Health Care Management Science (2020) 23:2–19
https://doi.org/10.1007/s10729-018-9460-8

Approaches for identifying U.S. medicare fraud in provider claims data

Matthew Herland[1] · Richard A. Bauder[1] · Taghi M. Khoshgoftaar[1]

Received: 15 May 2018 / Accepted: 8 October 2018 / Published online: 27 October 2018

Abstract

Quality and affordable healthcare is an important aspect in people's lives, particularly as they age. The rising elderly population in the United States (U.S.), with increasing number of chronic diseases, implies continuing healthcare later in life and the need for programs, such as U.S. Medicare, to help with associated medical expenses. Unfortunately, due to healthcare fraud, these programs are being adversely affected draining resources and reducing quality and accessibility of necessary healthcare services. The detection of fraud is critical in being able to identify and, subsequently, stop these perpetrators. The application of machine learning methods and data mining strategies can be leveraged to improve current fraud detection processes and reduce the resources needed to find and investigate possible fraudulent activities. In this paper, we employ an approach to predict a physician's expected specialty based on the type and number of procedures performed. From this approach, we generate a baseline model, comparing Logistic Regression and Multinomial Naive Bayes, in order to test and assess several new approaches to improve the detection of U.S. Medicare Part B provider fraud. Our results indicate that our proposed improvement strategies (specialty grouping, class removal, and class isolation), applied to different medical specialties, have mixed results over the selected Logistic Regression baseline model's fraud detection performance. Through our work, we demonstrate that improvements to current detection methods can be effective in identifying potential fraud.

Keywords Medicare · Big data · Machine learning · Fraud detection

1 Introduction

Healthcare is a vital component in the daily life of most citizens in the United States. Even given healthcare's prominence in society, costs and premiums continue to increase placing additional strain on those in need of medical services. In particular, the generally increasing life expectancy as well as increasing population size, with the rise of chronic diseases, puts additional burdens on programs focused on the health of the elderly [1–3]. As such, it is imperative that healthcare costs are kept fair and reasonable for quality medical services [4]. Medicare is a U.S. government program that provides healthcare insurance and financial support for the elderly population, ages 65 and older, and other select groups of beneficiaries [5].

Within the Medicare program, each covered medical procedure is codified for claims and payment purposes. The basic claims process entails a physician performing one or more procedures and then submitting a claim to Medicare for payment, rather than directly billing the patient, thus Medicare being a type of "middle man" in this process. A claim is defined as a request for payment for benefits or services received by a beneficiary. In this paper, we use the terms providers and other entities [6] interchangeably with physicians, and note any specific differences as needed. Additional information on the Medicare claims process can be found in [7].

In order to keep healthcare affordable, programs need to keep medical-related costs low. One way to do this involves reducing fraud, waste, and abuse (FWA) in medical practices and claims. Malicious or wasteful activities can lead to higher costs and the possibility of patients going without necessary medical care. In particular, fraud is prevalent within healthcare with about 10% of all U.S. medical claims being fraudulent [8, 9]. The group, Coalition Against Insurance Fraud [10], provides statistics on fraud

✉ Richard A. Bauder
rbauder2014@fau.edu

Matthew Herland
mherlan1@fau.edu

Taghi M. Khoshgoftaar
khoshgof@fau.edu

[1] Florida Atlantic University, Boca Raton, FL, USA

and abuse found in the U.S. healthcare system. Some of the more salient statistics include the recovery of $29.4 billion to Medicare since 1997 by the Health Care Fraud and Abuse Control program, the exclusion of 1,662 individuals and entities from Medicare and Medicaid claims and payments, and a nearly five-fold increase in the recovery of proceeds (i.e. civil recoveries [11]). The Office of Inspector General (OIG) created the Medicare Fraud Strike Force [12] in a concerted effort to find and bring to justice fraudulent physicians. The Fraud Strike Force employs data analytics to successfully cut down on FWA. As of May 31, 2017, this strike force has recovered a total of $2.52 billion dollars through 1,791 criminal actions and 2,326 indictments. Even with these successful recoveries, fraud is still prevalent. Fraud is not restricted to any one particular medical field (specialty or provider type), such as Urology [13], but adversely affects every field. It is important to emphasize that fraud can be perpetrated by any provider in any field and as such, machine learning and data mining techniques can be leveraged to detect fraud for every provider type within healthcare. Furthermore, in the U.S. alone, the application of machine learning and data mining approaches has the potential to save the healthcare industry up to $450 billion each year [14].

As with our previous works [15, 16], we use data released by the Centers for Medicare and Medicaid Services (CMS) [17–19]. In contrast to our other works that use smaller subsets of Medicare data, we use the complete 2012 to 2015 *Medicare Provider Utilization and Payment Data: Physician and Other Supplier* dataset, also known as Medicare Part B, which includes provider Medicare claims information for the U.S. and its commonwealths. Because the Medicare Part B dataset does not include labels indicating provider fraud, we use the List of Excluded Individuals and Entities (LEIE) [20] database to generate fraud class labels (i.e. fraud or no fraud) for each provider to assess fraud detection capabilities of our baseline model and proposed improvement strategies. The LEIE contains all physicians who are excluded from practicing medicine for federally funded programs, such as Medicare. It is important to note that in our study, fraud detection is flagging possible or suspected fraudulent activities (based on known LEIE exclusions) and any non-fraud provider claims should be considered either non-confirmed fraud or indicative of exhibiting no fraud. We use fraud and non-fraud labels interchangeably with possible/suspected fraud and non-confirmed fraud designations. The effective detection of fraud can help to reduce costs related to time and resources needed for further investigation, by focusing the efforts on candidates most likely to exhibit fraudulent behaviors. Our baseline model, closely based on prior work [16], predicts the expected medical specialty of a physician based on the type and count of procedures performed. For example, if our model predicts a physician as a Dermatologist but the actual specialty is Optometrist, then this physician could be performing procedures indicating possible fraud. This difference in expected and actual specialties, for a particular provider, could indicate possible fraud to include possible upcoding [21] or coding incorrect procedures. It is important to note that issues related to medical coding do not necessarily imply fraud, but are still problematic in submitting, accepting, and paying on claims. A medical coder may not have enough information (e.g. poor documentation or lack of easy access to the provider) regarding the procedure to document the coding at the highest level of specificity, or may be using incorrect medical coding code sets[1].

In this study, our primary contributions include determining the best baseline model and assessing our proposed improvement strategies which include: class grouping, class removal, and class isolation. The baseline model is based on our prior research in predicting provider specialties but takes into account providers practicing in offices and/or facilities, such as a hospital, since services can be offered at one or both locations. Class grouping takes similar specialties and combines them into a single class, whereas class removal entails removing a selection of low scoring specialties based on two specified criteria. Class isolation is a different approach than the previous two improvement strategies that randomly sample a percentage of non-fraud class labels, retaining all fraud labels, per specialty, building models for each specialty to predict fraudulent providers. Overall, the class grouping and removal strategies had inconsistent results, with limited improvements and some cases of worsening fraud detection performance versus the baseline model. The class isolation method, however, indicated good improvements in overall fraud detection.

The rest of the paper is organized as follows. Section 2 discusses works related to current research in this area. Section 3 details the methodology used in this paper including the data, learners, performance metrics, and baseline model selection and improvement strategies. Section 4 discusses the results of our baseline model and strategies, as well as some possible research limitations. In Section 5, we conclude and present ideas for future work.

2 Related works

CMS has been releasing a given year's data on average two and a half years after the end of a particular year. Therefore,

[1] https://www.medicalbillingandcoding.org/common-problems-coding/

research employing this data is relatively new with no comprehensive studies leveraging all the publicly available Medicare data. All related works use varying subsets of the full Medicare dataset with most providing preliminary assessments and results, with or without machine learning approaches. Two related works that use Medicare data employing more typical data analysis to include descriptive statistics and regression are by Feldman et al. [22] and Ko et al. [13]. Using 2012 Medicare data, Feldman et al. try to determine if there are any correlation between a physician's schooling against the way he or she practices. They compared medical school charges, procedures, and payments for a given physician researching whether they could identify possible anomalies in the data. With this information the authors could flag physicians that are at risk of performing fraudulent activities at the beginning of their careers. Another study by Ko et al. focused on the field of Urology. The authors analyze variability among Urologists within the field's service utilization and payment data (2012 Medicare) to determine the estimated savings from a standardized service utilization. They determined there was a strong correlation between the number of patient visits with reimbursement received from Medicare. They establish that in the specialty of Urology alone there could potentially be a $125 million savings, or about 9% of the total expenditure within the field. Neither of these studies though employs more advanced analytics or machine learning to predict fraudulent providers or behaviors, nor do they leverage the full scope of available Medicare data.

The idea of looking for deviations from normal or expected patterns is part of a study by Bauder et al. [23]. The authors built a multivariate regression model for each Medicare specialty, such as Cardiology. From this model, the studentized residuals were generated and used as inputs into a Bayesian probability model in order to produce the probability of an instance being an outlier, which indicates the likelihood of fraud. Sadiq et al. [24] used the 2014 CMS dataset in order to find anomalies that possibly point to fraudulent or other interesting behavior. The framework they employ is the Patient Rule Induction Method based bump hunting method attempting to determine peak anomalies by spotting spaces of higher modes and masses within the dataset. They explained that by applying their framework they could characterize the attribute space of the data helping uncover the events provoking financial loss.

