SMU Data Science Review

Volume 6
Number 2 *Summer 2022*

Article 18

2022

Exploration of Data Science Toolbox and Predictive Models to Detect and Prevent Medicare Fraud, Waste, and Abuse

Benjamin P. Goodwin
Southern Methodist University, bgoodwin@smu.edu

Adam Canton
Southern Methodist University, cantona@mail.smu.edu

Babatunde Olanipekun
Southern Methodist University, bolanipekun@mail.smu.edu

Exploration of Data Science Toolbox and Predictive Models to Detect and Prevent Medicare Fraud, Waste, and Abuse

Adam Canton[1], Eli Fuller[2], Benjamin Goodwin[3], Babatunde Olanipekun[4], Chris Papesh[5]
[1] Master of Science in Data Science, Southern Methodist University, Dallas, TX 75275 USA
{eli.fullerin, ovens, chris.papesh}@unlv.edu
{bgoodwin, cantona, bolanipekun}@smu.edu

Abstract. The Federal Department of Health and Human Services spends approximately $830 Billion annually on Medicare of which an estimated $30 to $110 billion is some form of fraud, waste, or abuse (FWA). Despite the Federal Government's ongoing auditing efforts, fraud, waste, and abuse is rampant and requires modern machine learning approaches to generalize and detect such patterns. New and novel machine learning algorithms offer hope to help detect fraud, waste, and abuse. The existence of publicly accessible datasets complied by The Centers for Medicare & Medicaid Services (CMS) contain vast quantities of structured data. This data, coupled with industry standardized billing codes provides many opportunities for the application of machine learning for fraud, waste, and abuse detection. This research aims to develop a new model utilizing machine learning to generalize the patterns of fraud, waste, and abuse in Medicare. This task is accomplished by linking provider and payment data with the list of excluded individuals and entities to train an Isolation Forest algorithm on previously fraudulent behavior. Results indicate anomalous instances occurring in 0.2% of all analyzed claims, demonstrating machine learning models' predictive ability to detect FWA.

1 Introduction

Medicare fraud, waste, and abuse (FWA) is a problem on the national scale, causing an enormous burden (e.g., loss of billions of dollars, Medicare Learning Network, 2021) on public finances and the allocation of resources for some of the most vulnerable populations (Medicare Learning Network, 2021). Actions that constitute FWA are poorly defined, but The Centers for Medicare & Medicaid Services (CMS) provides a list of the kinds of claims that provides a list of example claims that constitute FWA (Medicare Learning Network, 2021). Despite a multitude of prior studies concerning the topic of FWA, an exact amount of Medicare FWA is difficult to closely approximate because much of the FWA simply goes undetected. Conservative estimates claim 3-10% ($19 billion to $65 billion (about $200 per person in the US)) of all medical claims

fit into FWA (Bauder & Khoshgoftaar, 2018, p. 9), while other estimates claim the number is much higher and somewhere closer to $300 billion (about $920 per person in the US) (Nicholas, Segal, Hanson, Zhang, & Eisenberg, 2019, p. 788). Medicare is a large complex government program and in its current form has few controls in place to detect FWA. Unlike private insurers who use complex algorithms to detect FWA and subsequently deny such claims before they are paid, Medicare operates in an opposite fashion, paying providers first, and then investigating FWA (Pande & Mass, 2013, p. 10). The Centers for Medicare & Medicaid Services (CMS) provide publicly accessible data containing fields such as provider information and payment data. CMS also maintains a list of excluded providers, The List of Excluded Individuals and Entities (LEIE). The LEIE contains the individuals and entities reprimanded for FWA. CMS data provides an ample number of resources and records to train and develop a machine learning model to disseminate between legitimate Medicare claims and those considered FWA. The potential cost and labor savings for the Department of Health and Human Services from FWA detection via machine learning is significant.

Despite the variability of dollar estimates for FWA, one theme is clear, Medicare suffers from rampant FWA, and these costs are absorbed by the Federal Government as well as the taxpayers of the United States. CMS provides publicly available data on Medicare through various applications and databases and with this data this study proposes implementation of machine learning algorithms to generalize and detect FWA in current and future claims.

Machine learning techniques can complement the efforts of the Department of Health and Human Services in the investigation of potentially fraudulent claims. Such machine learning algorithms could serve as a first layer of detection, which human claims inspectors can further analyze to determine if the claim is a potential candidate for FWA. This step in the investigative process has the potential to save hundreds of thousands of labor hours per year as the algorithm has the capability to classify claims as “potentially fraudulent” or “not potentially fraudulent,” saving the humans from the task of manually reviewing each Medicare claim. FWA are often difficult to detect and pursue and the costs of pursuing individual cases many not always be worth the expense, especially in the case where the FWA is a legitimate provider making an honest mistake, or even the case when a dishonest provider only slightly overbills, invoking an investigation costing more than the FWA itself. In the scenario where an algorithm can make this first determination has the potential to save tremendous amounts of money and investigative labor.

The subject of FWA is an often researched and discussed problem within academia and beyond (Pande & Mass, 2013, p. 9). The topic has even finally reached the point of Federal Government admission, with President Obama making the elimination of healthcare FWA a top priority of his administration (Pande & Mass, 2013, p. 10). The Federal Government announced in 2011 that it would include predictive data modeling to assist in the fight against FWA prior to the payment of claims (Pande & Mass, 2013, p. 10). Yet despite this public affirmation of the issue of FWA, 12 years later the problem persists and my many accounts, only continues in increases prevalence (Pande & Mass, 2013, p. 9).

Among the various issues associated with solving such a problem revolves around access to data. Despite the large amount of publicly accessible data, this research is tasked with determining which data best pertains to the topic at hand and whether the

data answers the question of interest. The CMS offers robust publicly available data and will be utilized throughout the duration of this study.
The research team believes research in the field of Medicare FWA combined with machine learning will yield statistically significant and insightful results. Applying machine learning techniques to Medicare data, implemented according to the methodologies reported above can and will identify situations where fraud, waste, and abuse are present.

2 Literature Review

Much of the previous research on the topic of Medicare FWA have attempted to generalize fraud and train machine learning models to detect providers who commit Medicare FWA (e.g., Musal, 2010, Liu et al., 2016, Herland et al., 2017, Zhang & He 2017, Bauder & Khoshgoftaar 2018, Obodoekwe & van der Haar, 2019). Specifically, each prior paper has targeted a specific aspect of Medicare where the authors believe FWA could potentially be generalized, such as at the provider level (Bauder & Khoshgoftaar 2018, p. 9), or after the services have been rendered, at time of payment to the provider (Pande & Mass, 2013, p. 10). The researchers believe the prior studies have done an adequate job generalizing specific tenants of the problem but contain several shortcomings that need to be addressed. These shortcomings include, using multiple years of available data, different choices of classification algorithms, and linking the excluded providers with the claims data. The bulk of this analysis is centered around using machine learning to determine if a given claim is legitimate or FWA.

This research aims to use machine learning to identify possible fraudulent trends or activity using public Medicare Data. This research plans to use 3 years of Medicare Part B and Part D data along with labels collected from the LEIE to generate some insights on Medicare FWA. Since occurrences of FWA are rare when compared to legitimate claims, the researchers plan to mine these cases for insight and then use unsupervised techniques to identify FWA behavior.

