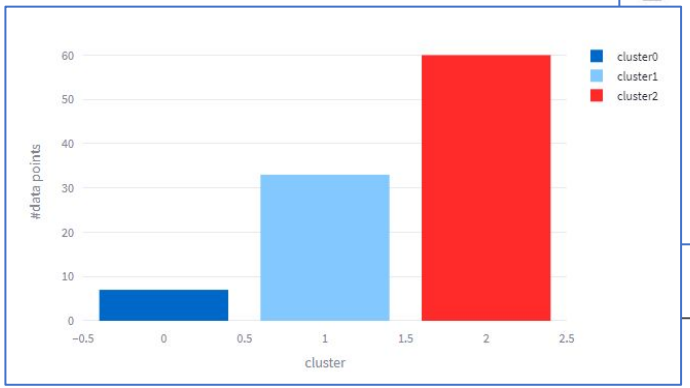


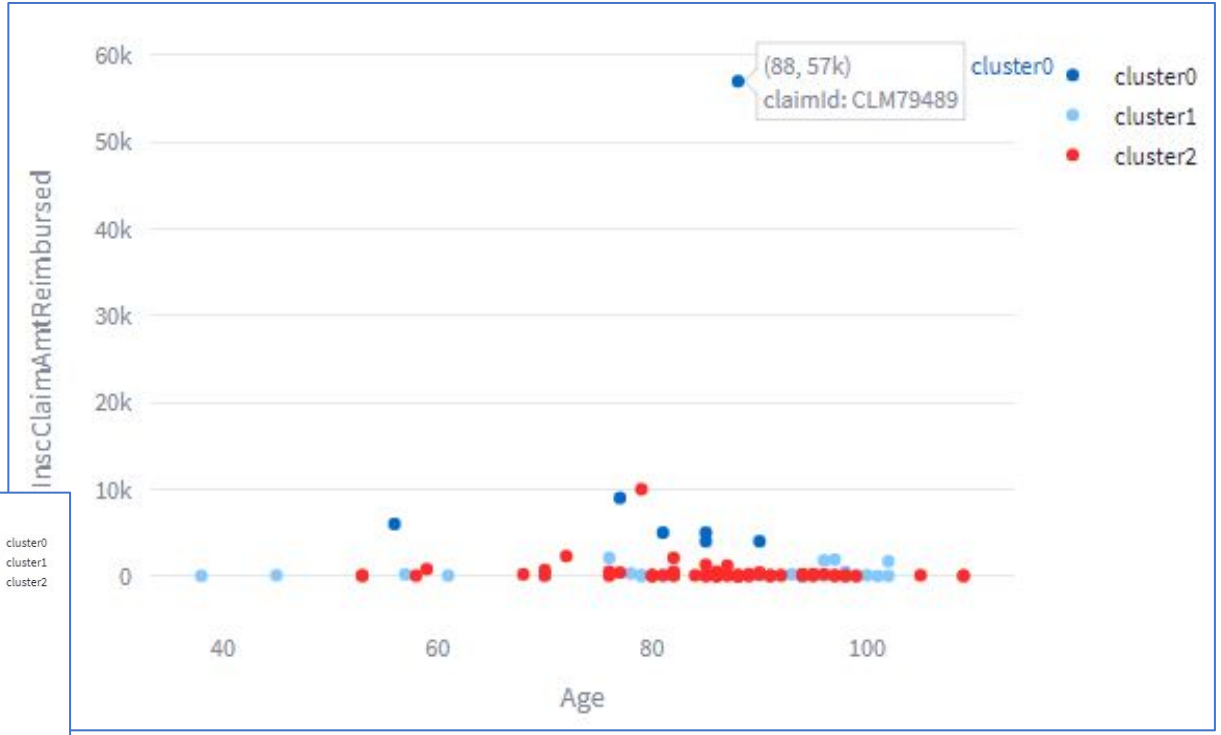




Findings: Unsupervised Model – KMeans Claim Labels







Current State: (envisioned as-is; do-nothing)

- Increasing overall annual # and $ claims
- Increasing associated annual #, %, and $ of anomalies and fraud cases
- Increasing avg $ cost per detection; i.e. evolving fraud sophistication