None of the previously discussed works, whether using descriptive statistics or machine learning, provide fraud detection performance validation using known fraudulent providers. In a general coverage paper by Chandola et al. [25], they used the CMS Medicare among other datasets to assess healthcare fraud using labeled data for fraudulent providers, primarily from the Texas Office of Inspector General's exclusion database. The authors employed several techniques including social network analysis, text mining, and temporal analysis. In particular, the authors discussed the use of typical treatment profiles, i.e. procedures performed. The idea is leveraging these profiles as the normal activity of physicians by comparing them with any provider in question which will determine possible issues or abuses in procedures. This is akin to predicting expected values and comparing these predictions to the associated actual values. In [26], Branting et al. use the 2012, 2013, and 2014 Medicare Part B and Part D data with the LEIE. They present a method for pinpointing fraudulent behavior by determining the fraud risk through the application of network algorithms from graphs. One algorithm, which they denote as Behavior-Vector Similarity, determines similarity in behavior for real-world fraudulent and non-fraudulent physicians using nominal values such as drug prescriptions and medical procedures. A group of algorithms makes up their Risk Propagation, which uses geospatial co-location (such as location of practice) in order to estimate the propagation of risk from fraudulent healthcare providers.

In our previous papers [15, 16], we experimented with whether or not it is possible to predict a physician's field of expertise based on only the procedures they perform. We found that there were a small percentage of specialties that could be accurately predicted. The overall goal of these papers and this current work is to better pinpoint fraudulent behavior. In essence, if we can predict a physician's specialty accurately (determined by F-score), then we could potentially find anomalous behaviors in a physician's procedures if they are predicted to be a specialty other than their own (e.g. Dermatologist as a Rheumatologist). The idea is that if a physician is predicted as having a specialty other than their actual specialty, they could be behaving in fraudulent, wasteful or abusive ways. The strategies we constructed were successful in increasing the prediction capabilities for many specialties, using 2013 Florida-only Medicare data. Our results showed we were able to detect 12 of the 18 known fraudulent physicians. These studies are limited in the location and amount of claims data used, and do not encompass the full scope of Medicare claims for the United States. Our previous works applied a different approach in trying to detect Medicare provider claims fraud, versus other related works, predicting a provider's specialty using the procedure type and utilization. We extend and improve upon our prior research considering procedures performed in an office and/or facility, and use data for the entire U.S. (not just Florida) over all available years. In this paper, we evaluate fraud detection performance of our proposed improvement strategies by leveraging the LEIE excluded providers as fraud or no-fraud labels.

Table 1 Sample of dataset used for this study

PROVIDER_TYPE	99222	99223	...	62311	64483
Internal Medicine	142	96	...	0	0
Pathology	0	0	...	0	0
Anesthesiology	0	0	...	56	16

3 Methodology

This section includes discussions on the datasets, learners, and metrics used in our study. Additionally, we detail the baseline model selection and the proposed improvement strategies. In particular, for the baseline model and class isolation method, we use different machine learning algorithms to build predictive models using the Part B data with LEIE fraud labels. These algorithms include Multinomial Naive Bayes, Logistic Regression, Support Vector Machines, and Random Forest.

3.1 Data

In our study, we use the Medicare Part B claims dataset which includes the 2012 to 2015 calendar years [19]. The Part B dataset provides information pertaining to the number of times each provider (or physician) billed a specific procedure within a given year, for medical claims only. We combined the four available years of Part B data into a single dataset. Table I summarizes the full Medicare Part B data for each year. Each physician is denoted by his or her National Provider Identifier (NPI) [27] and each procedure is denoted by a Healthcare Common Procedure Coding System (HCPCS) code [28]. The Part B dataset contains a number of features, such as the average amount billed and paid by Medicare for each physician and procedure per year. However, for this study, we are only interested in the provider's specialty, procedures performed, number of each procedure performed, and place of service (office or facility). Using only the procedure codes and associated count of the procedures performed, we transformed each physician entry into a vector where the class label for each instance is the physician's specialty and the features are the available procedures, identified by unique HCPCS codes. Therefore, the value of each feature and specialty is the number of times that provider billed Medicare for a specific procedure. This results in a sparse vector, since most physicians only use a small number of procedure codes as needed by their specialty. Table 1 shows a small example of the sparse matrix where each line is a physician, indicated by *provider_type* (i.e. specialty), with the remaining attributes (codes 99222 through 64482 in this example) being the procedures. For every instance in this sparse vector, there is a value for the number of times the given physician performed that procedure for that given year. Table 2 shows a small example of the sparse matrix where each line is a physician, indicated by provider_type (i.e. specialty), with the remaining attributes (codes 99222 through 64482 in this example) being the procedures. For every instance, there is a value for the number of times the given physician performed that procedure per year.

In order to validate fraud detection performance, we need labels indicating fraudulent provider claims. The Medicare Part B dataset does not include fraud labels; thus, we incorporate the List of Excluded Individuals and Entities (LEIE) database [20], which includes physicians who have been found to be in violation of one or more rules within Sections 1128 and 1156 of the Social Security Act [29]. The LEIE contains all current physicians who have been found unsuited to practice medicine and thus excluded from practicing in the United States for a given period of time. The Office of the Inspector General (OIG) is responsible for maintaining the LEIE, where the individuals on the exclusion list, under Section 1128, have convictions for crimes related to a healthcare program or patient abuse and neglect. This includes being convicted of a felony for fraud or the misuse of controlled substances, and are considered mandatory exclusions. After reviewing the violations under the aforementioned sections, we decided to only incorporate physicians with mandatory exclusions (Section 1128). Even though the LEIE database provides known provider-level exclusions, it is not a complete record of all known provider fraud, where 38% with fraud convictions continue

Table 2 Sample of dataset by year

PROVIDER_TYPE	2012	2013	2014	2015	Full
Number of Physicians	874,743	909,606	938,147	947,824	1,120,904
Number of Procedures	5,949	5,983	5,973	5,983	7,023
Provider Types (Specialties)	89	90	90	91	89

to practice medicine and 21% were not suspended from medical practice despite their convictions [30]. This lack of knowledge regarding all possible fraudulent providers could lead to predicting a provider as fraudulent when they are not, or vice versa, which may reduce the overall accuracy of a prediction model. Even so, fraud cases, like most criminal cases, are only known because those individuals were caught by law enforcement. There are many cases for which the perpetrators are never caught, thus we have no record of these activities. The exact size of annual theft is unknown and is the subject of debate, for which healthcare fraud likely costs tens of billions of dollars a year[2].

The LEIE database does not include NPI numbers for all physicians and after preliminary analysis, we found that combining first name, last name, and address is not 100% reliable in determining identity. Therefore, we used only those physicians with NPI numbers to identify matches for mapping fraud labels to the Part B data. We supplemented these 1,310 physicians with two other documented fraud cases, found during our previous research [16], giving us 1,312 fraudulent physicians. One important item to mention is that because physicians can be added to the LEIE at any time of the year and some instances in the exclusion dataset may not have complete years (e.g. if they were put on the LEIE on February 1, 2015, they would only have submitted procedures for January 2015), but we decided to retain all relevant and available instances. We generate two datasets (exclusion and non-exclusion) from the combined Medicare Part B data with the LEIE excluded physicians as fraud labels. If a physician is in the LEIE, they are put in the exclusion dataset, otherwise they are placed in the non-exclusion dataset. Each of these generated datasets are used in our experiments to build and test models, and validate fraud detection performance. Table 3 provides a high-level summary of each dataset.

Table 3 Dataset descriptions

Dataset	Description
Part B	– Released by CMS. – Provides claims information for each procedure a Physician/provider performs within a given year. – Oriented by: 1) NPI, 2) HCPCS, and 3) Place of Service
LEIE	– Released by OIG. – Contains physicians/providers that have committed real-world fraud.

3.2 Learners and performance metrics

We perform the data manipulations in order to generate each of the datasets using the R [31] and Python [32] programming languages. For building and testing our models, we use either Weka [33] or PySpark [34] depending on the improvement approach and the size of the dataset. The PySpark library in Python is used to interface with and leverage the capabilities of Spark, which is a unified analytics engine for large-scale data processing [35]. Weka is an application providing a suite of machine learning algorithms that can be applied either via a graphical interface or the command line, and is suitable for smaller datasets. More specifically, for the class grouping and removal strategies, we use PySpark for building models due to the large size of the full Medicare Part B dataset. The class isolation method, however, generates much smaller datasets and we can leverage the larger variety of tools for machine learning in Weka. In this study, we use four machine learning models: Multinomial Naive Bayes (MNB), Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF). Additional details on each learner can be found in Appendix A.

In order to provide a robust experiment on fraud detection, we use several different metrics to assess model and improvement strategy performance. We leverage Recall, Precision, F-score, G-measure, and overall accuracy to assess model performance, with F-score being the primary metric for general performance comparisons. We also use the so-called Inverse Overall Accuracy (IOA), for each specialty, and the overall weighted average IOA (owaIOA). For our study, we are interested in the incorrectly classified (as we consider these fraud) which is given by the IOA and owaIOA, which is normalized by the number of fraudulent instances for each specialty. Additionally, for the class isolation method, we use Type I (false positive) and Type II (false negative) error rates to assess the predictive capabilities of the models for detecting Medicare Part B fraud. More information on performance metrics used in our study can be found in Appendix B.