The data sets are quite large and combining them presents many challenges. The public Medicare data sets are not released at the patient/event level. They are instead aggregated on National Provider Identifier (NPI) (a unique identification number assigned to all covered health care providers) and other characteristics (Bauder & Khoshgoftaar 2018 p. 2). This means the researchers will have to join some large datasets together to examine the relevant features required for this analysis. Anomaly detection, one of the preeminent methods of fraud detection is employed in many different areas such as procurement fraud, credit card fraud, and Medicare fraud detection (Zhang & He 2017 p. 310). The assumption being that anomalous events or activity is likely to be fraudulent when compared with the rest of the body (Bauder & Khoshgoftaar 2018 p. 3). Previous researchers have used Spatial Density using imLOF (Improved Local Outlier Factor) (Zhang & He 2017 p. 311). As well as unsupervised methods such as Isolation Forest and Unsupervised Random Forest (Bauder, Rosa, & Khoshgoftaar 2018 p. 285), Deviation Clustering, Gaussian Mixture

Models, and Bayesian Co-clustering (Ekina, Leva, Ruggeri & Soyer 2013 p. 151). Further, past researchers have seen that Local Outlier Factor (non-improved), K-Nearest Neighbors, and autoencoders are suboptimal performers (Bauder, Rosa, & Khoshgoftaar 2018 p. 286). Though there is some discussion over LOF (Bauder, Rosa, & Khoshgoftaar 2018 p. 287), (Zhang & He 2017 p. 310). The researchers involved in the study "An Anomaly Detection Method for Medicare Fraud Detection" designed a new LOF metric designated imLOF for improved Local Outlier Factor. This metric is designed to detect excessive medical treatment and decomposing hospitalization using spatial density information. (Zhang & He 2017 p. 312) The original measure, developed by Breunig et al. (Breunig, Kriegel, Ng, & Sander 2000 p. 94), gives anomaly scores based on the density of observations; they noted that the density of anomalous events would be less than that of its normal neighbors. This metric has issues with small clusters making it suboptimal for healthcare use (Bauder, Rosa & Khoshgoftaar 2018 p. 288), (Zhang & He p. 312). The authors give an example of a small cluster of hypertension patients with a great deal of fraud. Since the cluster is small and the point's neighbors are also likely to be anomalous the metric scores a low chance of anomalous activity. The authors then suggested an improvement to the LOF score by adding the size of a cluster into consideration instead of only density, with the additional use of the DBSCAN algorithm the improved LOF score performed much better on healthcare data (Zhang & He 2017 p. 312)

Some research has only focused on single years (Bauder, Rosa, & Khoshgoftaar, 2018 p. 288) (Gordon & Siegel 2020 p. 1), (Hancock & Khoshgoftaar 2020 p. 572) and used either a Supervised Learning design or a combination of unsupervised and supervised (Bauder, Rosa & Khoshgoftaar 2018 p. 288) (Meyers 2017 p. 251) Our design will follow the latter. There have also been studies directed at specific portions of Medicare/Medicaid such as dental, otolaryngology (Ekina, Leva, Ruggeri & Soyer 2013 p. 151) and dermatology (Gordon & Siegel 2020 p. 1). The use of a single year is due to the size of the data, given that a single year's Medicare Provider Utilization and Payment Data for Part B is around 10 million records, 29 columns, and about 3 GB of memory by itself.

An issue with the current data is proper class balance in distribution of the target classes. This was an issue in all of the studies that used supervised methods. Such severe imbalance requires careful data adjustments to somehow increase the representation of the minority class, in this case fraudulent activity. This means some type of special sampling method. Bauder et al. (2018) indicated that random under sampling provided the best results followed by SMOTE (Synthetic Minority Oversampling Technique) (Bauder & Khoshgoftaar 2018 p. 3). In random under sampling (RUS), the algorithm throws out random events from the majority class, thereby increasing the representation of the fraudulent cases. In the case of SMOTE, a more advanced algorithm is used to create new minority class instances by first finding a minority class instance and its k nearest neighbors. Then a new instance is created by choosing a random neighbor and combining it with the original instance. In a further study by Bauder and Khoshgoftaar an RUS method was used along with an adjusted cost function (Bauder and Khoshgoftaar 2018 p. 5)

This research will follow heavily in the footsteps of prior research in the field of Medicare FWA. Modern machine learning and data exploration techniques will be

exploited for the purposes of better understanding the drivers and factors behind Medicare FWA. Several of the techniques and methodologies referenced above will be modified and adapted for the purposes of this research into Medicare FWA.
Using machine learning algorithms, prior research has indicated FWA is possible to detect, but often these prior studies are inconclusive in their test for statistical significance when testing against the hypothesis of a difference between a fraudulent claim and a legitimate claim. This is primarily due to a handful of challenges split between Medicare data and the current limitations of machine learning. The first challenge concerns the balance of legitimate claims against illegitimate ones. For example, if 3-10% ($19 billion (about $58 per person in the US) to $65 billion (about $200 per person in the US)) of all medical claims fit into FWA (Bauder & Khoshgoftaar, 2018, p. 9), approximately 90% of claims are legitimate. This indicates that a machine learning model could perform reasonability well by simply classifying all transactions as legitimate. Class imbalance is a significant issue with Medicare data and will require new and novel approaches to overcome. Fortunately, credit card companies and other large organizations that process huge volumes of transaction data have deeply studied such topics and have developed formidable and complex anti-fraud and theft systems. Unfortunately, much of implementation of these systems is proprietary, nonetheless this indicates that large class-imbalances can be overcome.
In addition to class-imbalance, the issue of training data presents a significant barrier. Medicare claims data is just that, information on claims and aggregated to the procedure and National Provider Identifier Standard (NPI), there is no comprehensive database of claims data complete with an indicator if the transaction is a legitimate one or categorized as FWA. The proposed solution involves integration of claims data and the excluded provider list. The intention is to examine those who populated the excluded provider list, and then determine the last year they submitted claims, and find those claims, add the claims and provider to the training data to train the model based on the claims behaviors of known excluded providers. Adding the excluded providers in addition to the legitimate claims will provide the model with the ability to differentiate between the behaviors of FWA and real claims data.
A few examples of fraudulent Medicare claims include:

1. Claims for appointments that were not attended by the patient.
2. Claims for more complex services than those performed or required.
3. Claims for services that that were not carried out (Johnson & Khoshgoftaar 2019, p. 18).

On the other hand, Medicare abuse includes the practice of knowingly providing medically unnecessary services to patients against recognized standards. For example, misusing billing codes for personal gain. There are applicable Federal laws that prohibit Medicare fraud and abuse. These include the False Claims Act (FCA) and Anti-Kickback Statute (Johnson & Khoshgoftaar, 2019, p. 18).

