Final Report (Integrity M)

SeungHeon Han, Yihang Zhao, Yuwen Luo

1. Executive Summary 2
Goals 2
Success Criteria 2
Project Plan 2
2. Problem Understanding 2
Background 2
Objective 2
Business Success Criteria 2
3. Methodology 2
Data Analyzed 2
Analytics Techniques Used 5
4. Results, Conclusion, and/or Recommendations 5
Logistic Regression 5
Random Forest 6
Decision Tree 7
Explainable Boosting Machine(EBM) 8
eXtreme Gradient Boosting（XGBoost） 9
Evaluation on Test Set 11
5. Potential Next Steps for future 12
6. Appendices including code developed for the project 12

1. Executive Summary

1) Goals

Develop a predictive model which can predict fraud, waste, and abuse with a high accuracy rate

2) Success Criteria

- The predictive model can be replicated in other places without losing much accuracy
- The model can predict fraudulent patterns with high F1, AUC, ROC scores

3) Project Plan

- Understand the business problem and goal with clients
- Collect datasets, review documents
- Wrangle with datasets and clean datasets
- Label fraud patterns within the data frame
- Data balancing (undersampling)
- Data partitioning (train, test)
- Data partitioning (train, validation)
- Apply machine learning algorithm to the data
- Measure model performance on a test dataset
- Visualize the results

2. Problem Understanding

1) Background

Emerging technologies like AI and ML can help agencies like CMS to mitigate fraud, waste, and abuse, expedite the claims process, reduce errors, and lower the costs of Medicare claims management

2) Objective

Develop a predictive model prototype to identify fraudulent patterns in healthcare data

3) Business Success Criteria:

The predictive model can be implemented to reduce fraud, waste, and abuse which can lead to reduced patient harm and financial costs

3. Methodology

1) Data Analyzed

- **Data sets**

- PartB [Main Dataset]
  - Medicare Physician & Other Practitioners - by Provider and Service
  - 2013 - 2019
  - Source: CMS website
  - Size: 67,764,122 * 29

- LEIE Databases [Labeling]
  - 08-2021 Updated LEIE Database
    - Source: Office of Inspector General Website
    - Size: 74,584 * 18
  - 2020-2021 LEIE with Reinstatements
    - Source: Office of Inspector General Website
    - Size: 484*21
  - LEIE Plus
    - Source: IntegrityM
    - Size: 937*18

- **Used Features (PartB)**

Features Used	**Data Dictionary**
Rndrng_NP	National Provider Identifier
Rndrng_Prvdr_Type	Type of the Provider
HCPCS_Cd	Healthcare Common Procedure Coding System (HCPCS) cod
Place_Of_Srvc	either a facility (F) or non-facility (O)
Tot_Benes	Number of Medicare Part B fee-for-service beneficiaries utilizing the drug
Tot_Srvc	Number of services provided
Tot_Bene_Day_Srvcs	Number of Distinct Medicare Beneficiary/Per Day Services
Avg_Sbmtd_Chrg	Average Submitted Charge Amount
Avg_Mdcr_Pymt_Amt	Average Medicare Payment Amount
Avg_Mdcr_Alowd_Amt	Average Medicare Allowed Amount
Avg_Mdcr_Stdzd_Amt	Average Medicare Standardized Payment Amount

- **Preprocessing**

- **Data Exploration**
    - No null/invalid values across every used feature
    - Definition of Outlier: |z-score| > 3
    - **Multicollinearity:**
        - Tot_Bene_Day_Srvcs & Tot_Benes (r = 0.97)
        - Avg_Mdcr_Alowd_Amt & Avg_Mdcr_Stdzd_Amt - Average (r = 0.99)
        - Avg_Mdcr_Pymt_Amt & Avg_Mdcr_Stdzd_Amt (r = 0.99)
        - Avg_Mdcr_Pymt_Amt & Avg_Mdcr_Alowd_Amt (r = 1)

- **Data Cleaning & Labeling**
    - **Filtering out the rows that contain HCPCS code related to prescription**
        - Drug Average Sales Pricing File (DASP)
        - Removing the instances that have HCPCS_cd matching to the DASP file
        - 2,074,502 rows were removed from PartB