3.3 Baseline model selection and improvement strategies

First, we determine the best model to be used as our baseline model for assessing the change in fraud detection performance using the class grouping and removal strategies. Recall that the baseline model is based on our prior research in producing models that predict provider specialties to assess possible fraudulent behaviors. There are a couple of requirements for selecting a baseline model: 1) able to support multi-class classification, and 2) able to output the precision, recall, F-score and overall accuracy in order to make fair comparisons. Due to the large data size

[2] https://www.insurancefraud.org/statistics.htm#13

and through preliminary analysis, we found the PySpark versions of MNB and LR met these criteria.

The methods for handling different locations, office or facility, from which physicians perform medical procedures is also tested in order to get the best overall baseline model using either office- or facility-based procedures. The problem with having different places of service is that a physician can bill the same procedure in either an office or facility. To account for these possible multiple entries, we devised two methods of handling this by either Combining Office and Facility (COF) or Separating Office and Facility (SOF). COF is where each procedure performed by each physician, per year, is summed, whether they billed from an office or facility. For example, in a given year, physician X performed code A 100 times in their office and 500 times in a facility, such as a hospital, thus these counts would be summed together with physician X performing code A 600 times. The second method, SOF, is where the codes are treated as separate codes depending on where they were administered. Using the same example with physician X, for a given year, code A would be transformed into two different codes such that code A would be codified as code AO, being done 100 times, and code AF, being done 500 times. Therefore, COF leaves the data with the original number of procedure codes, while SOF increases the number of procedures as seen in Table 4.

The best performing PySpark model using either COF or SOF is to be used as the baseline model for the class grouping and removal strategies. Additionally, the COF or SOF method which gives better performance results is also used with the Weka models for the class isolation method. In order to select the best baseline model and office or facility method, we use Precision, Recall, F-score, G-measure, and overall accuracy as metrics for performance evaluation. For this experiment, we use the non-exclusion dataset for training and the exclusion dataset for testing, in order to validate the models and assess the detection of fraud instances. When testing the models, the exclusion dataset is split into a number of smaller datasets, one for each physician type (specialty), and each of these datasets is used as the test set. We use IOA to assess performance with the best models having the highest IOA per specialty. The baseline model is thus the best combination of MNB or LR and COF or SOF. With this baseline model, we experiment with the class grouping and removal strategies, as well as the combination of both strategies. Lastly, the class isolation method also compares performance against that of the baseline model.

Table 4 Sample of full dataset

PROVIDER_TYPE	COF	SOF
Number of Physicians	1,120,904	1,120,904
Number of Procedures	7,023	10,029
Provider Types (Specialties)	89	89

The class grouping strategy groups similar specialties into a single group to reduce redundant specialties and decrease model variance. This is based on research in our previous work [16], where we found that specialties that practice on similar parts of the body or have similar medical descriptions provide improvements to prediction results. The creation of the groupings, in this study as in [16], are done manually on a specialty-by-specialty basis. Note that because these groups are manually formed, there could be other groupings, but our goal is to demonstrate the effectiveness of grouping for improving prediction and fraud detection performance and not specifically the formation of groups. For future work, we will consider different clustering approaches. These groupings are generated based on the assumption that if the specialties are similar then they share a significant number of HCPCS codes, thus indicating overlapping procedures. Classes such as Internal Medicine, which overlap with many classes and can be confused for other classes, were not considered due to the large group generated and low prediction performance. A very large group that consists of many classes would defeat the purpose of class grouping, as we want to find small groups of specialties where every class within this group is similar to each other in practice. An example of a reasonable grouping involves Ophthalmology and Optometry which both provide medical procedures focused on the eyes and thus makes sense as a group. With this in mind, we decided to test fourteen different groups as shown in Fig. 1 under the *Group Name* column. In order to evaluate any performance improvements via class grouping, we take a two stage approach. The first stage consists of creating datasets for each of the 14 groupings. We evaluated the performance of each group, using Recall, Precision, F-score, G-measure, and accuracy, versus the individual results for each of the members of the group, using the baseline model. In the second stage, we choose the groups that showed improvement over their individual members (per specialty) and conduct two tests determining improvement in fraud detection: 1) for each group individually and 2) all together (combined groups dataset).

The class removal strategy removes certain specialties from the dataset prior to building a model. We have two different sets of classes (specialties) for removal to test model performance. The first, referred to as the 'original four classes', is from our previous work and was based on unique procedures that have both a high number of instances and poor classification performance. These four classes were chosen for removal from the non-exclusion and exclusions datasets and include the following: Family Practice, Nurse Practitioner, Internal Medicine, and

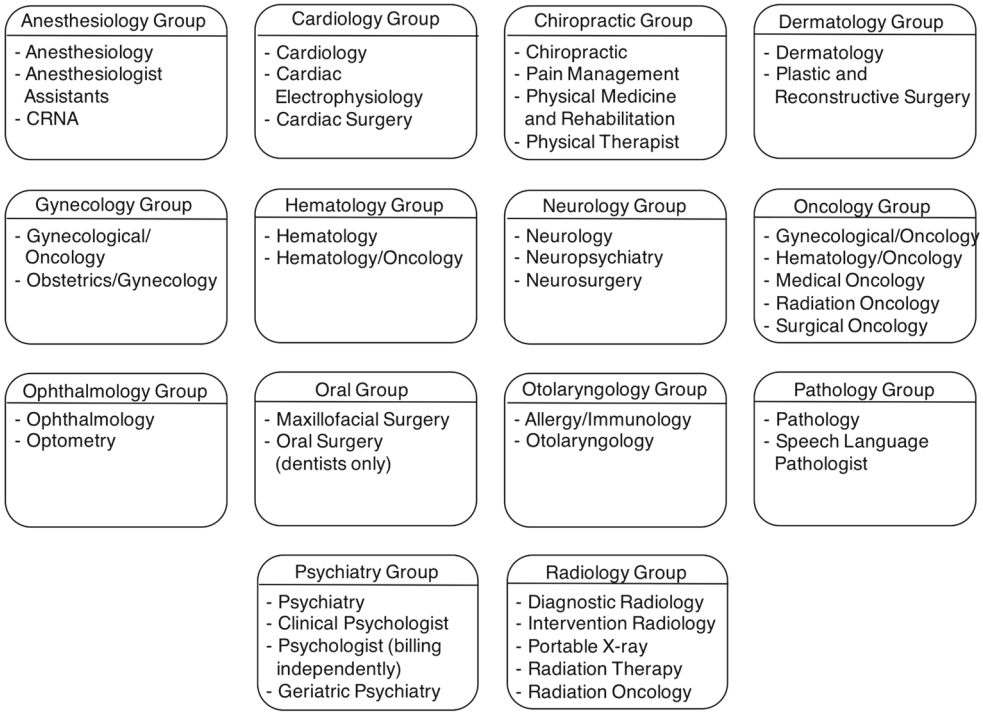

Fig. 1 List of Groupings and Group Members

Physician Assistant. The choice to remove these four classes was determined by reviewing our prior work's confusion matrix[3] of the 2013 Florida data and confirming that all removed classes do indeed cause a relatively large number of misclassifications. The second criteria for removal includes those removed via the first criteria plus specialties with low scores (precision and recall) and containing the words 'medicine', 'general', or 'unknown' (i.e. Unknown Supplier/Provider), indicating less specific practices (e.g. family practice) or ambiguous and less defined specialties (via the 'unknown' word). Table 5 shows the specialties for both criteria of class removal.

The class isolation method improvement strategy differs from the other two approaches by employing data sampling and adjusting the costs associated with Type I and Type II errors. With this method, we split the Medicare Part B dataset and build fraud detection models for each specialty. In the previous strategies, as with the baseline model, the result is a prediction of the physician's specialty from which we can validate fraud detection using the LEIE database. For these methods, to assess a correct prediction, the predicted specialty is compared to the actual specialty and if they differ, this could indicate possible fraud. This possible fraud result is compared to the LEIE to determine whether this prediction captured an actual fraudulent provider or not. For class isolation, however, the model results are not the physician's specialty but rather 'fraud' or 'non-fraud' derived from the mapped LEIE fraud labels. In order to provide both fraud and non-fraud labels, the exclusion and non-exclusion datasets are combined. For our study, we chose the following specialties, from the original dataset, based on having 50 or more available LEIE fraud labels: Chiropractic, Family Practice, General Practice, Internal Medicine, Physician Assistant, and Psychiatry. For each of these specialties, the class isolation method uses the SOF subset of data to build models, as chosen in the baseline

Table 5 Specialties used in the class removal strategy

Criteria 1: Four classes	Criteria 2: Chosen classes
Class removal strategy	
Family Practice	Certified Clinical Nurse Specialist
Nurse Practitioner	General Surgery
Internal Medicine	General Practice
Physician Assistant	All Other Suppliers
	Unknown Physician Specialty Code
	Unknown Supplier/Provider
	Nuclear Medicine
	Osteopathic Manipulative Medicine
	Sports Medicine
	Geriatric Medicine
	Preventative Medicine
	Addiction Medicine
	Pediatric Medicine

[3] A tabular representation comparing the predicted class membership against the actual class membership for each instance present in the dataset, denoting true positives, true negatives, false positives, and false negatives.

model selection process, and can be compared to the other improvement strategies via the F-score measure.