Bauder et al. (2018) applied several anomaly detection techniques to segment medical provider fraud in the 2012 to 2015 Medicare Provider Utilization and Payment Data: Physician and Other Supplier which is publicly available from the Center for Medicare and Medicaid Services. To evaluate the performance of candidate learners, the authors mapped fraud labels dataset using the List of Excluded Individuals and Entities (LEIE). The novelty in their study is the application of

Isolation Forest and Unsupervised Random Forest on this big Medicare dataset. Bauder et al., 2018 worked with only half of their preprocessed dataset as they reduced the dataset from 3.7 million to 1.8 million due to hardware limitations. Prior research on Medicare FWA has studied algorithms such as XG-Boost, CatBoost, and Gradient Boosted Decision Trees (Hancock & Khoshgoftaar, 2019, p. 572). While extremely useful the machine learning space, these algorithms have proven to be inconclusive on Medicare FWA (Hancock & Khoshgoftaar, 2019, p. 578). Other techniques such as regression and clustering analysis have been deployed with comparable results (Musal, 2010, p. 2828). Chief among these algorithms is a basis in frequentist statistics. Looking at Medicare FWA through a Bayesian lens could provide the missing link between sporadic classicization and a clearly defined approach to detecting FWA. The notion of prior probabilities can help better train models, and approach the data with a unique perspective, treating both model parameters and data as random. As the algorithm evaluates an individual claim, the assistance of a prior probability related to its legitimacy has the potential to facilitate updating the likelihood of flagging transactions as FWA or legitimate with more precision. Better performance was reported by Johnson and Khoshgoftaar 2019 who reported AUC score > 85% from similar Medicare dataset. They tackled the class-imbalance with random over-sampling and random under-sampling techniques prior to fitting it on a 3-layer dense neural network. Bauder et al., 2018 could have achieved better performance by using random under-sampling and/or random over-sampling to manage the severe class-imbalance in the dataset. Other methods that have been applied to the Medicare dataset for fraud detection. Liu et al., (2016) utilized isolation forest; Bauder et al., 2016 compared supervised learning techniques: Gradient Boosted Machine, Random Forest, Deep Neural Network, and Naive Bayes and a suite of unsupervised learning techniques: autoencoder, Mahalanobis distance, KNN, and local outlier factor, and hybrid (multivariate regression and Bayesian probability) machine learning approaches.

3 Data

The Center for Medicare and Medicaid Services maintains an extensive repository of claims data, which the researchers used as the basis for modeling. The Medicare Provider Utilization and Payment Data, hereafter referred to as MPUPD, contains Medicare data aggregated to the NPI-procedure level, in less formal language this data contains pertinent details “on services and procedures provided to Medicare beneficiaries by physicians other healthcare professionals.”(Red Hat Marketplace, 2022, “Overview” section) It is important to note that CMS maintains claims data for several years, and the currently available data was collected between 2013 and 2019.

In addition to claims data, CMS also provides the List of Excluded Individuals and Entities. This list is updated monthly by CMS and the OIG (Office of the Inspector General) to reflect any actions taken against individuals or entities committing fraud, waste, or abuse towards the system, this dataset is an important piece of the analysis and is paired with claims data to develop a sense for behaviors that could indicate FWA. This can cause issues for supervised learning methods because the LEIE is

aggregated only to the NPI level. Thus, to get a one-to-one relationship, the MPUPD data must be aggregated up to the same level (Bauder & Khoshgoftaar 2018 p. 2). Among the various challenges with Medicare data are its sheer size, complexity, and attempts at anonymity. Two of these concerns (size and complexity) were mitigated through techniques such as aggregation to the NPI and year level to condense the data into a more manageable size. Breaking apart the data allowed the researchers the benefit of reduced computational complexity as well as the ability to utilize k-fold cross validation across the dispersed data. The other and more complex challenge concerned matching MPUPD data with LEIE data. The LEIE data is maintained as a running list. Individuals or entities that enter the list are not usually excluded permanently. This means that in a year or set of years an entrant may be removed from the list, and any interested parties viewing the list in the future would not know the entrant had ever been excluded. These movements on and off the LEIE are kept track of in monthly supplements. However, CMS only keeps these supplemental files for a period of one year. Since the researchers' data runs from 2013 to 2019 and the last supplemental files for the LEIE contain 2021 data, the researchers failed to completely capture all excluded individuals in the data. The only excluded individuals or entities the researchers have access to are those that have been excluded permanently or are currently excluded and have not been reinstated. This leaves a large information gap where the researchers cannot see the individuals or entities who were once on the list but have now been reinstated.

To illuminate the issue, suppose Dr. X is found to have committed fraud in 2015 and is entered into the LEIE. Their exclusion causes privileges to be suspended for two years. The LEIE records their NPI, year of a fraudulent judgment, and level of fraud infraction amongst other features. The MPUPD will have recorded that Dr. X provided some number of services for each HCPCS code aggregated that year. Dr. X the mandatory two years and is reinstated in 2017 and thus no longer in the LEIE. The only record available of Dr. X's exclusion now that they are not on the list is the supplemental record of their reinstatement, which will no longer be kept after 2018. A researcher investigating fraud in 2020 uses the LEIE to label data and Dr. X is included in the MPUPD, but is no longer included in the LEIE, and there is no record available to correct this, since the supplemental reinstatement record was lost over two years ago. Consequently, this would lead to fraudulent patterns being identified as non-fraudulent.

The final challenge of the data is that fraud occurring is rare, in the context of machine learning this is considered a class imbalance and in terms of the CMS data the rough split between legitimate transactions and FWA is 90% and 10% respectively. Unfortunately, a data split such as the one present in CMS data will cause undesirable performance by machine learning algorithms and requires some manipulation of the data.

Data strategy: Merging MPUPD and LEIE datasets

To produce meaningful results, the data required significant manipulation to correctly drive inference from machine learning algorithms. The exact steps are listed below:

1. The data was first merged, combining MPUPD with LEIE to form a new dataset titled, "MergedWithLables" with the intention of having a sole source of data for the algorithms to process.
2. The next step included generating feature columns and aggregating all numerical columns so that each column contains a minimum, maximum, mean, and standard deviation as included feature to assist in aggregation.
    a. Categorical columns with keyword, "type" necessitated cleaning, as an example of this is the designation "MD" which occurs in the data as "MD"," M.D.," and "M.D," each indicates medical doctor, however categorical breakout considers these separate designations.
    b. All instances of these keywords were corrected and aggregated.
3. Additionally, categorical variables were on-hot encoded to further facilitate modeling efforts and ensure interpretability of the model.
4. This new dataset was then split into several smaller datasets to ensure less computation time and to provide a variety of training and testing datasets for the researchers to monitor results.

Figure one visually describes the basic data structure and preparation roadmap.

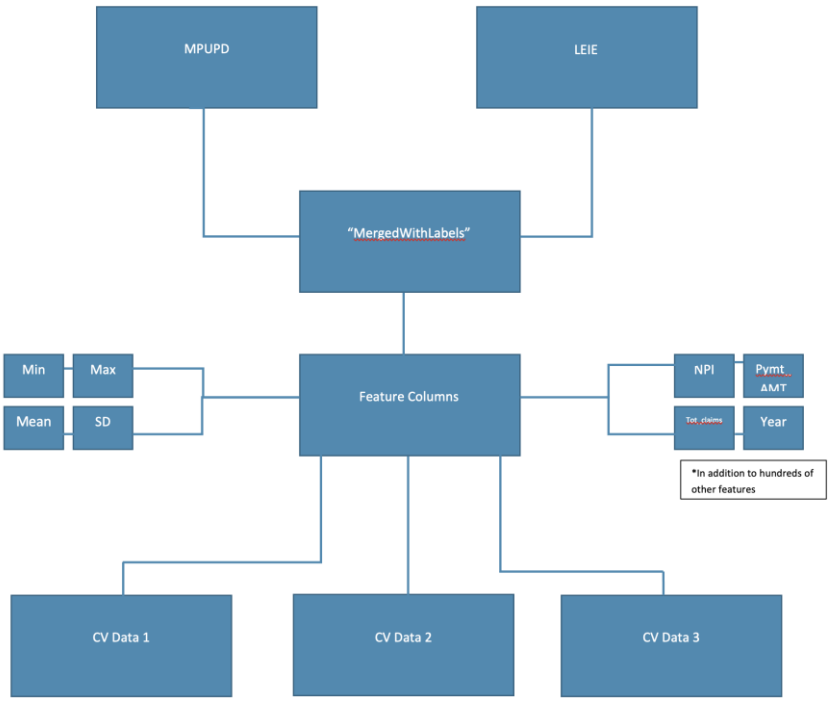


Fig 1. Data Structure and Preparation Roadmap.