    - **Sorting out the NPIs in the LEIE files matching to the fraud-related exclusion types referencing the OIG Acts**
        - **Total number of LEIE data:** 76,005
        - **Number of invalid unique NPI:** 69,465
        - **Number of valid unique NPI:** 6,540
        - **Number of valid unique NPI matching to frauds:** 5,489

    - **Labeling each instance of PartB (Fraud = 1/ Non-fraud = 0)**
        - **Fraud:** 33,638
        - **Non-fraud:** 65,655,982

    - **Undersampling the majority group**
        - Decrease the number of examples in the Non-fraud group to 10 times as many examples as the minority group
        - Undersampled PartB:
            - **Fraud:** 33,638 (9.09%)
            - **Non-fraud:** 336,380 (90.91%)
    - **One-Hot Encoding**
        - Removed HCPCS_cd variable (due to its more than 3000 dummies)
        - Dummy variables for
            - *Rndrng_Prvdr_Type*

        - *Place_Of_Srvc*
    - Number of features
        - 11 ⇒ 132

- **Data Splitting**
    - **Train:** 80% / **Test:** 20%
    - **Stratified random sampling**
        - Fraud data were evenly separated into train and test sets

2) Analytics Techniques Used

- **Validation Split:**
    - Splitting the whole dataset into train and test sets with 7:3 ratio

- **Modeling Technique**
    - Logistic Regression
    - Random Forest
    - Decision Tree
    - Explainable Boosting Machine (EBM)
    - eXtreme Gradient Boosting (XGBoost)

- **Modeling assumptions for a logistic regression model**
    - The random errors have a constant standard deviation
    - The random errors follow a normal distribution
    - The data are randomly sampled from the process

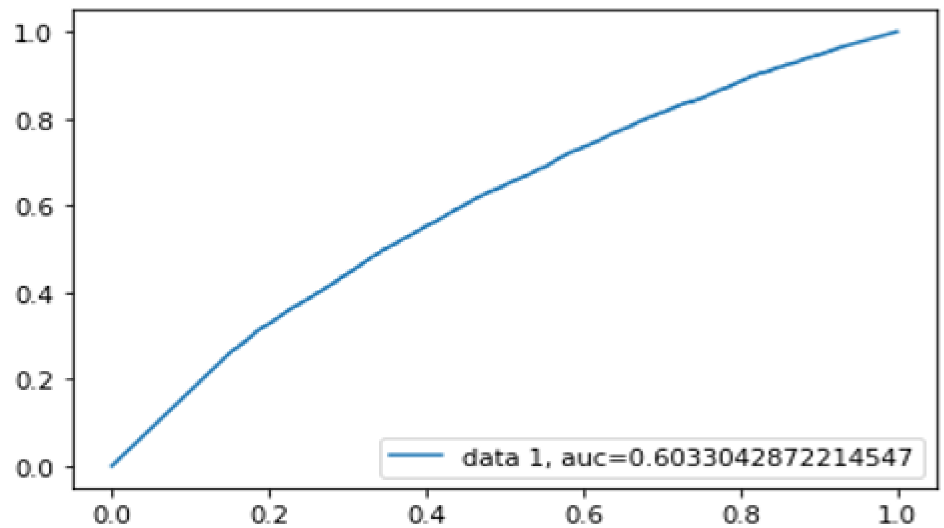
4. Results, Conclusion, and/or Recommendations

1) Logistic Regression

Below table include all continuous variable and their coefficients with significant p-values

Intercept	Tot_Benes (coefficient)	Tot_Srvcs (coefficient)	Tot_Bene_Day_Srvcs (coefficient)	Avg_Sbmtd_Chrg (coefficient)	Avg_Mdcr_Alowd_Amt (coefficient)
-2.52	-0.69	0.43	0.53	-0.95	0.96