For the first class isolation approach, we employ Random Under Sampling (RUS) with a 3:1 ratio (class distribution), or 75% non-fraud and 25% fraud instances. RUS removes samples from the majority class (non-fraud), while keeping all of the minority class (fraud) observations. For example, the Chiropractic specialty had 86 excluded physician instances, which are retained, with 258 non-excluded Chiropractic instances chosen at random from the entire dataset. We decided to use a 3:1 ratio in order to retain a larger amount of non-fraud instances, thus retaining more information relative to the fraud instances. Based on our previous research in applying machine learning techniques with varying levels of imbalanced datasets [36, 37], we found that RUS is an effective method for mitigating the adverse effects associated with class imbalance, including higher levels of imbalance (i.e. a low number of fraud labels relative to non-fraud labels) such as in our study. Note that the goal in our study is not to find the best class distribution, but to determine the effectiveness of this particular class isolation approach in Medicare fraud detection. In order to reduce bias due to lucky or unlucky draws in the RUS process, we generated ten different datasets (RUS repeats) for the six specialties, each with the 3:1 ratio. For each of these ten datasets, we built each model (MNB, LR, RF, and SVM), per specialty, using 10-fold cross-validation. The performance results were averaged over the 10 repeats for the final model evaluation. Type I and II error rates are used as the primary metrics, while F-score is used to validate the overall fraud detection performance. The second class isolation approach uses a cost-sensitive classifier [38]. This classifier is used to find the model with the lowest combination of Type I and II error rates, while minimizing the Type II error rate. To do this, only the cost of a Type II error is varied, which changes the ratio between Type I and II errors. In our case, the Type II error is more important as it can indicate money lost due to fraudulent activities, thus we increase the cost associated with this error to help the model better discriminate fraud instances.

4 Results and discussion

In this section, we present the fraud detection results for the baseline model and each of the improvement strategies. For ease of presentation, we organize this section into three parts corresponding to the experimental flow and design as outlined in Section 3. The three parts include: 1) baseline model selection with SOF and COF, 2) class grouping and removal strategies with combinations, and 3) class isolation method. For each of these parts, we present the salient results with discussions pertaining to the Medicare Part B fraud detection performance and validity. Figure 2 depicts a high-level flow of these parts with inputs, descriptions, and possible output behaviors. Note that these output behaviors are examples of a possible predicted specialty versus actual specialty.

4.1 Baseline model selection with place of service

During the baseline selection process, we select the best model using the SOF or COF configurations[4] between MNB and LR. Table 6 shows the average scores for each performance metric by learner and SOF/COF method. These results indicate that LR outscores MNB for all but recall. Since LR, in general, is shown to perform better than MNB, we focus on the SOF and COF configuration results for LR only for which SOF has higher scores (possibly due to the larger number of features to draw from). Given these results, we selected LR and SOF as the baseline model for the class grouping and removal strategies, and the SOF configuration for the class isolation method.

We evaluate the baseline model's fraud detection performance for each specialty (IOA) and overall (owaIOA). Table 7 shows the results for the baseline model, per specialty, with F-scores at or above 0.75 indicating good detection performance. In our previous work [16], we group different scores into performance groups and from these, we selected the 0.75 minimum threshold indicating moderate to high prediction performance. Note that only specialties with moderate to high prediction performance can accurately determine our model's ability to detect fraud given our hypothesis of "if mislabeled, then fraudulent", and thus only tested on these specialties. In order to summarize the by-specialty results into a single metric representing the overall fraud detection performance of the model, we calculate the owaIOA over the IOA values seen in Table 7. The owaIOA for the baseline model, which is the model created to compare our improvement strategies, is 0.231 indicating 23.1% of the instances were correctly labeled as fraud.

4.2 Class grouping and removal strategies

Given the baseline model, in this section, we present the results of the class grouping and removal strategies and discuss any areas of improvement to this baseline. Unfortunately, both strategies produce mixed results and do not demonstrate general improvements on the baseline model. In Table 8, we present the results for the class grouping strategy. In this table, the Group Name is the name we denoted to this selected grouping and the Group Members are the specialties that compose this grouping, with changes in performance shown in the 'Overall Accuracy

[4] Separate Office or Facility or Combined Office or Facility.

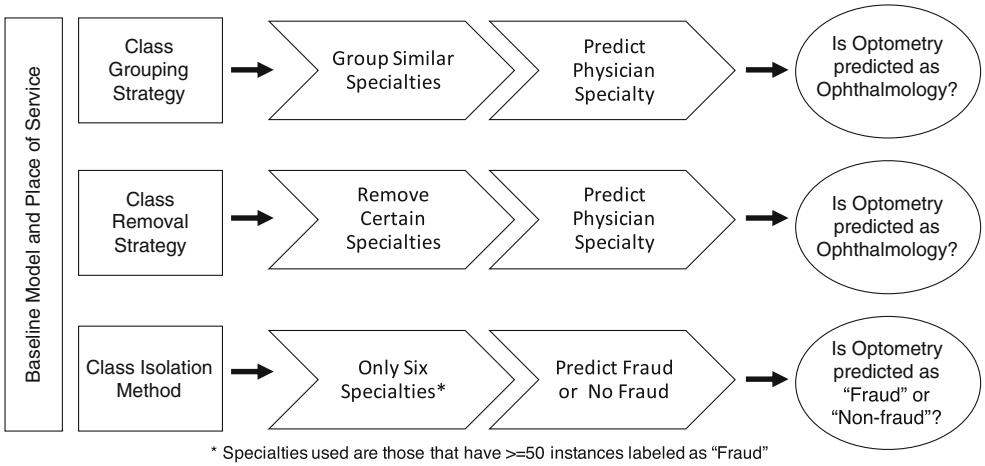

Fig. 2 Improvement Approaches

(Improvement)' column. The results are heavily influenced by the physician specialties in each of the groupings, with some groups having better performing individual specialties, thus better group performance and vice versa. Therefore, the class grouping strategy could be effective with some classes and boost fraud detection performance. In general, we do not discriminate between different provider types (specialties) with the assumption that each provider (with a specific provider type) has legitimate claims (payment and utilization information). Furthermore, each class, or specialty, can be tied back to an NPI for a particular claim for further investigations.

There were improvements from the baseline model compared to individual members within each group, with good group improvements seen for Anesthesiology, Cardiology, Gynecology, Ophthalmology, Oral, and Otolaryngology. Each of these groups showed good overall improvements, the F-scores are considerably higher than most of the individual member's results indicating significant improvements due to class grouping for these specialties. These particular specialties show promise employing a grouping strategy. The Chiropractic, Dermatology, Hematology, Neurology, Psychiatry, and Radiology groups showed moderate improvements from the class grouping strategy. These groups had members with both high and low F-scores, with the group score being either slightly above or slightly below the groups highest scoring member. The remaining groups, Pathology and Oncology, showed no useful improvements. The Pathology group F-score was quite high at 0.95, but the individual members had F-scores of 0.95 and 0.97, thus no improvement was found via grouping. The Oncology group had good results, but this group has a lot of overlapping specialties. Specifically, the Oncology group overlaps classes with the Gynecological, Hematology, and Radiology

Table 6 Average metric scores

	Naïve bayes		Logistic regression	
	COF	SOF	COF	SOF
Precision	0.442	0.447	0.512	0.514
Recall	0.546	0.550	0.499	0.510
F-score	0.407	0.410	0.487	0.497
G-Measure	0.436	0.440	0.495	0.504
Accuracy	0.517	0.520	0.679	0.682

Table 7 Baseline fraudulent detection results

PROVIDER_TYPE	F-score	Instances	IOA
Ambulance Service Supplier	1.00	2	0.00
Chiropractic	1.00	86	0.00
Audiologist (billing independently)	0.99	8	0.00
Physical Therapist	0.99	17	0.00
Speech Language Pathologist	0.97	3	0.33
Pathology	0.95	4	0.00
Diagnostic Radiology	0.94	4	0.25
Occupational therapist	0.92	6	0.17
Podiatry	0.90	47	0.09
Urology	0.90	13	0.23
Gastroenterology	0.88	7	0.29
Cardiology	0.85	13	0.46
Ophthalmology	0.84	20	0.25
Optometry	0.84	8	0.25
Dermatology	0.83	11	0.10
Otolaryngology	0.82	13	0.62
Emergency Medicine	0.81	32	0.53
Obstetrics/ Gynecology	0.80	38	0.58
Allergy/Immunology	0.79	4	0.00
Independent Diagnostic Testing Facility	0.78	5	0.00
Nephrology	0.78	1	0.00
CRNA	0.76	12	0.17
Orthopedic Surgery	0.76	11	0.82