4 Methods

Using modern and novel approaches to machine learning, the goal is to analyze publicly accessible Medicare data to determine patterns behind FWA. While the algorithms play an extremely key component in this research, the dissection of claims data and subsequent paring with excluded provider information is a unique approach to the problem of detecting Medicare FWA. Pairing claims data with excluded provider information is a key differentiator between this study and those proceeding and should result in significantly different results.

Prior studies have identified some of the pitfalls of attempting to identify Medicare FWA. Among those pitfalls concern the Local Outlier Factor (LOF) producing a 63% Area Under the Curve, just 13% higher than random. As also averred by the study, such a low score makes it difficult to deploy this study for real world application. Situations like this usually demands scrutiny of the data and its source/s. One explanation for the low AUC scores could be a lack of known fraudulent providers to use as fraud labels for validation, creating a highly imbalanced dataset (Bauder & Khoshgoftaar 2017, p. 860).

The imbalance issue is further complicated by how the data is kept. The data is stored by separate institutions, CMS hosts MPUPD and the OIG maintains LEIE. The LEIE does not have event level data. The Medicare Physician Utilization and Payment Data also does not contain event level data, but aggregated data. The LEIE uses NPI-Year as a key, whereas the MPUPD uses NPI-Year-HCPCS code. Where HCPCS code is the type of service being provided. To transform the MPUPD data into a form where labels can be connected, the HCPCs codes and other values are aggregated over the year. Continuing with the example of Dr. X as a reference, the researchers were put in the position of instead of knowing Dr. X performed 10 of service y and 14 of service z the data only identifies Dr. X as performing 24 services. This loss of information is necessary to connect labels. To limit the loss of information, the team takes several statistics of the aggregated values, such as min, median, mean, max, and standard deviation. Once the label was connected, the researchers determined that the connected NPI had committed a fraudulent act in that service year. However, it is not known which set of services in that year were, in fact fraudulent.

The challenges created by the data set inconsistencies have pushed the researchers toward unsupervised models which do not rely on labels to analyze patterns. Anomaly detection through several types of clustering and other unsupervised methods are popular for this space and well represented in open-source tools.

This research will use entirely open-source models and publicly accessible data to embrace transparency in claim evaluation criteria.

Given the size (5.5 million rows and over 100 features) of the dataset, Apache Spark libraries were used for tasks associated with importing the data into the model as well as exploration and feature engineering. For the task of feature engineering, the

mean and standard deviation attributes of the aggregates, where applicable, were used while the other minimum, maximum, and median were removed from the dataset. Special consideration had to be made to the multitude of categorical variables due to the many designations of categorical variables at the provider specialty level. Categorical columns with less clear patterns of designation were corrected, as an example “Obstetrics & Gynecology” and “Obstetrics /Gynecology” were corrected to show a single designation of “Obstetrics/Gynecology.”

The Isolation Forest model was selected as the analysis model due to its performance characteristics for tasks such as anomaly detection. The first strategy considered by the researchers was K-means clustering, but this was abandoned due to the sparse, abnormal behavior of fraudulent claims. The primary concern of the researchers with this approach was K-means inability to correctly cluster fraudulent claims among legitimate claims. According to *Towards Data Science* “A lot of machine learning algorithms suffer in terms of their performance when outliers are not taken care of.” (Lewinson, 2021, “Introduction” section) Considering this research aims to detect and classify outliers in the Medicare space, this was a huge concern. The Isolation Forest was selected because of the novel approach to outlier detection. Again, *Towards Data Science* offers a clear and concise explanation of the advantages of using The Isolation Forest as a means of outlier detection, “Isolation Forest explicitly identifies anomalies instead of profiling normal data points. Isolation Forest, like any tree ensemble method, is built based on decision trees. In these trees, partitions are created by first randomly selecting a feature and then selecting a random split value between the minimum and maximum value of the selected feature.” (Lewinson, 2021, “Some Theory First” section)

Model Performance

After training, the Isolation Forest returns values between 0 and 1, these classification bins indicate the extent of deviation from normal instances with values closer to 1 can be considered anomaly instances. The baseline version of the Isolation Forest utilized 1.110 million rows (35%) of the aggregated Medicare dataset for the modeling. Sub-setting the dataset was necessary as required by the researcher's data architecture (see figure 1) and reduced the computational requirements of running the analysis on the full dataset. The number of trees and subsamples were set to 1500 and 4096 respectively, these tuning parameters were optimized based on different modeling runs and their subsequent output. The results from the baseline version detected 2,240 anomalous instances in the dataset containing 1.110 million rows of Medicare claims data. In terms of performance, the model indicates that 0.2% of all claims contained within this reduced dataset can be considered abnormal and serve as a starting point for CMS to further investigate these claims. Figure 2 contained within the technical appendix contains the yearly breakout of anomalous and normal instances.

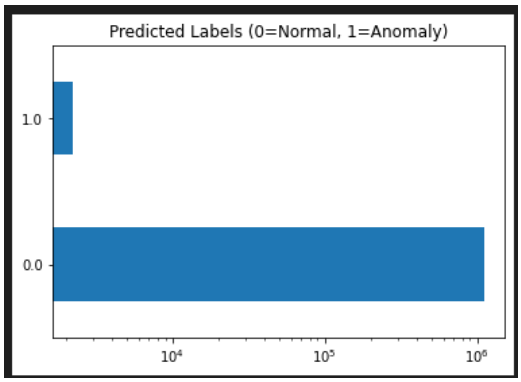


Fig 2. Log plot showing distribution of predicted anomaly labels for 1.110M rows on Medicare claims dataset. 1 represents anomalous instances and 0 represents normal instances

Machine learning is a subjective science and criteria for success can be measured on any number of levels. Since this data is extremely imbalanced and an algorithm that simply classifies all claims as legitimate would perform at around 90% accuracy. As a solution to this problem, F1 score will be used as the primary metric for success, F1 is an evaluation metrics that balances precision and recall. F1 is the harmonic means of both precision and recall and will balance the overabundance of legitimate claim data with the need to detect and accurately classify transactions categorized as FWA. For the purposes of this research, the threshold for a successful algorithm will achieve an F1 score of >0.6. In addition to the F1 metric, the research team will also look at the contextual implications of additional criteria to evaluate performance of different algorithms.

5 Context Architecture

Through the research of Medicare data, the research team hopes to generalize the common occurrences of fraud, abuse and overbilling throughout the Medicare system using machine learning. The research team will determine “good” results using various statistical tests as well as participating in a continuous feedback loop with stakeholders who regularly handle and identify fraud, abuse, and overbilling.

