COMPARISON OF DIFFERENT MACHINE LEARNING ALGORITHMS AND SAMPLING METHODS FOR MEDICAL FRAUD DETECTION

Ahmed Burak Gulhan, Kaan Kurama, Kadir Bulut Ozler

Abstract Provider fraud is a growing problem that has impactful costs on healthcare institutions and on Medicare. Fraud detection is an issue that has its uncommon problems, and different approaches must be taken to obtain successful results. This paper presents a comprehensive evaluation of various machine learning algorithms and data analysis techniques applied on the public Medicare Provider Data. To mitigate class imbalance, random undersampling and oversampling methods have been used. After the analysis and processing of the data, Logistic Regression, C4.5 and Random Forest algorithms have been applied to detect fraudulent providers. Each algorithm has its own advantages and setbacks. Each algorithm has been tested on different class ratios and the results have been displayed in the confusion matrix. The goal of this paper is to review how algorithms perform under different class ratios.

Keywords: Fraud Detection, Class Imbalance, Supervised Learning, Healthcare Provider Fraud

I. Introduction and Motivation

Healthcare is a vital system that is essential for the wellbeing of the population. Medical treatment is specifically important for the elderly population, and according to the U.S. Census Bureau, 21.8% of the U.S. population is projected to be older than 65 years of age, in 2045. Thus, healthcare industry is set to scale and more people will be dependent on these services in the future. Medicare is a U.S. government program providing insurance to beneficiaries older than 65 years of age, or to people with special medical conditions. It was stated that in 2016, U.S. Healthcare spending increased 4.3% from 2015, reaching $3.3 trillion in expenditures [1]. Medicare expenditure, in particular grew to $672.1 billion. Evidently, healthcare industry is costly for the economy, and it would be for the profit of the population if these costs were reduced. The Centers for Medicare & Medicaid Services (CMS) states that the rising costs of medical goods and services, increased spending on prescription drugs, and growth on the health industry enrollments are key factors of this increase [1]. Inevitably, Healthcare industry is highly valuable and critical, and thus it is subject to fraud. he Federal Bureau of Investigation (FBI) suspects that 3-10% of healthcare expenditure are wasted on fraud [2]. Given that Medicare accounts to 20% of overall spendings, in our research we have focused on provider fraud on Medicare Suppliers. Fraudsters profit from this vulnerable field. Providers, and even beneficiaries are involved in healthcare fraud, and this generates a hefty burden on the industry. Regrettably, even a small portion of physicians exploit the integrity of their field by exploiting the trust of their patients. Clearly, their financial interests far outweigh the integral values of their profession.

The methods of fraudsters are constantly changing and adapting to current conditions of the industry. Organized crime, schemes that span multiple states and even patients are involved in healthcare fraud. Thus, development of fraud detection methods and algorithms are essential for the population, and for the economy. Fraud detection can reduce wasted money, and can significantly benefit the healthcare industry and its beneficiaries. Healthcare is a necessary, essential need for everyone, and especially for the elderly population. Its value and importance grows every day.

Several Medicare datasets have been publicized by CMS with intention of combatting fraud. These datasets contain information about services conducted by physicians, other medical professionals and suppliers, provided to Medicare

G. A. Author. He is now studying in the Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey (e-mail: gulhan16@itu.edu.tr).

K. K. Author. He is now studying in the Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey (e-mail: kurama16@itu.edu.tr).

O. B. Author. He is now studying in the Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey (e-mail: ozler16@itu.edu.tr).

beneficiaries. The data contains financial information about the products or services, about the transactions, and it also displays properties of the supplier. In our research, we have used the *Medicare Provider Utilization and Payment Data: Physician and Other Supplier* dataset, specifically the versions released in 2016 and 2017. The dataset does not class the data as fraud or non-fraud. Thus we could have implemented an anomaly detection algorithm to work with this unlabeled dataset. Fortunately, the Office of Inspector General's (OIG) List of Excluded Individuals/Entities (LEIE) database contains information about providers and entities who have committed fraud and who are thus exempt from providing services to Medicare. However, it must be noted that 38% of convicted physicians continued to practice their profession, and 21% were not punished [3]. Both datasets include National Prover Identifier (NPI) as a feature. NPI has been used to label the unlabeled Physician and Supplier dataset.

A. Challenges with healthcare fraud detection

Fraud detection is an exceptionally difficult task in the field of machine learning. It has many constraints and challenges. Due to its potential impact and adapting adversaries, fraud is an issue that has to be solved. In this section, the issues regarding this domain are discussed.

Unavailability of Datasets: There is a limited amount of publicized fraud data available. One main requisite of machine learning is the need of data. With more data, more accurate models can be built. However, in the field of fraud, companies and organizations tend to not release fraud datasets. One reason is confidentiality. Releasing confidential data would cause security problems, and would potentially expose the privacy of the entities in datasets [4,6]. In brief, the limited amount of data available may not be sufficient to detect the distinct and countless different fraud patterns.