Table 8 Grouping strategy performance results

Group name	Group members	Recall	Precision	F-score	G-measure	Overall accuracy (Improvement)
Anesthesiology	Anesthesiology	0.75	0.65	0.70	0.70	0.701 (0.19)
	Anesthesiologist Assistants	0.05	0.01	0.02	0.02	
	CRNA	0.71	0.82	0.76	0.76	
	Grouped Values	0.98	0.97	0.97	0.97	
Cardiology	Cardiology	0.86	0.85	0.85	0.86	0.683 (0.01)
	Cardiac Electrophysiology	0.48	0.52	0.50	0.50	
	Cardiac Surgery	0.39	0.34	0.36	0.36	
	Grouped Values	0.86	0.86	0.86	0.86	
Chiropractic	Chiropractic	1.00	1.00	1.00	1.00	0.682 (0.00)
	Pain Management	0.24	0.15	0.18	0.19	
	Physical Medicine and Rehabilitation	0.50	0.48	0.48	0.49	
	Physical Therapist	0.98	0.99	0.99	0.99	
	Grouped Values	0.94	0.95	0.94	0.94	
Dermatology	Dermatology	0.78	0.88	0.83	0.83	0.682 (0.00)
	Plastic and Reconstructive Surgery	0.49	0.39	0.42	0.43	
	Grouped Values	0.76	0.75	0.75	0.75	
Gynecology	Gynecological/ Oncology	0.30	0.32	0.31	0.31	0.683 (0.01)
	Obstetrics/ Gynecology	0.86	0.76	0.80	0.81	
	Grouped Value	0.87	0.76	0.81	0.81	
Hematology	Hematology	0.03	0.03	0.03	0.03	0.682 (0.00)
	Hematology/ Oncology	0.46	0.62	0.53	0.53	
	Grouped Values	0.44	0.67	0.53	0.54	
Neurology	Neurology	0.69	0.71	0.70	0.70	0.682 (0.00)
	Neuropsychiatry	0.02	0.01	0.01	0.01	
	Neurosurgery	0.50	0.57	0.53	0.53	
	Grouped Values	0.61	0.71	0.66	0.66	
Oncology	Gynecological/ Oncology	0.30	0.32	0.31	0.31	0.684 (0.02)
	Hematology/ Oncology	0.46	0.62	0.53	0.53	
	Medical Oncology	0.22	0.12	0.15	0.17	
	Radiation Oncology	0.96	0.94	0.95	0.95	
	Surgical Oncology	0.34	0.16	0.21	0.23	
	Grouped Values	0.62	0.76	0.68	0.69	
Ophthalmology	Ophthalmology	0.85	0.83	0.84	0.84	0.688 (0.06)
	Optometry	0.85	0.84	0.84	0.84	
	Grouped Values	0.97	0.95	0.96	0.96	
Oral	Maxillofacial Surgery	0.48	0.14	0.21	0.26	0.682 (0.00)
	Oral Surgery (dentists only)	0.56	0.40	0.47	0.47	
	Grouped Values	0.69	0.64	0.67	0.67	
Otolaryngology	Allergy/ Immunology	0.79	0.79	0.79	0.79	0.682 (0.00)
	Otolaryngology	0.76	0.89	0.82	0.82	
	Grouped Values	0.86	0.87	0.87	0.87	
Pathology	Pathology	0.96	0.93	0.95	0.95	0.682 (0.00)
	Speech Language Pathologist	0.98	0.96	0.97	0.97	
	Grouped Values	0.96	0.93	0.95	0.95	
Psychiatry	Psychiatry	0.59	0.66	0.62	0.62	0.683 (0.01)
	Clinical Psychologist	0.62	0.51	0.56	0.56	
	Psychologist (billing independently)	0.08	0.00	0.00	0.01	

Table 8 (continued)

Group name	Group members	Recall	Precision	F-score	G-measure	Overall accuracy (Improvement)
	Geriatric Psychiatry	0.00	0.00	0.00	0.00	
	Grouped Values	0.65	0.60	0.62	0.62	
Radiology	Diagnostic Radiology	0.94	0.94	0.94	0.94	0.683 (0.01)
	Interventional Radiology	0.29	0.18	0.22	0.23	
	Portable X-ray	0.90	0.93	0.92	0.92	
	Radiation Therapy	0.71	0.84	0.77	0.77	
	Radiation Oncology	0.96	0.94	0.95	0.95	
	Grouped Values	0.96	0.96	0.96	0.96	

groups. But by removing the overlapping groups, we would be removing the three highest scoring classes within this group leaving only the two lowest scoring specialties (Medical Oncology and Surgical Oncology), thus no notable improvements. Even with improved F-scores, class grouping does not seem to improve overall detection results, as indicated by the overall accuracy. Only two groups show good improvement, Anesthesiology and Ophthalmology, with improvements over the baseline model of 0.19 and 0.06, respectively. Otherwise, the accuracy remained the same or showed negligible improvement. Therefore, the improvements shown by only grouping may not actually increase the overall effectiveness of our baseline model.

As with the baseline model, the class grouping strategy's fraud detection performance is assessed with the IOA and owaIOA metrics. To better evaluate possible improvements made by this strategy, we look at two different grouping tests. The first group test is used to assess performance of each group separately, keeping those groups exhibiting improvements and testing only those with an F-score at or above 0.75. These results are shown in Table 15 in Appendix C, where the number of total number of instances is calculated by taking the sum of instances from each group member. There are 293 instances between these groups with an owaIOA across the eight groups of 25.2% (a 2% increase over the baseline model). In general, some groups show noticeable improvements whereas others do not. The Cardiology and Chiropractic groups showed improvements due to class grouping, demonstrating that this strategy can improve fraud detection for certain groupings of specialties and groups. With that said, there were two groups, Otolaryngology and Ophthalmology, that showed noticeable decreases in performance because of class grouping. One possible reason for the negative results after grouping could be that the members within these groups are extremely similar and, with the baseline model, these classes were being labeled as another class within the group.

The second test, with results shown in Table 16 in Appendix C, contains all specialties and groups with an F-score at or above 0.75. This includes the groups from the group one combinations plus the remaining specialties not included in a group. There are 444 instances across these specialties with an owaIOA of 26.1%. Thus, there is an increase of 3% over the baseline model using all the groups that had individually good or mediocre results. In particular, two specialties showed improvement for this grouping: Independent Diagnostic Testing Facility and Orthopedic Surgery. Otolaryngology showed an improvement in IOA compared to when grouping was done individually. Conversely, the Chiropractic group had a slight decrease in IOA. Based on our results for class grouping, in order to detect fraudulent behavior for specialties within groups, these specialties would need to be contained within a group (e.g. radiation therapy should be grouped in the Radiation group). The exception could be the Otolaryngology group. The reason is that when using only the group containing the specialty of the physician in question, there are fewer procedure codes available with which to be confused. Future research would involve improved grouping methods to maximize procedure code similarities among grouped specialties. The best situation, for class grouping, is to retain the largest number of specialties with the best fraudulent detection results.

Table 9 presents the averaged results based on the removal of specific classes for each of the criteria[5] compared to the baseline model. Using the original four classes criteria, we notice that there is decent improvement in the overall performance, especially regarding accuracy which was improved by 0.136 over the baseline model. Further investigation into these results show the baseline model had 15 classes with an F-score greater than 0.90, with the original four class criteria improving the baseline result to 21 classes having over a 0.90 F-score. The chosen classes criteria demonstrates

[5]Original four classes are from our previous work and based on unique procedures that have both a high number of instances and poor classification performance, and Chosen classes removes the original four classes plus twelve additional specialties.

Table 9 Class removal results and improvements by criteria

	Original four	Chosen	Original four improvement	Chosen improvement
Precision	0.572	0.641	0.058	0.127
Recall	0.546	0.623	0.036	0.113
F-score	0.544	0.621	0.047	0.124
G-measure	0.551	0.626	0.047	0.122
Accuracy	0.818	0.843	0.136	0.161

a larger performance improvement over the removal of the original four classes only. This improvement is most likely due to fact that some of the classes removed had very low individual performance scores. Thus, these low-scoring removed classes had no effect on the remaining high-scoring classes. Even with the demonstrated improvements using the chosen classes criteria, there was no noticeable improvement beyond the original four classes criteria. The classes with relatively large improvements most likely have procedures that are easily confused with those in the removed classes, whereas the classes with little to no improvement indicate that more specialized services are not well represented in the removed classes.

The detailed performance results for the original four classes and chosen classes removal criteria are listed in Tables 17 and 18, found in Appendix C. Similar to the previously discussed class grouping strategy, we only test specialties with an F-score 0.75 or above for evaluating fraud detection results. The baseline model results included 365 instances with an owaIOA of 23.1%, with the removal of the original four classes having 378 instances and a lower owaIOA of 15.9% and the chosen classes criteria, with 397 instances, having a 14.1% owaIOA. For both criteria, we found the IOA for each individual specialty was either decreased or stayed the same when compared to the baseline model. The reason that fraud detection performance decreased using the class removal strategy is because class confusion, related to each specialty's procedures performed, is removed from the dataset giving less specialties for classification. In other words, the only reason specialties were mislabeled was due to the confusion of the low scoring fields and not because these specialties were actually performing as any of the removed classes. Therefore, given our need to predict a physician's specialty, the class removal strategy does not provide any meaningful improvements.