Exploratory Data Analysis

The dataset used for this project contains 3,330,000 instances and 105 Attributes. These were aggregated to show the distribution of claims among the uniquely identified providers (Figure 3). Most of the claims studied occurred are categorized in the diagnostic radiology and internal medicine. Interestingly podiatry,

gastroenterology, and urology, respectively, have the least claims (Figure 4). To provide a perspective on the claims submitted, in 2019 (Figure 5), 2.59 billion services were performed on 879 million individuals at a cost of $3.81 billion (about $12 per person in the US) (about $12 per person in the US). Of these charges, Medicare made a payment of $816 million. Of particular interest is the distribution of providers in the LEIE. Figure 6 shows that the proportion of these individuals correlates with the population distribution of states in the USA. For instance, the most populous states of the United States (California, New York, and Florida) record the highest number of providers found in the Excluded List.

Fig 3. Aggregations Over National Provider Identifier (NPI) 2017-2019

NPI	Rndrng_NPI	Count of Rndrng_NPI	Tot_Benes
1003000407	1003000407	13	607
1003066838	1003066838	19	617
1003127655	1003127655	8	2218
1003811167	1003811167	2	78
1003850603	1003850603	1	90
1003870239	1003870239	7	350
1003892746	1003892746	5	120
1003904830	1003904830	14	234
1013009729	1013009729	6	213
1013069780	1013069780	21	1305
1013074525	1013074525	6	150
1013087741	1013087741	11	227
1013095975	1013095975	5	154
1013167311	1013167311	1	67
1023006129	1023006129	1	16
1023031481	1023031481	5	215
1023077047	1023077047	9	666
1023156320	1023156320	4	1172
1023230562	1023230562	4	1116
1043241870	1043241870	9	716
1043299159	1043299159	19	645
1043319866	1043319866	28	3382
1043445927	1043445927	13	911
1043470370	1043470370	4	91
1053360966	1053360966	5	409
1053423764	1053423764	8	638
1053486704	1053486704	16	940
1053533919	1053533919	31	1834
1063420529	1063420529	1	25
1063452167	1063452167	1	18
1063568475	1063568475	24	961
1063575561	1063575561	7	157
Total		**4287**	**448048**

NPI	Rndrng_NPI	Count of Rndrng_NPI	Tot_Benes
1003000407	1003000407	10	305
1003066838	1003066838	23	956
1003127655	1003127655	9	1724
1003850603	1003850603	2	94
1013087741	1013087741	4	92
1023031481	1023031481	5	187
1023156320	1023156320	2	746
1023230562	1023230562	4	647
1043319866	1043319866	17	932
1043445927	1043445927	15	756
1043470370	1043470370	5	89
1053423764	1053423764	4	200
1053486704	1053486704	8	284
1053533919	1053533919	13	758
1063568475	1063568475	7	239
1063728483	1063728483	9	379
1073557047	1073557047	3	70
1073593919	1073593919	22	810
1073624912	1073624912	3	80
1073645545	1073645545	1	12
1083607741	1083607741	13	774
1083687859	1083687859	6	269
1093750804	1093750804	12	948
1093800609	1093800609	34	3968
1093904914	1093904914	16	1084
1104829639	1104829639	43	3906
1114964442	1114964442	34	4143
1124292966	1124292966	14	508
1124369863	1124369863	6	267
1134192370	1134192370	2	40
1134221351	1134221351	10	607
1134228794	1134228794	5	136
Total		**2111**	**219428**

NPI	Rndrng_NPI	Count of Rndrng_NPI	Tot_Benes	Tot_Srvs
1104829639	1104829639	43	3698	54(
1285673012	1285673012	38	2107	1078
1154334381	1154334381	30	1334	65
1114964442	1114964442	25	2785	38
1497910111	1497910111	25	1302	53
1598746919	1598746919	21	913	9(
1134232887	1134232887	20	1225	25
1770505612	1770505612	20	1715	31
1871571406	1871571406	20	941	401
1093904914	1093904914	19	812	16
1386634293	1386634293	18	1010	26
1528139722	1528139722	18	923	19
1851340731	1851340731	18	1228	24
1386626968	1386626968	17	737	16
1609801877	1609801877	16	505	23
1821009390	1821009390	16	2934	39
1871531350	1871531350	15	398	13
1922005990	1922005990	14	496	67
1679754725	1679754725	13	951	25
1922280254	1922280254	13	1449	33
1568485233	1568485233	12	1535	58
1922080415	1922080415	12	876	38
1295831659	1295831659	11	537	35
1366638736	1366638736	11	1938	205
1639113012	1639113012	11	221	25
1801860507	1801860507	11	580	15
1982662722	1982662722	11	779	10
1144363011	1144363011	10	161	16
1164587697	1164587697	10	930	42
1326209354	1326209354	10	780	14
1528045333	1528045333	10	374	4(
1629183223	1629183223	10	1044	35
Total		**843**	**52966**	**3134**

Fig 4. HCPCS Codes by Provider Type – an indication of volume 2019

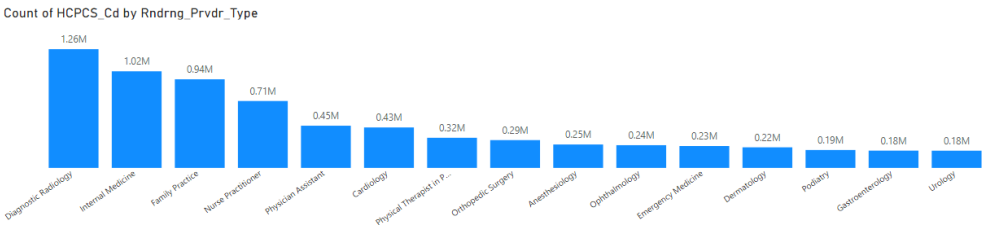


Fig 5. Some Summary Stats from 2019 MPU and PD

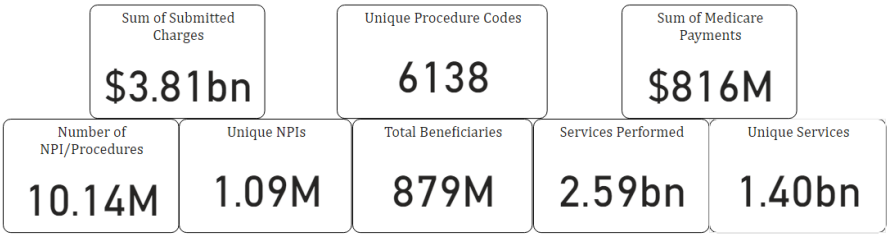


Fig 6. Excluded Individuals and Entities by Year (Compare to NPIs (National Provider Identifier) in 2019)