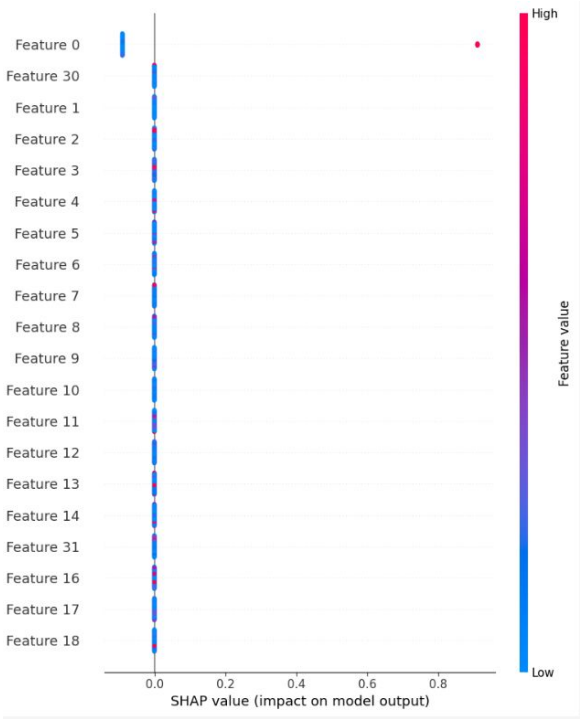
- (Relatively) Flat or decreasing % analyst manhours per claim
- (Relatively) Flat or increasing data, and analysis tools (receptive)

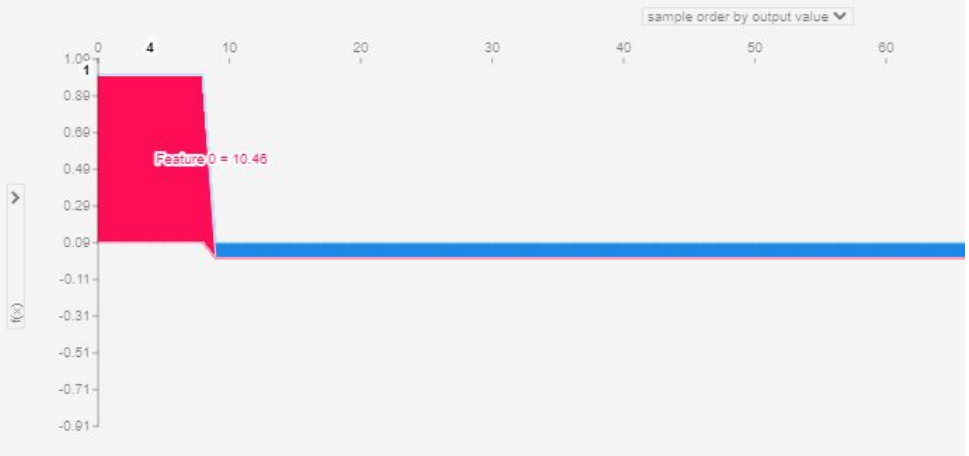
Future State: (envisioned to-be)

- Q: For Claims Anomalies, can Data Science:
  - **Recommend** coarse guidelines for claim anomaly filters and rule sets?
  - **Provide deeper insights**, i.e. supplementary details and key contributing attributes (features)?
  - **Weigh** the % likelihood of anomaly and/or fraud to assist with case prioritization?

Approach: Supervised Model – Labelled Providers

- Data: Kaggle Dataset
- Exploratory Data Analysis, Feature Engineering
- Model Training: Supervised – Provider predictions, labels
  - Logistic Regression, Support Vector Machines, XG Boost, Auto-encoders, Auto-ML (TPOT)
- Model Selection: XG Boost





Findings: Claims EDA

- Training data
  - Age: skewed towards patients >65 yrs
  - Geo: Most beneficiaries are from states 5, 10, 33, and 45
  - Procedures: ids 66, 3893, 8154, and 9904 have the highest cases of fraud
  - Out patients: some indications of repeat offense by provider
  - Insurance claim fraud – distribution of $ reimbursed
    - 8% by provider
    - 7% by attending physician
    - 6% by operating physician
    - 4% by claim diagnosis

Future Work

- Streamlit:
  - Provide capabilities to dig into claim to explore the contributing features of the anomaly
  - Upgrade to a HTML5 front-end
  - Separate hosting for front-end from back-end
- FastAPI:
  - Expand api to include claims data verification, and json data request/response
  - Expand api to include model updates, retraining, and published model performance
- Infrastructure:
  - Promote local Docker environments to hosted in Amazon EC2, or Kubernetes
- MLOps:
  - Upgrade from Level0 to Level1; introduce aspects of retraining, monitoring
  - Host the model seperately in a service such as SageMaker