AUC chart of the logistic regression model



2) Random Forest

Features selection on Train Set using Elastic Net Logistic GLM

```
Variable Importances:
```

	variable	relative_importance	scaled_importance	percentage
0	Avg_Mdcr_Alowd_Amt	0.939274	1.000000	0.077023
1	Avg_Mdcr_Pymt_Amt	0.775738	0.825891	0.063613
2	Type_Diagnostic Radiology	0.512556	0.545694	0.042031
3	Type_Nephrology	0.424577	0.452026	0.034816
4	Type_Internal Medicine	0.420153	0.447317	0.034454
5	Place_Of_Srvc	0.411892	0.438521	0.033776
6	Avg_Sbmtd_Chrg	0.294762	0.313819	0.024171
7	Type_Family Practice	0.294504	0.313544	0.024150
8	Type_Centralized Flu	0.267694	0.285001	0.021952
9	Type_Interventional Pain Management	0.255371	0.271881	0.020941
10	Type_Physical Therapist in Private Practice	0.252693	0.269029	0.020721
11	Type_Podiatry	0.245579	0.261456	0.020138
12	Type_Anesthesiology	0.244442	0.260246	0.020045
13	Type_Pathology	0.234414	0.249569	0.019223
14	Tot_Bene_Day_Srvcs	0.230888	0.245816	0.018933
15	Type_Clinical Laboratory	0.220627	0.234891	0.018092
16	Type_Mass Immunizer Roster Biller	0.202039	0.215101	0.016568
17	Type_General Practice	0.200167	0.213108	0.016414
18	Type_Mass Immunization Roster Biller	0.198659	0.211503	0.016291
19	Type_Interventional Radiology	0.195609	0.208256	0.016040

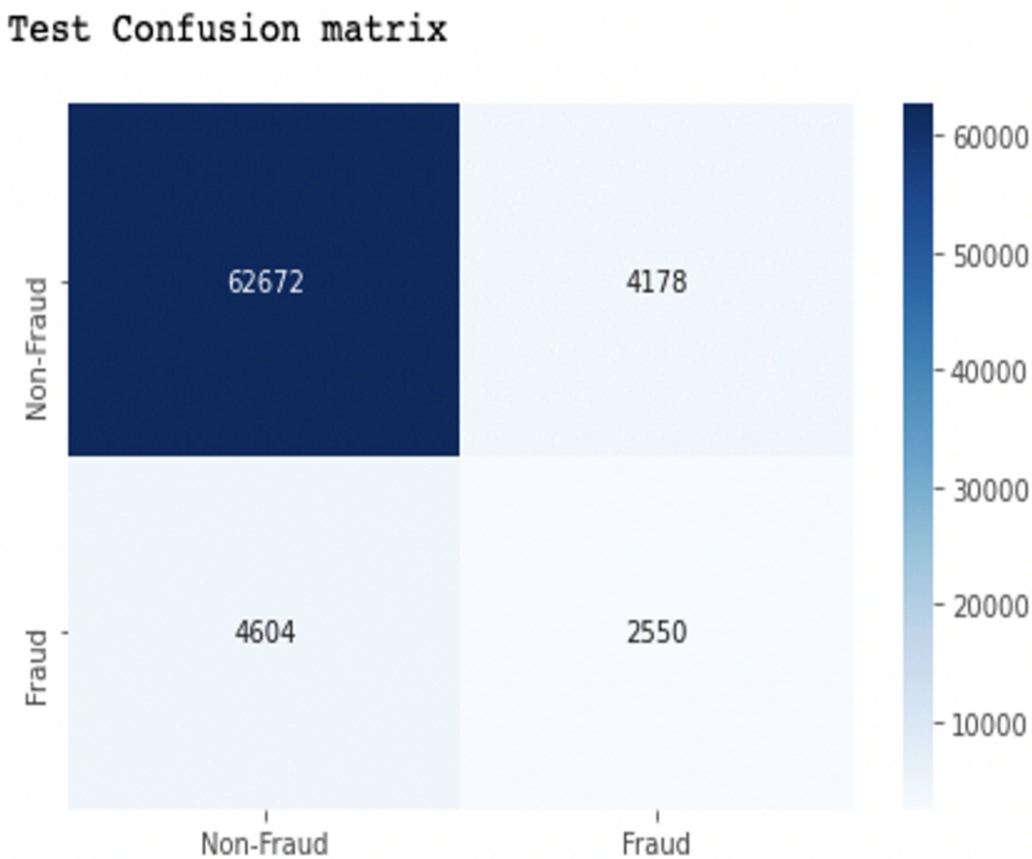
Modeled Random Forest using selected features and list test set performance