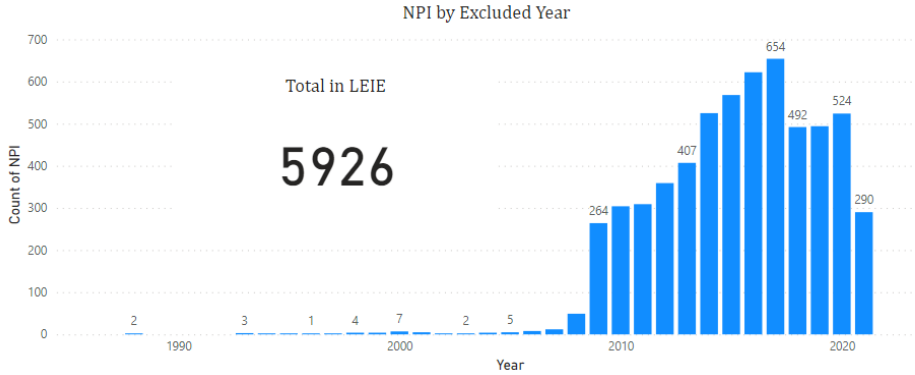


Fig 7. Excluded Individuals and Entities by State

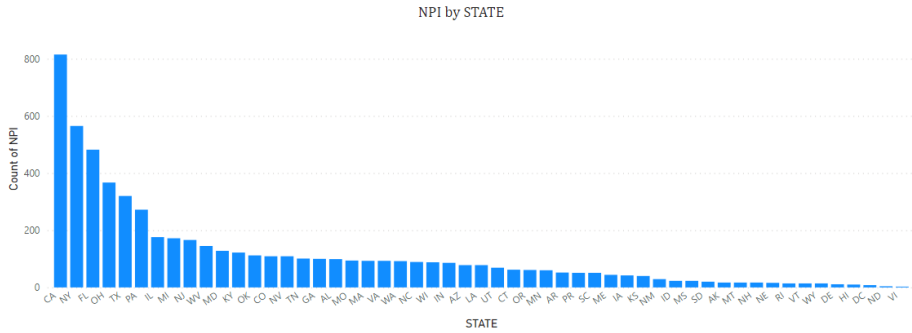


Fig 8. Differences and Errors in Credentials and Specialties MPUPD 2019 v LEIE

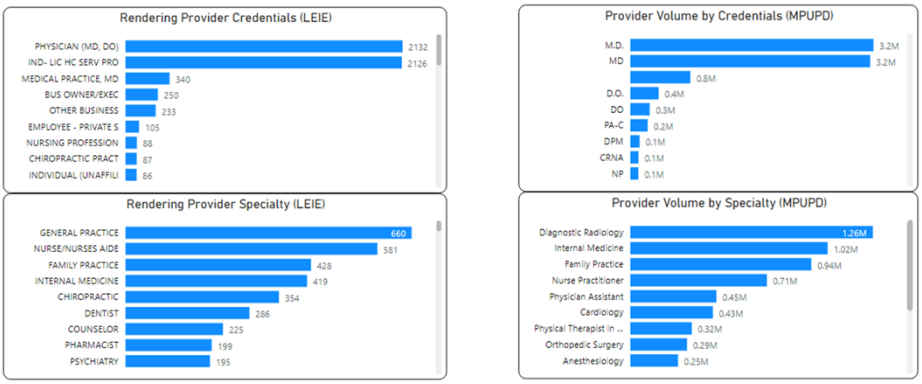


Fig 9. Word Cloud of HCPCS Code Descriptions (Stop Words Removed)

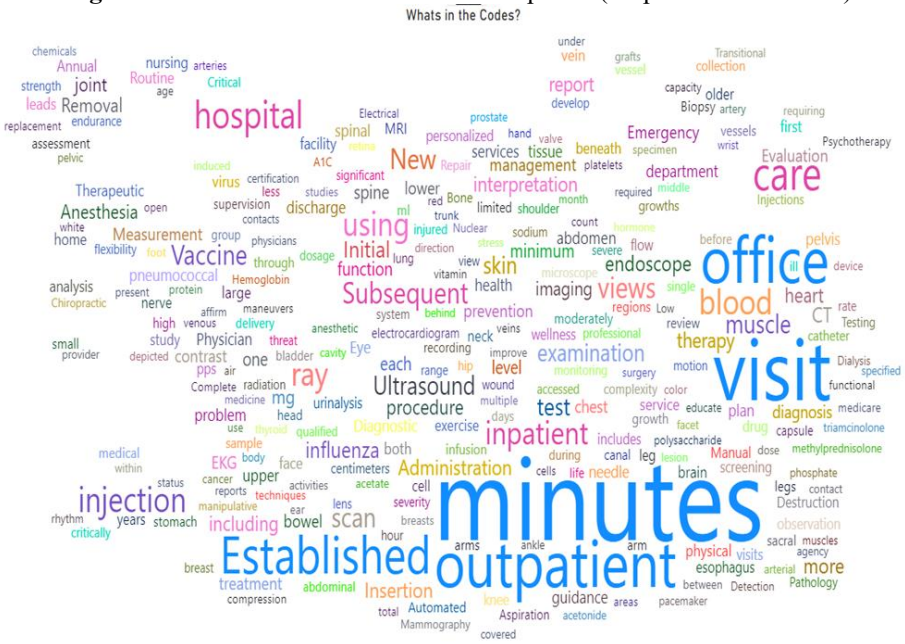


Fig 10. Median Unique Services Performed by NPI YoY 2013-14 (Top 50)

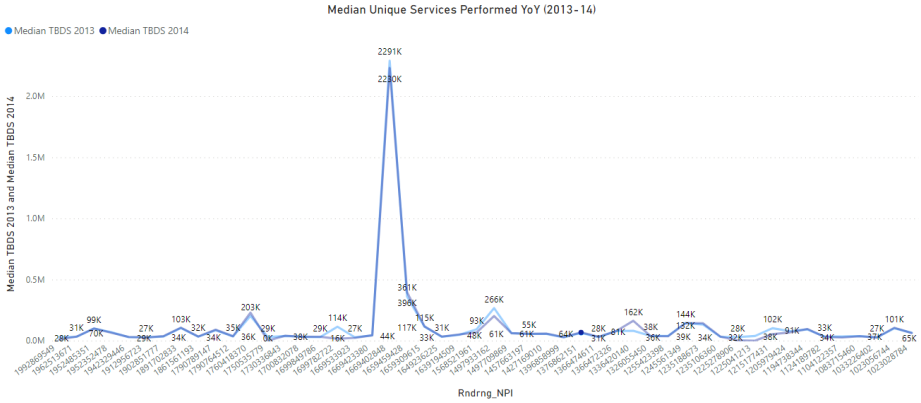


Fig 11. Average Medicare Payments by NPI YoY 2013-14 (Top 50)