In order to understand the impacts for class removal and grouping, we leverage the best results from both and evaluate the fraud detection performance. Table 10 shows the overall performance when mixing the combined groups with the original four and the chosen classes removal criteria. We notice that for both there is a greater prediction improvement over the summation of each individual strategy. However, the owaIOA for the combined groups and original four classes is 20.8% with 457 instances, and the combined groups and chosen classes owaIOA is 21.1% with 484 instances. Both results indicate a decrease in performance when compared to the baseline model in terms of owaIOA. Thus, mixing class grouping and removal strategies does not have a positive performance impact.

Table 10 Combined removal and grouping strategy performance results

	Combined averages		Improvement	
	Original four	Chosen	Original Four	Chosen
Precision	0.599	0.706	0.085	0.192
Recall	0.579	0.693	0.069	0.183
F-score	0.573	0.688	0.076	0.192
G-Measure	0.580	0.694	0.092	0.19
Accuracy	0.864	0.891	0.182	0.209

4.3 Class isolation method

When employing the class isolation method, we build a model per physician specialty (Chiropractic, Family Practice, General Practice, Internal Medicine, Physician Assistant, and Psychiatry), each with a 3:1 RUS class distribution. Table 11 focuses on the results employing data sampling showing the Type I and Type II error rates for all six classes (specialties) separated by learners, with the best performers denoted in boldface. Family Practice using MNB is the only class that fits our previously mentioned goal of balancing error rates while minimizing Type II error, with a Type I error rate of 0.284 and Type II error rate of 0.256. No other learner was able to create a comparable model.

In addition to using data sampling, via RUS, to improve fraud detection performance, we use a cost sensitive classifier[6]. The purpose of a cost sensitive classifier is to determine the optimal model by finding the best learner and cost ratio combination for each specialty. Again, we want to have the lowest possible Type I and Type II error rates, thus minimizing both while treating the Type II error as the more important metric. In order to find the best model for each specialty, we first determined the intersection of the Type I and Type II errors for each of the four tested models. We found the intersections for each learner and

[6] A model built by adjusting the costs associated with Type I and Type II errors.

Table 11 Summary of class isolation results with data sampling

Provider type	MNB		LR		RF		SVM	
	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
Chiropractic	**0.363**	**0.639**	0.022	0.914	0.174	0.696	0.052	0.992
Family Practice	**0.284**	**0.256**	0.235	0.491	0.174	0.696	0.032	0.759
General Practice	0.137	0.601	**0.339**	**0.583**	0.183	0.520	0.059	0.747
Internal Medicine	0.056	0.569	0.196	0.444	0.101	0.589	**0.351**	**0.667**
Physician Assistant	**0.235**	**0.349**	0.250	0.498	0.052	0.689	0.059	0.633
Psychiatry	**0.220**	**0.485**	0.133	0.649	0.078	0.717	0.023	0.847

specialty by randomly choosing one of the ten datasets for a given class. The variation between the ten datasets for each learner was small, thus this process is considered reliable. We found that Random Forest had the best results for all six specialties, and therefore, we use RF to find the optimal cost ratio. After determining where the Type I and Type II error rates intersected using a single dataset from a given specialty, we ran that same cost ratio over all ten datasets, checking for an optimal model. If that particular cost ratio was found to not produce an optimal model, we either increased or decreased the cost for the Type II error only until we reached the optimal cost. Table 12 presents the final cost ratios and the F-scores for the optimal models, per specialty. The results for Chiropractic, Physician Assistant, and Psychiatry were similar with error rates around 0.300 and 0.400. Family Practice and General Practice had better results with error rates around 0.200. Internal Medicine had very good results with the lowest error rates. Overall the results of the class isolation method demonstrate strong fraud detection performance. It is particularly notable when compared to the F-score results of either the class removal or grouping strategies.

4.4 Summary of improvements

A summary of the fraud detection performance results comparing each strategy and the baseline model is shown in Tables 13 and 14. The results shown in Table 13 are for the class grouping and removal strategies with baseline model comparisons only. Table 14 summarizes the class isolation method results. With regards to improving performance, the class grouping strategy has produced results indicating improved detection. In class grouping, there were some large improvements over individual specialty results, such as the Ophthalmology and Optometry, and Anesthesiology and Anesthesiologist Assistants groups, while the other groups had moderate to minimal improvements. The high scoring groups, with only the group in the dataset, had an owaIOA of 25.2% that was slightly higher than the baseline model result. When the eight specialties were in the dataset for combined class grouping, which combined groups with moderate to high improvements, provided similar results to grouping separately and produced minimal overall improvement, with a 26.1% owaIOA. The class removal strategy, with the two criteria, showed a significant decline in

Table 12 Class isolation cost sensitive classifier results with RF

Provider Type	Cost		Error		F-Score	
	Non-fraud	Fraud	Type I	Type II	Non-fraud	Fraud
Chiropractic	1	8	0.397	0.405		
	1	8.3	0.403	0.394	0.69	0.43
Family Practice	1	8	0.212	0.216		
	1	8.35	0.222	0.206	0.82	0.68
	1	8.5	0.215	0.223		
General Practice	1	4	0.236	0.293		
	1	5	0.274	0.238	0.8	0.59
Internal Medicine	1	8	0.193	0.126	0.85	0.71
Physician Assistant	1	10	0.262	0.324		
	1	13.5	0.361	0.291	0.74	0.51
	1	14.5	0.379	0.284		
Psychiatry	1	8	0.336	0.297	0.75	0.52

Table 13 Summary of fraud detection results for baseline and improvement strategies

Experiment	Description	owaIOA
Baseline	The learner used is Logistic Regression. The O/F method used was separate (SOF). Only classes that had a F-score above 0.75 were tested for fraudulent detection.	23.1%
Groups	Similar specialties were grouped together. All fourteen groups were tested separately. In this experiment only classes within groups were tested for fraudulent to see how the grouping affected the member within the groupings. Only groups with an F-score above 0.75 were chosen for testing fraudulent detection.	25.2%
Combining Groups	All groups found to yield improved prediction results in grouping were used in one dataset in order to determine the overall effects of grouping. All classes and groupings with an F-score above 0.75 are used for testing fraudulent detection.	26.1%
Class Removal (Original Four)	Classes were removed that were considered to have high overlap of procedures and low F-score. Only classes that had a F-score above 0.75 were tested for fraudulent detection.	15.9%
Class Removal (Chosen)	Along with the O4, classes were removed that created ambiguity, scored low in precision and recall, contained medicine, general or unknown. Only classes that had a F-score above 0.75 were tested for fraudulent detection.	14.1%
Mixed (Original Four)	This experimented combined the class removal O4 and combined grouping. Only classes that had a F-score above 0.75 were tested for fraudulent detection.	20.8%
Mixed (Chosen)	This experimented combined the class removal chosen and combined grouping. Only classes that had a F-score above 0.75 were tested for fraudulent detection.	21.1%

fraud detection capability compared to the baseline with an owaIOA of 15.9% for the original four classes and 14.1% for the chosen classes criteria. Mixing both combined grouping and class removal showed that the predictive results were increased more than the sum of each strategy alone. The fraud detection results, however, both decreased in comparison to the baseline model with owaIOA scores of 20.8% and 21.1%. However, the class isolation method, summarized in Table 14, shows good performance, especially for Internal Medicine with a Type I error of 0.193 and a low Type II error at 0.126. The IOA of fraudulent instances for Internal Medicine is 87.4% indicating that percentage of fraudulent behaviors would be correctly identified as fraudulent, exhibiting promising detection performance. Note that these lower error rates are only achieved using a cost sensitive classifier with an RF model.

4.5 Research limitations

Given the difficulty in detecting Medicare fraud, we briefly summarize some possible limitations of our models and proposed improvement strategies. Overall, our study focuses only on claims information in the Medicare Part B dataset. Also, our fraud detection results depend on possible fraudulent activities being represented by a physician/provider showing claims patterns (from procedures performed) outside of their primary specialty. This, however, does not specifically account for or consider other possible confounding variables that could impact service utilization. For the baseline, grouping, and class removal strategies, we can only detect certain types of fraud where procedure codes are used that do not align with a primary specialty. In grouping similar classes, we could potentially remove a specialty with which another specialty could be classified as, such as Optometrist and Ophthalmology. Finally, class isolation only uses a subset of the entire Part B data, limited by physician type and number of non-fraudulent instances.

Table 14 Isolation method error rates

Specialty	Type I	Type II
Chiropractic	0.403	0.394
Family Practice	0.215	0.223
General Practice	0.274	0.238
Internal Medicine	0.193	0.126
Physician Assistant	0.361	0.291
Psychiatry	0.336	0.297

5 Conclusion and future work

Medicare fraud continues to be a problem for the U.S. government and its beneficiaries. Reducing the impact of fraud is critical in helping to reduce costs and provide high quality of service. In our study, we demonstrate, through the use of data mining and machine learning, the successful detection of Part B provider fraud for different medical specialties. In our previous research, we created a unique model to detect fraud by predicting a physician's specialty. If this predicted (or expected) specialty differs from that physician's actual specialty, as listed in the Medicare Part B data, this could be indicative of possible fraud. The reason is that this misclassified physician is not performing

procedures in a manner similar to their peers, which is considered to be anomalous. For instance, a physician who's expected specialty differs from their actual specialty could be performing fraudulent acts such as double billing, upcoding [21], or otherwise purposefully coding incorrect procedures. There are many examples of these types of fraudulent behaviors, and the interested reader can find a sample real-world Medicare conviction for upcoding at [39].