Imbalanced Datasets: A very important and unique property of fraud databases is the skew in the binary fraud class. In other words, the majority of the data is non-fraud data [4,5], and a very small portion of the data represents fraud. This causes an imbalance between classes and has an adverse effect on the classifiers. In the case of financial fraud, fraudulent cases represent less than 1% of the entire dataset [7].

Adaptive Fraud Techniques: Fraud is a profitable business that has a lot of money at stake. Thus, fraudsters continue to create and modify their complex techniques to overcome the current detection systems [4-5]. Also, how fraud is defined in a field can change over time [8].

Noise: Presence of noise in a highly imbalanced dataset further damages the classifier. This may cause the classifier to overfit the data, and result in mispredictions [4].

Comprehensibility: Another factor for the lack of data is that some frauds have not yet been detected. Fraud takes effort to be discovered. According to another research, simply flagging fraudulent cases are not sufficient. The fraudulent claims undergo a series of investigations before a decision is taken on them [5]. Also, several investigations and fraudulent claims, may conflict with each other

The issues stated in this section display the difficulty of fraud detection domain.

B. Our approach

Training classifiers on a highly imbalanced dataset would be pointless, so initially we modify the dataset using sampling methods. In our project, we have tested our algorithms on the following class ratios: 90:10, 75:25, 65:35, and 50:50. Ratio on the left represents non-fraudulent cases, and the one on the right represents fraudulent cases. In order to obtain these ratios, we have used different sampling methods, namely Random Under Sampling (RUS), Random Over Sampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE). We have chosen these techniques as they have been proposed as being successful in handling imbalanced datasets [10]. We also would have preferred to use two additional methods, SMOTE + Tomek Link Removal, and ADASYN (Adaptive Synthetic Sampling Method for Imbalanced Data), however our machines' computational power did not suffice to train methods and learners using these two last samplers.

Then, we apply various supervised classifier methods on the new dataset to detect illegitimate cases. Logistic Regression (LR) and C4.5 Decision Tree learners have been proven to yield decent results on imbalanced datasets so we employ them [9]. Additionally, we use Random Forest (RF) algorithm as our third learner, in order to observe the outlier detection performance.

Our results display a variety of different performances. For the combined dataset, by employing several learning algorithms with multiple sampling methods, we reached a 0.746 score for ROC-AUC. We calculated average precision and recall scores of classes. In this metric best score is 0.79 and 0.75 respectively. For macro-F1 score, our best result has been 0.67. As you can see in table 1,2,3 and 4, Random Forest Classifier performs significantly better than other learning algorithms.

As stated before, fraud datasets are scarce. Yet, datasets representing medical fraud are even more rare. Thus, the primary contribution of our work is the comparison of various supervised learners, and sampling methods on Medicare data. We obtain our results by applying different learners to datasets with different class ratios. We achieve each class ratio using different sampling methods. Not only our results display how different learners perform in

Medicare dataset, but they also display how different sampling methods affect the precision. We discuss the sensitivity of each method, and we obtain a general idea about the issues regarding classification of highly skewed imbalanced datasets.

Also, to the best of our knowledge, a research about SMOTE+Tomek sampling method on medical fraud dataset has not been published before.

The remainder of the paper is organized as follows. In Section II, related works to our current research is discussed. In Section III, the datasets are discussed in depth. Additionally, we give a brief introduction to the learners and to the sampling methods that we use. In Section IV, we show the results of our research, examine and compare each result. Finally, our conclusions and possible future work are presented in Section V.

II. Related works

Since the release of datasets by The Centers for Medicare and Medicaid Services there has been a number recent research papers that analyze comparisons between different supervised learning algorithms to detect medical fraud. There are also a number of papers that analyze different sampling methods and how they effect the performance of an algorithm. In our research we will analyze the performance between different supervised learning algorithms, sampling methods and class ratios. In most papers on medical fraud detection at least one of these are not taken into account. For example Bauder and Khoshgoftaar [1] compare class ratios with different supervised learning algorithms, however they only use random undersampling (RUS) as their sampling method. On the other hand Branting et al. [18] only use a balanced 50:50 ratio when training their classifier. A 50:50 class ratio is not necessarily the best class ratio to use in medical fraud detection. Bauder and Khoshgoftaar [19] show that for the algorithm Random Forest, using a balanced 50:50 class distribution gives the same performance as using a 99:1 class distribution. In fact the best fraud detection performance for Random Forest was found to be given by the class distribution ratio of 90:10.