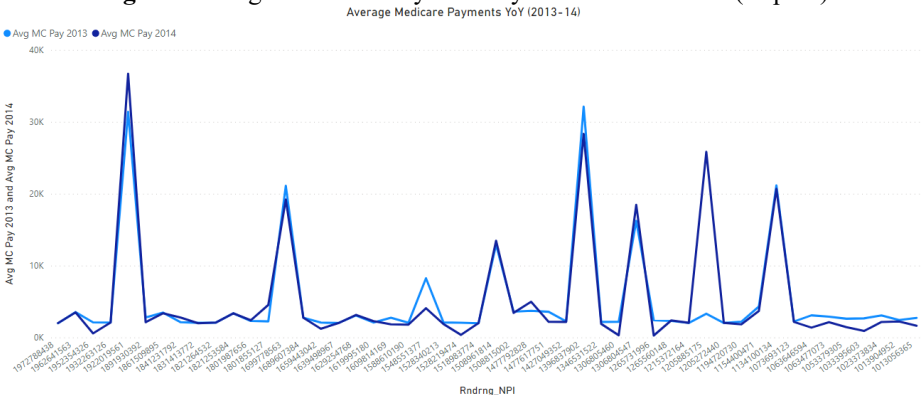


Fig 12. Average Submitted Amount from Provider YoY 2013-14

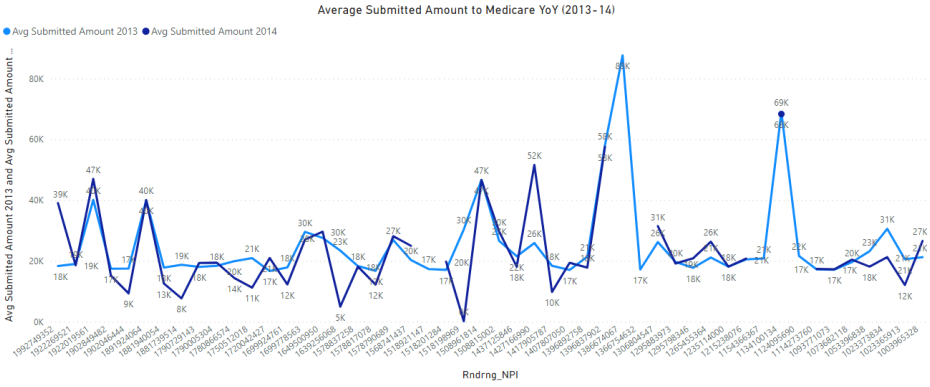
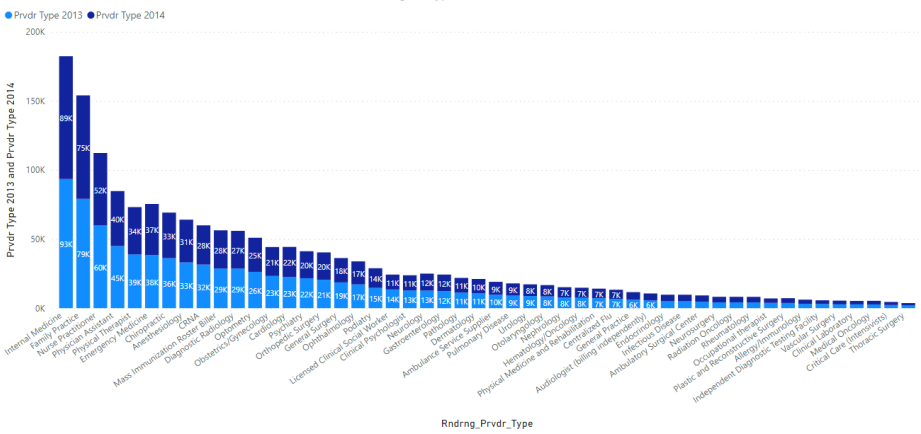


Fig 13. Change in Overall Provider Specialty Mix YoY 2013-2014



6 Discussion

Interpretations

Given the relative rarity of Medicare FWA, and its associated class imbalances, the results are reasonable. Furthermore, the first generation of the model utilizes some 35% of all claims data, further demonstrating the model's ability to detect fraudulent behaviors. Prior studies estimate FWA occurring at or around 10% of all transactions, and assuming the researchers' model is scaled to 100% of all claims data, the algorithm should classify around 7% of all transactions as FWA, safely within estimates of previous academic research. Classifying 7% of all Medicare claims as FWA would represent a change in thinking in CMS's ability flag potentially fraudulent claims. This research is important because even a small reduction in FWA in Medicare represents significant savings to the federal government.

Implications

Research into Medicare FWA is a pressing issue for the country, the amount of tax dollars sequestered due to these actions far exceeds the cost of mitigating them. By proving that detecting Medicare fraud, waste, and abuse is possible, the researchers have shown that CMS should be able to utilize tools such as this to help prevent continued losses into the future. An important implication of this study is that the researchers used machine learning and predictive analytics on previous years Medicare data and solely relies on the output from that model to determine if current claims can be determined to be fraudulent or not and could potentially flag non-fraudulent claims as fraudulent. The researchers believe that CMS could utilize a tool such as this and conservatively save some $50 to $100 billion (about $310 per person

in the US) a year that would normally be incurred due to loss attributed to fraud waste and abuse. Private healthcare companies as well as credit card companies have successfully employed these techniques for many years and do not incur the same levels of FWA as Medicare programs. A reasonable expectation is that CMS would employ some basic level of mitigation against FWA.

Like previous studies in this field, accurately matching claims data with proven fraud is a formidable task. The adversaries often produce new and novel means for evading detection such as running offshore operations and finding new medical practitioners to recruit into fraudulent activities. All the while, the officially released data is several years old at time of publishing and requires machine learning techniques to determine patterns of FWA in the data that may no longer be in use in the current year and into the future.

Ethical Considerations

Some ethical considerations include issues of redlining and ensuring only intentional fraud is actively pursued and prosecuted.

Within the publicly available claims data, location information related to these claims is readily available. Use of this data alongside the claims data could easily result in redlining, however this research did not make use of any location information and instead aggregated all the data at the national level, furthermore, all personally identifying information is scrubbed from the data at the CMS level. Potential issues that may arise from redlining could be bias toward PoC (People of Color) and other protected classes.

This research is primarily interested in bad actors and malicious providers seeking to exploit the Medicare system or otherwise extract money through the program against its actual intentions. Unfortunately, as Medicare is an extremely complex system, there are incidents where providers mistakenly misbill and commit Medicare FWA, the research team does not intend to pursue this type of behavior and feels that this behavior, while harmful doesn't typically warrant prosecution in the same way as intentional fraud. Any anomalous behavior found by the researchers' model whether this is an individual or entity are innocent until otherwise proven guilty.

Future Research

The researchers also considered using the Medicare Open General Payments Data set, but this dataset did not contain the correct features to be joined to the List of Excluded Individuals and Entities. These sets were also double and triple the size of the MPUPD. Ideally, In the future when computing resources are more readily available, an excellent expansion of this study would be to apply similar techniques to the MPUPD that were not included in this study.

Limitations and Challenges

This research, along with previous attempts at using Machine Learning to detect FWA suffers from the affliction of data. CMS data must be paired with LEIE data to determine rough estimates of fraudulent claims, this process is inherently error prone

and realistically touches only an exceedingly small proportion of the total available data.

As stated, claims data is only available for a period of years before being replaced with newer data, the risk posed here is that bad actors whose techniques were identified by the algorithm may no longer be employing the same tactics for committing Medicare FWA and the algorithm will break on future data. This is a classic game of "cat and mouse" with the central question centering around, "does past fraudulent behavior clearly predict future fraudulent behavior?"

It is important to note that the data contained in this research utilizes CMS claims data from 2013-2019, with the onset of the global COVID-19 pandemic, the state of healthcare has changed significantly. Chief among those changes has been the rapid adoption of tele-medicine, a subject that has little mention of in the current datasets. If this analysis were to be performed again with data from 2020 and beyond, the research team fully expects the results of this study to change significantly. A speculative hypothesis based on the current state of Medicare FWA would indicate an uptick in FWA based on the previous prevalence of FWA combined with the confusing nature of healthcare in a COVID-19 world.

7 Conclusion

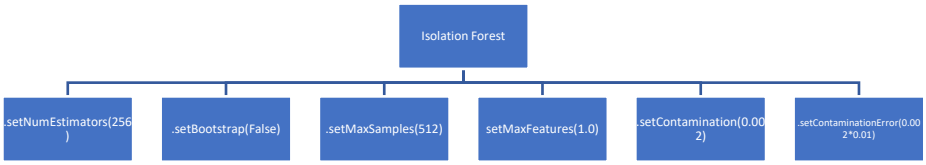
A clear need exists for CMS to put more resources into the detection and prevention of Medicare FWA, private insurers do not suffer from the same level of theft and provide a precedent for the ability to identify and stop this behavior with a key notable difference of questioning each claim prior to paying out. Due to the nature of the federal government, this approach is not possible with CMS. The publicly available data suffers from many afflictions including quality issues, mismatched claims information, and a serious lack of current information of providers on the LEIE. This research has shown, however, evaluating claims data to detect FWA is possible. Once the data has been thoroughly cleaned and merged, machine learning offers many opportunities to study patterns contained within Medicare data. At a minimum, an algorithm like the one developed by this research can drive insights into disseminating between legitimate Medicare transactions and ones that constitute Medicare FWA. An automated approach has the potential to assist human analysts comb through the millions of claims records and further investigate claims that are classified as anomalous. Chief among the benefits to an automated machine learning approach include cost savings, increased integrity, and a better public facing image for CMS.