Due to the small size of the dataset used in our previous research, we were limited in performing robust fraud detection validation of the models and unable to assess any proposed improvement strategies. With that said, our current research incorporates the full Medicare Part B data thus has both baseline model and strategy assessments and validation. Because our experiments used big data, we employed different methods in building and testing baseline models. We used PySpark to build the Multinomial Naive Bayes and Logistic Regression models, where each of these models can effectively perform multi-class classification on the 89 different specialties. Along with building a baseline model, we also addressed the concern regarding a physician's place of service. We tested both office and facility by either combining (COF) or separating (SOF) the procedures and split the results based on where the procedures were performed. From the model selection process, accounting for the place of service, the Logistic Regression model using the Separate Office or Facility (SOF) was chosen as the best baseline model. From this baseline model configuration, we tested two strategies specifically to assess improvement to the baseline model. Specifically, we tested two strategies in order to increase the performance of the baseline model. Grouping similar specialties and removing specialties with a large number of possibly overlapping procedures, as well as the combination of grouping and removal strategies, produced mixed results. Even so, class grouping showed some improvements but only for certain physician specialties. In addition to the aforementioned improvement strategies, we assessed the class isolation method, using the SOF selected dataset. This method builds separate models per specialty by either using 3:1 (non-fraud to fraud instances) sampled datasets or via cost sensitive classification. Both of these methods perform well and show improved fraud detection performance over the baseline model and associated improvement strategies.

Our proposed fraud detection method and improvement strategies can be used to flag possibly fraudulent physicians based on the procedures performed. These flagged providers are then the focus of further investigation to determine any actual fraudulent behaviors and legal culpability. It is important to note that our detection approach reduces the efforts and resources required to investigate possible fraud by limiting the number of fraud instances to a small subset of all possible providers. Even so, further scrutiny is still recommended to confirm and document any fraud. Through our research, we found that the improvements from our class grouping and class removal strategies were minimal, except for a few of the specialties, thus not a viable approach for continued research in this area. The class isolation method, however, is promising and applying new approaches and data mining techniques to this method could yield increased fraud detection performance. These improvements to class isolation, such as the use of grouping, are left as future work. Other areas of future work involve adding additional features, as well as increasing the pool of known fraudulent providers.

Acknowledgments The authors would like to thank the anonymous reviewers and associate editor for their insightful evaluation and constructive feedback of this paper, as well as the members of the Data Mining and Machine Learning Laboratory, Florida Atlantic University, for their assistance in the review process. We acknowledge partial support by the NSF (CNS-1427536). Opinions, findings, conclusions, or recommendations in this paper are the authors' and do not reflect the views of the NSF.

Compliance with Ethical Standards

Conflict of interests Author 1 declares that he has no conflict of interest. Author 2 declares that he has no conflict of interest. Author 3 declares that he has no conflict of interest.

Ethical Approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Appendix A

Multinomial Naive Bayes classifies new instances, in our case these instances consist of new Medicare claims per year for a particular provider, by finding the posterior probabilities of class membership based on each feature value, which is learned from a set of labeled training instances. The approximation is done using Bayes' rule by assuming conditional independence. Conditional independence is the idea that each feature in the dataset is independent from one another which is rarely true in practice, however, the model is very effective and is used extensively in the field of data mining and machine learning [33]. Naive Bayes is used with both PySpark and Weka.

Logistic Regression predicts probabilities for which class a categorically distributed dependent variable belongs to by using a set of independent variables employing a logistic function. Multinomial Logistic Regression (MLR) is an extension of binomial Logistic Regression that allows for more than two categories of the dependent variable. Unlike Naive Bayes, there is no requirement for statistical independence between independent variables, though there is an assumption of collinearity [40, 41]. We used Logistic Regression in both Weka and PySpark. Logistic

Regression with the Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (LBFGS) is the version used in our study to improve memory usage [42]. In this research, we will be using data for both binomial and mulitnomial Logistic Regression.

Support Vector Machine models create a space consisting of training instances portraying them as points, mapping them in a way that best creates linearly separable categories, where the goal is to have the largest gap between them. The specific implementation of SVM used in Weka is called Sequential Minimal Optimization (SMO), which uses this optimization algorithm as the training method for Support Vector Classification [40, 43].

Random Forest is an ensemble learning method that generates a large number of trees. The class value that appears most often among these trees, or mode, is the class predicted as output from the model. As an ensemble learning method, RF is an aggregation of various tree predictors where each tree within the forest is dependent upon the values dictated by a random vector that is independently sampled [40, 44]. We use the Random Forest learner only in Weka.

Appendix B

Type I error rate (false positive rate) is the percentage of instances that are actually non-fraud but marked as fraud, in relation to the number of actual non-fraud instances. A fire alarm going off indicating a fire when in fact there is no fire would be an example of this kind of error. Type II error rate (false negative rate) is the percentage of instances that are actually fraud but marked as non-fraud, in relation to the actual number of actual fraud instances. As an example, a fire breaking out and the fire alarm does not ring would be considered a false negative. Note that in binary classification, finding a balance between the error rates, while minimizing the Type II error rate, is generally preferred. Recall measures the ability of a classifier to determine the rate of positively marked instances that are in fact positive; therefore, in our study, recall is the fraction of physicians labeled correctly and not as any of the other specialties. Precision indicates how well a classifier has predicted a class by finding the ratio of actually positive instances from the pool of instances that it has marked as part of the positive class; therefore, precision shows the fraction of physicians marked correctly against the number of physicians, from any of the other specialties, also marked as the class in question.

F-score (also known as F1-score or F-measure) is the harmonic mean of both precision and recall, generating a number between 0 and 1, where values closer to one indicate better performance. For this study, we assume equal weighting between precision and recall, with $\beta = 1$, as seen in Eq. 1.

$$F_1 = (1 + \beta^2) \times \frac{Recall \times Precision}{(\beta^2 \times Recall) + Precision} \tag{1}$$

G-measure, also known as the Fowlkes-Mallows index, gives the geometric mean of precision and recall giving the central point between the values as seen in Eq. 2.

$$\text{G-measure} = \sqrt{Recall \times Precision} \tag{2}$$

Finally, in order to leverage the successes from our prior works, we manipulated the datasets, by filtering out certain specialties only, so that we could test one fraudulent specialty at a time (based on the model predicting the physician's specialty). Since the test dataset contains only one specialty at a time, the overall accuracy would be the percentage of real-world fraudulent physicians that are considered not fraudulent. In order to capture the the percentages of classes that are labeled as another class, we incorporate the Inverse Overall Accuracy (IOA) performance measure. IOA, where IOA = 1 − overall accuracy, is the percentage of fraudulent physicians marked as fraudulent for a given specialty. As shown in Eq. 3, to calculate the model's overall weighted average IOA (owaIOA), we take the IOA for a specialty, the number of fraudulent instances (n) for that specialty and the total number of instances (N), and sum over the total number of fraudulent instances between all specialties with F-score of 0.75 or above (NoS).

$$owaIOA = \sum_{i=1}^{i=NoS} \frac{n_i}{N} \times IOA_i \tag{3}$$

Appendix C

Table 15 Fraud detection results with groups only (group test one)

Group name	Group F-score	Instances	IOA
Anesthesiology Group	0.97	44	0.36
Cardiology Group	0.86	13	0.54
Chiropractic Group	0.94	148	0.19
Dermatology Group	0.75	11	0.09
Gynecology Group	0.81	40	0.55
Ophthalmology Group	0.96	28	0.04
Otolaryngology Group	0.87	17	0.18
Radiology Group	0.96	4	0.00

Table 16 Fraud detection results with groups and non-grouped specialties (group test two)

PROVIDER_TYPE	F-score	Instances	IOA
Ambulance Service Supplier	1.00	2	0.00
Audiologist (billing independently)	0.99	8	0.00
Anesthesiology Group	0.97	44	0.36
Speech Language Pathologist	0.97	3	0.33
Radiology Group	0.96	4	0.00
Pathology	0.95	4	0.00
Opthalmology Group	0.94	28	0.04
Chiropractic Group	0.94	148	0.18
Occupational therapist	0.92	6	0.17
Podiatry	0.90	47	0.09
Urology	0.90	13	0.23
Otolaryngology Group	0.88	17	0.24
Gastroenterology	0.88	7	0.29
Cardiology Group	0.86	13	0.54
Emergency Medicine	0.81	32	0.47
Gynecology Group	0.81	40	0.55
Nephrology	0.78	1	0.00
Independent Diagnostic Testing Facility	0.78	5	0.40
Dermatology Group	0.76	11	0.09
Orthopedic Surgery	0.76	11	0.91