	non_fraud	fraud
non_fraud	65426	1850
fraud	5507	1221

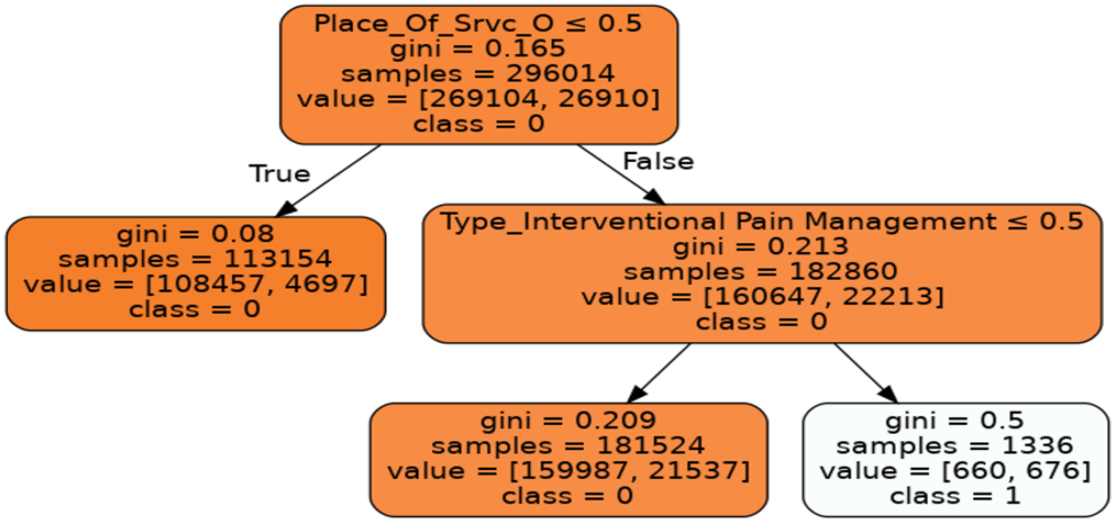
	Matrix
rf_precision	0.659977
rf_roc_auc	0.576991
rf_accuracy	0.900586
rf_f1	0.249209
rf_logloss	3.433642
rf_mse	0.099414

3) Decision Tree

Confusion matrix result on test set



Visualization of decision tree after pruning



4) Explainable Boosting Machine(EBM)

- **Hyperparameters:**
  - Random Grid Search
  - Best Hyperparameters:

Parameter	value	parameter	value
n_jobs	4	outer_bags	4
early_stopping_rounds	100	inner_bags	0
random_state	1234	learning_rate	0.001
max_bins	128	validation_size	0.25
max_interaction_bins	16	min_samles_leaf	10
interactions	15	max_leaves	3

- **Running Time:** 2531.81 sec

- **Average of the five fold evaluation:**

ACC	AUC	F1-Score	Logloss	MSE
0.909	0.792	0.350	0.339	0.102

- **Confusion Matrix with the best cutoff**

Cut-off = 0.44		Prediction	
		1 (Fraud)	0 (Non-fraud)
Actual	1 (Fraud)	3,159	3,569
	0 (Non-Fraud)	8,351	58,925

5) eXtreme Gradient Boosting（XGBoost）

- **Hyperparameters:**
  - Random Grid Search
  - Best Hyperparameters:

-

Parameter	value	parameter	value
booster	GBT	learning_rate	0.5
evaluation_metric	auc	max_depth	7
nthread	4	reg_alpha	0.005
min_child_weight	1	reg_lambda	0.005
colsample_bytree	0.7	subsample	0.9
colsample_bylevel	0.9	gamma	0.2