8 Technical Appendix

The appendix is to describe and delve into further detail on certain technical aspects discussed in the paper.

Isolation Forest Architecture:

Figure 1: Isolation Forest Architecture



As per the diagram above, the Isolation Forest was fed the following parameters:
Number of estimators: 256
SetBootstrap: False
SetMaxSamples(512)
SetMaxFeatures(1.0)
SetContamination(0.002)
SetContaminationError(0.002*0.01)

Figure 2: Table Containing Anomalous and normal instances by year

	Year	prediction	count
0	2013	0	680656
1	2013	1	913
2	2014	0	700445
3	2014	1	1284
4	2015	0	723105
5	2015	1	1375
6	2016	1	1373
7	2016	0	749956
8	2017	0	775856
9	2017	1	1249

10	2018	1	2719
11	2018	0	798212

Label: 0 = Normal, 1 = Anomaly by Year.

References

0. R. Bauder and T. Khoshgoftaar, "A Survey of Medicare Data Processing and Integration for Fraud Detection," 2018 IEEE International Conference on Information Reuse and Integration (IRI), 2018, pp. 9-14, doi: 10.1109/IRI.2018.00010.
1. R. Bauder, R. da Rosa and T. Khoshgoftaar, "Identifying Medicare Provider Fraud with Unsupervised Machine Learning," 2018 IEEE International Conference on Information Reuse and Integration (IRI), 2018, pp. 285-292, doi: 10.1109/IRI.2018.00051.
2. Bauder, R. A., & Khoshgoftaar, T. M. (2018). The effects of varying class distribution on learner behavior for Medicare fraud detection with imbalanced big data. Health Information Science and Systems, 6(1), 1–14. https://doi.org/10.1007/s13755-018-0051-3
3. Bauder, R.A., & Khoshgoftaar, T.M. (2017). Medicare Fraud Detection Using Machine Learning Methods. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 858-865.
4. Buppert, C. (2001). Avoiding Medicare Fraud part 2. The Nurse Practitioner, 26(2), 34–41. https://doi.org/10.1097/00006205-200102000-00005
5. CMS. (n.d.). *Medicare provider utilization and payment data: Physician and other supplier*. CMS. Retrieved December 7, 2021, from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.
6. Ekina, T., Leva, F., Ruggeri, F., & Soyer, R. (2013). Application of Bayesian Methods in Detection of Healthcare Fraud.
7. Gad, B., Warren, C., & Paskhover, B. (2020). Why Otolaryngologists Get Excluded from Medicare and Medicaid. Ear, Nose, & Throat Journal, 14556132093304–145561320933040. https://doi.org/10.1177/0145561320933040
8. Gordon, D., & Siegel, D. M. (2020). Machine learning and the future of Medicare fraud detection. Journal of the American Academy of Dermatology, 83(2), e133. https://doi.org/10.1016/j.jaad.2020.03.059
9. J. Hancock and T. M. Khoshgoftaar, "Performance of CatBoost and XGBoost in Medicare Fraud Detection," 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), 2020, pp. 572-579, doi: 10.1109/ICMLA51294.2020.00095.
10. Herland, M., Bauder, R. A., & Khoshgoftaar, T. M. (2017). Medical provider specialty predictions for the detection of Anomalous Medicare Insurance claims. *2017 IEEE International Conference on Information Reuse and Integration (IRI)*. https://doi.org/10.1109/iri.2017.29

11. Lewinson, E. (2021, August 26). *Outlier detection with Isolation Forest*. Medium. Retrieved February 27, 2022, from https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e
12. Liu, J., Bier, E., Wilson, A., Guerra-Gomez, J. A., Honda, T., Sricharan, K., Gilpin, L., & Davies, D. (2016). Graph analysis for detecting fraud, waste, and abuse in healthcare data. *AI (Artificial Intelligence) Magazine*, *37*(2), 33–46. https://doi.org/10.1609/aimag.v37i2.2630
13. M. Johnson and Taghi M. Khoshgoftaar. Medicare fraud detection using neural networks. Journal of big data. https://doi.org/10.1186/s40537-019-0225-0. 2019. pp. 6–63.
14. Kose, Ilker & Gokturk, Mehmet & Kilic, Kemal. (2015). An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. Applied Soft Computing. 36. 283–299. 10.1016/j.asoc.2015.07.018.
15. Markus M Breunig, Hans Peter Kriegel, Raymond T Ng, and Jrg Sander. Lof: identifying density-based local outliers in ACM SIGMOD International Conference on Management of Data May 16-18, 2000, Dallas, Texas, Usa, pages 93-104, 2000.
16. Meyers, T. J. (2017). Examining the network components of a Medicare fraud scheme: the Mirzoyan-Terdjanian organization. Crime, Law, and Social Change, 68(1), 251–279. https://doi.org/10.1007/s10611-017-9689-z
17. Musal, R. M. (2010). Two models to investigate Medicare fraud within unsupervised databases. Expert Systems with Applications, 37(12), 8628–8633. https://doi.org/10.1016/j.eswa.2010.06.095
18. Nicholas, L. H., Segal, J., Hanson, C., Zhang, K., & Eisenberg, M. D. (2019). Medicare beneficiaries' exposure to fraud and abuse perpetrators. Health Affairs, 38(5), 788-793,793A-793C. doi:http://dx.doi.org/10.1377/hlthaff.2018.05149
19. Obodoekwe, N., & van der Haar, D. T. (2019). A comparison of machine learning methods applicable to healthcare claims fraud detection. *Advances in Intelligent Systems and Computing*, 548–557. https://doi.org/10.1007/978-3-030-11890-7_53
20. Pande, V., & Maas, W. (2013). Physician Medicare fraud: Characteristics and consequences. International Journal of Pharmaceutical and Healthcare Marketing, 7(1), 8-33.doi:http://dx.doi.org/10.1108/17506121311315391
21. Red Hat Marketplace. (n.d.). *Medicare provider utilization and payment data* . Medicare Provider Utilization and Payment Data - Physician and Other on Red Hat Marketplace - United States. Retrieved February 26, 2022, from https://marketplace.redhat.com/en-us/products/medicare-provider-utilization-and-payment-data-physician-and-other
22. Westerski, Kanagasabai, R., Shaham, E., Narayanan, A., Wong, J., & Singh, M. (2021). Explainable anomaly detection for procurement fraud identification—lessons from practical deployments. International Transactions in Operational Research, 28(6), 3276–3302. https://doi.org/10.1111/itor.12968
23. W. Zhang and X. He, "An Anomaly Detection Method for Medicare Fraud Detection," 2017 IEEE International Conference on Big Knowledge (ICBK), 2017, pp. 309-314, doi: 10.1109/ICBK.2017.47.
24. Medicare Learning Network (2021). “Medicare Fraud & Abuse: Prevent, Detect, Report.” ICN MLN4649244 January 2021, https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Downloads/Fraud-Abuse-MLN4649244.pdf. Accessed on: Dec. 4, 2021.