Table 17 Class removal (original four classes) fraud detection results

PROVIDER_TYPE	F-score	Instances	IOA
Ambulance Service Supplier	1.00	2	0.00
Chiropractic	1.00	86	0.00
Audiologist (billing independently)	0.99	8	0.00
Physical Therapist	0.99	17	0.00
Speech Language Pathologist	0.97	3	0.33
Dermatology	0.96	11	0.00
Urology	0.96	13	0.08
Pathology	0.95	4	0.00
Diagnostic Radiology	0.94	4	0.25
Otolaryngology	0.94	13	0.23
Cardiology	0.93	13	0.31
Podiatry	0.93	47	0.02
Emergency Medicine	0.93	32	0.50
Gastroenterology	0.92	7	0.14
Occupational therapist	0.92	6	0.17
Nephrology	0.91	1	0.00
Orthopedic Surgery	0.87	11	0.55
Obstetrics/Gynecology	0.86	38	0.42
Allergy/Immunology	0.85	4	0.00
Optometry	0.82	8	0.25
Ophthalmology	0.82	20	0.15
Pulmonary Disease	0.81	10	0.20
Independent Diagnostic Testing Facility	0.78	5	0.00
Clinical Laboratory	0.77	3	0.00
CRNA	0.76	12	0.17

Table 18 Class removal (chosen classes) fraud detection results

PROVIDER_TYPE	F-score	Instances	IOA
Ambulance Service Supplier	1.00	2	0.00
Chiropractic	1.00	86	0.00
Audiologist (billing independently)	0.99	8	0.00
Physical Therapist	0.99	17	0.00
Speech Language Pathologist	0.97	3	0.33
Urology	0.96	13	0.00
Dermatology	0.96	11	0.00
Gastroenterology	0.95	7	0.14
Diagnostic Radiology	0.94	4	0.25
Podiatry	0.94	47	0.02
Pathology	0.94	4	0.00
Emergency Medicine	0.94	32	0.34
Otolaryngology	0.94	13	0.23
Cardiology	0.93	13	0.23
Occupational therapist	0.92	6	0.17
Nephrology	0.91	1	0.00
Orthopedic Surgery	0.89	11	0.45
Allergy/Immunology	0.86	4	0.00
Obstetrics/Gynecology	0.86	38	0.37
Optometry	0.84	8	0.25
Ophthalmology	0.83	20	0.15
Pulmonary Disease	0.81	10	0.30
Clinical Laboratory	0.79	3	0.00
Independent Diagnostic Testing Facility	0.79	5	0.00
CRNA	0.76	12	0.17
Neurology	0.75	19	0.26

References

1. HHS.gov. Hhs report: Average health insurance premiums doubled since 2013. [Online]. Available: https://www.hhs.gov/about/news/2017/05/23/hhs-report-average-health-insurance-premiums-doubled-2013.html
2. Forbes. Healthcare – 5, 10, 20 years in the past and future. [Online]. Available: https://www.forbes.com/sites/singularity/2012/07/02/healthcare-5-10-20-years-in-the-past-and-future/#4d2c89b4310b
3. U.S. Department of Health and Human Services. Health, United State, 2016. [Online]. Available: https://www.cdc.gov/nchs/data/hus/hus16.pdf
4. Feldstein M (2006) Balancing the goals of health care provision and financing. Health Aff 25(6):1603–1611

5. U.S. Centers for Medicare & Medicaid Services (2018) What's Medicare. [Online]. Available: https://www.medicare.gov/sign-up-change-plans/decide-how-to-get-medicare/whats-medicare/what-is-medicare.html
6. U.S. Centers for Medicare & Medicaid Services (2017) Other entities frequently asked questions. [Online]. Available: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/other-entities-faqs.pdf
7. U.S. Centers for Medicare & Medicaid Services. (2012) Medicare claim submission guidelines fact sheet. [Online]. Available: http://www.nacns.org/wp-content/uploads/2016/11/CMS_ReimbursementClaim.pdf
8. Joudaki H, Rashidian A, Minaei-Bidgoli B, Mahmoodi M, Geraili B, Nasiri M, Arab M (2016) Improving fraud and abuse detection in general physician claims: a data mining study. Int J Health Policy Manag 5(3):165
9. Pawar MP (2016) Review on data mining techniques for fraud detection in health insurance. IJETT 3:2
10. Coalition Against Insurance Fraud. By the numbers: fraud statistics. [Online]. Available: http://www.insurancefraud.org/statistics.htm
11. Lambert J, Dunstan R Civil recovery schemes: for or against? [Online]. Available: https://www.theguardian.com/law/2010/dec/07/civil-recovery-schemes-for-or-against
12. Medicare Fraud Strike Force. Office of inspector general. [Online]. Available: https://www.oig.hhs.gov/fraud/strike-force/
13. Ko JS, Chalfin H, Trock BJ, Feng Z, Humphreys E, Park S-W, Carter HB, Frick KD, Han M (2015) Variability in medicare utilization and payment among urologists. Urol 85(5):1045–1051
14. Santa Clara, Oct 6, 2013, in Conjuction with the IEEE International Conference on BigData. Bigdata in bioinformatics and health care informatics. [Online]. Available: http://www.ittc.ku.edu/jhuan/BBH/
15. Bauder RA, Khoshgoftaar TM, Richter A, Herland M (2016) Predicting medical provider specialties to detect anomalous insurance claims. In: 2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI). IEEE, pp 784–790
16. Herland M, Bauder RA, Khoshgoftaar TM (2017) Medical provider specialty predictions for the detection of anomalous medicare insurance claims. In: 2017 IEEE 18th international conference information reuse and integration (IRI). IEEE, pp 579–588
17. Bauder RA, Khoshgoftaar TM (2018) A survey of medicare data processing and integration for fraud detection. In: 2018 IEEE 19th international conference on information reuse and integration (IRI). IEEE, pp 9–14
18. CMS. Research, statistics, data, and systems. [Online]. Available: https://www.cms.gov/research-statistics-data-and-systems/research-statistics-data-and-systems.html
19. CMS. Medicare provider utilization and payment data: Physician and other supplier. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html
20. LEIE. Office of inspector general leie downloadable databases. [Online]. Available: https://oig.hhs.gov/exclusions/authorities.asp
21. Bauder RA, Khoshgoftaar TM, Seliya N (2017) A survey on the state of healthcare upcoding fraud analysis and detection. Health Serv Outcome Res Methodol 17(1):31–55
22. Feldman K, Chawla NV (2015) Does medical school training relate to practice? Evidence from big data. Big Data 3(2):103–113
23. Bauder RA, Khoshgoftaar TM (2017) Multivariate outlier detection in medicare claims payments applying probabilistic programming methods. Health Serv Outcome Res Methodol 17(3–4):256–289
24. Sadiq S, Tao Y, Yan Y, Shyu M-L (2017) Mining anomalies in medicare big data using patient rule induction method. In: 2017 IEEE third international conference on multimedia big data (BigMM). IEEE, pp 185–192
25. Chandola V, Sukumar SR, Schryver JC (2013) Knowledge discovery from massive healthcare claims data. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1312–1320
26. Branting LK, Reeder F, Gold J, Champney T (2016) Graph analytics for healthcare fraud risk estimation. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 845–851
27. CMS. National provider identifier standard (npi). [Online]. Available: https://www.cms.gov/Regulations-and-Guidance/Administrative-Simplification/NationalProvIdentStand/
28. CMS. HCPCS - General Information. [Online]. Available: https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html?redirect=/medhcpcsgeninfo/
29. OIG. Office of inspector general exclusion authorities. [Online]. Available: https://oig.hhs.gov/exclusions/index.asp
30. Pande V, Maas W (2013) Physician medicare fraud: characteristics and consequences. Int J Pharm Healthc Mark 7(1):8–33
31. The R Foundation. What is r? [Online]. Available: https://www.r-project.org/about.html
32. Python Software Foundation. Python. [Online]. Available: https://www.python.org/
33. Witten IH, Frank E, Hall M, Pal CJ (2016) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, San Mateo
34. Apache. Welcome to spark python api docs! [Online]. Available: http://spark.apache.org/docs/2.1.0/api/python/index.html
35. Shanahan JG, Dai L (2015) Large scale distributed data science using apache spark, pp 2323–2324, 08
36. Khoshgoftaar TM, Seiffert C, Van Hulse J, Napolitano A, Folleco A (2007) Learning with limited minority class data. In: 2007 sixth international conference on machine learning and applications, ICMLA 2007. IEEE, pp 348–353
37. Van Hulse J, Khoshgoftaar TM, Napolitano A (2007) Experimental perspectives on learning from imbalanced data. In: Proceedings of the 24th international conference on machine learning. ACM, pp 935–942
38. Weka. Costsensitiveclassifier. [Online]. Available: https://weka.wikispaces.com/CostSensitiveClassifier
39. Department of Justice U.S. Attorney's Office (2016) Federal jury convicts tinley park physician in medicare fraud scheme. [Online]. Available: https://www.justice.gov/usao-ndil/pr/federal-jury-convicts-tinley-park-physician-medicare-fraud-scheme
40. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. ACM SIGKDD Explor Newsl 11(1):10–18
41. Le Cessie S, Van Houwelingen JC (1992) Ridge estimators in logistic regression. Appl Stat 41(1):191–201
42. Malouf R (2002) A comparison of algorithms for maximum entropy parameter estimation. In: Proceedings of the 6th conference on natural language learning, vol 20. Association for Computational Linguistics, pp 1–7
43. Platt J, Smola A (1998) Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf B, Burges C (eds) Advances in Kernel methods - support vector learning. MIT Press. [Online]. Available: http://research.microsoft.com/~jplatt/smo.html
44. Breiman L (2001) Random forests. Mach Learn 45(1):5–32