-

- **Running Time:** 340 sec

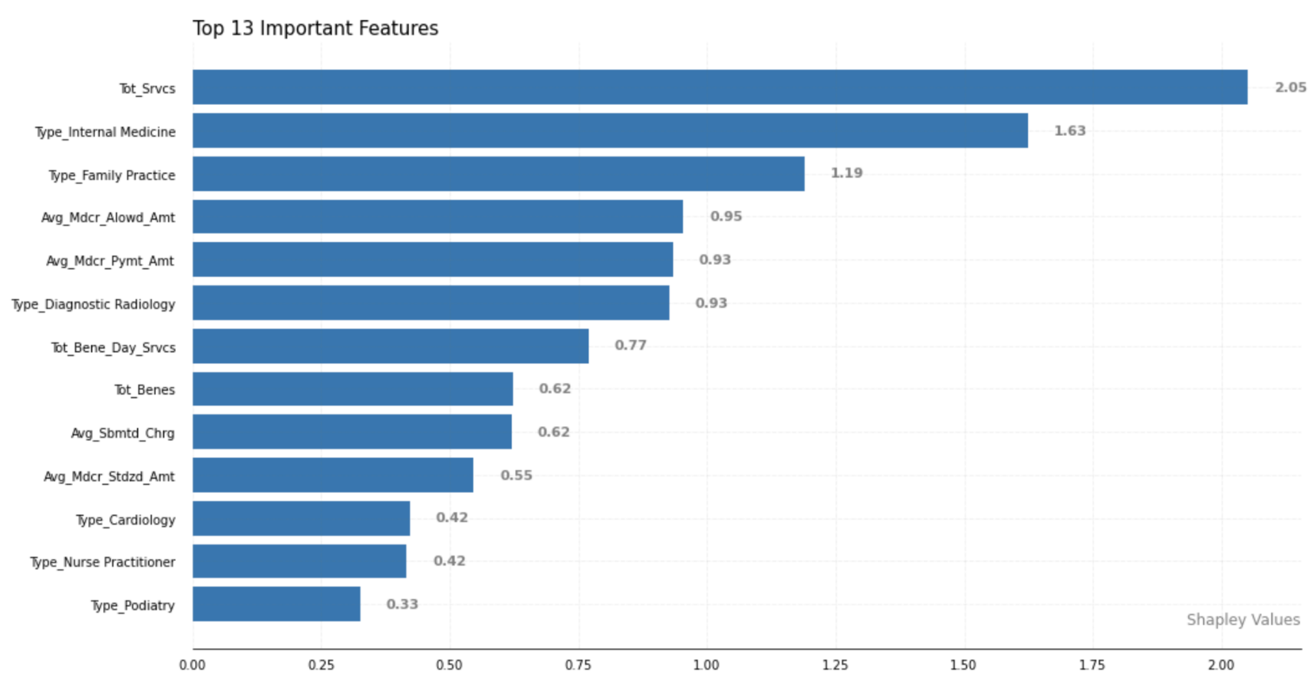
- **Average of the five fold evaluation:**

ACC	AUC	F1-Score	Logloss	MSE
0.914	0.836	0.411	0.274	0.076

- **Confusion Matrix with the best cutoff**

Cut-off = 0.2		Prediction	
		1 (Fraud)	0 (Non-fraud)
Actual	1 (Fraud)	3,163	3,565
	0 (Non-Fraud)	5,686	61,590

- **Global Feature Importance based on Shapley Values**
  - Features with the shapley values above 90th percentiles:



1. **Tot_Srvcs**
2. **Type_Internal Medicine**
3. **Type_Family Practice**
4. **Avg_Mdcr_Alowd_Amt**
5. **Avg_Mdcr_Pymt_Amt**
6. **Type_Diagnostic Radiology**
7. **Tot_Bene_Day_Srvcs**
8. **Tot_Benes**
9. **Avg_Sbmtd_Charg**
10. **Avg_Mdcr_Stdzd_Amt**
11. **Type_Cardiology**
12. **Type_Nurse Practitioner**
13. **Type_Proiatry**

6) Evaluation on Test Set

Evaluation Metric	Logistic Regression	Random Forest	Decision Tree	eXtreme Gradient Boosting (XGBoost)	Explainable Boosting Machine(EBM)
Accuracy	0.70	0.90	0.88	0.914	0.909
AUC	0.603	0.57	0.655	0.836	0.792
F1	0.198	0.24	0.367	0.411	0.350
Log Loss	10.34	3.42	4.099	0.274	0.3390
MSE	0.299	0.09	0.123	0.076	0.102

- **Our recommending classification model**
  - XGBoost model

5. Potential Next Steps for future

- Apply string matching method (like Fuzzywuzzy package in Python) to receive more fraud labels from LEIE Plus.

- Check if the predicting fraud labels from XGBoost match the fraud labels in reality. This could help us to examine the predictability of the XGBoost model.

6. Appendices including code developed for the project

- ipynb files are attached independently