We found two papers that evaluate the performance by comparing between different supervised learning algorithms, sampling methods and sampling ratios which is what we have done in our research [14,f]. In [14], the authors compare and analyze three different machine learning algorithms (Logisitic Regression, Random Forest, and Gradient Boosted Trees) along with six different sampling techniques (Random Undersampling, Random Oversampling, Synthetic Minority Oversampling Technique (SMOTE), Boderline-SMOTE type 1, Boderline-SMOTE type 2 and Adaptive Synthetic (ADASYN)) and five different class ratios (99:1, 90:10, 75:25, 65:35, 50:50). The authors used three Medicare datasets, from the years 2013 to 2015, which are provided by the Centers for Medicare and Medicaid Services (CMS) [20], that include Part B, Part D, and DMEPOS. Each learning algorithm was evaluated by performing 5-fold cross-validation 10 times and getting the average over all 10 passes. For the performance metric Area Under the Receiver Operating Characteristic Curve (AUC) was used. The results show that the learner Linear Regression has the best overall performance while the learner ,Gradient Boosted Trees has the worst performance. The best single result, an AUC of 0.82793, was obtained by using the Random Forest along with the sampling method Random Under Sampling with a ratio of 90:10. Linear Regression had the best performance using the sampling method SMOTE while Gradient Boosted Trees and Random Forest had the best performance using Random Under Sampling. In [21] Herland compares six different machine learning algorithms (Multinomial Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine, and Gradient Boosted Tree) along with three sampling algorithms (Random Over Sampling, Random Under Sampling, and SMOTE). The database used was the three CMS datasets: Part B, Part D, and DMEPOS [20] and their combination along with LEIE [22] for labeling fraudulent practitioners. The best result found was by using Logistic Regression with Random Under Sampling with a 90:10 class ratio. The best sampling method was Random Under Sampling however in some cases SMOTE performed better. In [21] Herland also tried training separate classifiers for each profession type. This gave a much better performance compared to training one classifier for all professions. In this case Multinomial Naive Bayes had the best performance.

We found two papers that compare learning algorithms together with class distribution ratios without comparing different sampling methods[1][16]. In [1] Bauder et al. compares different supervised learning algorithms along with different class distribution ratios. For the sampling method, only Random Under Sampling is used. The dataset used here is from Centers for Medicare and Medicaid Services (CMS) [20] which from the years 2012 to 2015 combined with the Medicare LEIE database from the year 2017 [22] which contains all medical practitioners who have committed fraud up to that date. From the LEIE only fraudulent incidences severe enough to get a 5-year exclusion period or more were used. The final Medicare dataset, used had 3,331 instances labeled as fraudulent and 37,143,882 instances of non fraudulent. The authors compared three different learners (C4.5 decision trees, Support Vector Machines and Logistic Regression) along with four different class distribution ratios (80:20, 75:25, 65:35, 50:50). Each learning algorithm was evaluated by performing 5-fold cross-validation. The learning method C4.5 decision trees was found to have the highest AUC score across all class distribution ratios with Linear Regression being a close second. The single highest AUC score, which is 0.883, was gotten by using C4.5 with a 80:20 class distribution ratio. In [16] Herland et al. compares several learners (Linear Regression, Random Forest, and Gradient Boosted Trees) together with different class distribution ratios (99:1, 90:10, 75:25, 65:35, 50:50). The sampling method Random Under Sampling was the

only sampling method used. Each learning algorithm together with the datasets was evaluated by performing 5-fold cross validation. For the performance metric AUC was used. Linear Regression showed the overall best results across all class distribution ratios, with the highest score obtained using a 90:10 class distribution ratio.

We found one paper that compares learning algorithms together with sampling methods without comparing class distribution ratios[24]. In [24] Bauder and Khoshgoftaar compares the following supervised learning algorithms: Gradient Boosting Machines; Random forest; The Naive Bayes algorithm. The authors also compared several unsupervised learning algorithms. The sampling methods used were Random Under Sampling and Random Over Sampling. The class distribution ratio was set as 80:20. The database used is the CMS Physician and Other Supplier Data 2015 database [20] along with the List of Excluded Individuals/Entities (LEIE) database [22]. Each learning algorithm was evaluated by performing 5-fold cross validation. For the performance metrics AUC and G-score were used. After training and testing each model, Random Under Sampling with a 80:20 class distribution ratio was found to give the best results while Random Over Sampling gave significantly worse results.

We found one paper that compares sampling methods together with different class distribution ratios without changing the type of learner used[15]. In [15] Johnson and Khoshgoftaar used a deep neural network along with several different sampling methods (Random Under Sampling (RUS), Random Over Sampling (ROS), and ROS-RUS hybrids) and several class distribution ratios (99:1, 80:20, 60:40, 50:50, 40:60). The CMS Medicare Provider Utilization and Payment Data: Physician and Other Supplier database was used [20] along with the List of Excluded Individuals and Entities (LEIE) database [22] to label the data. 80% of the data was used with 10% used for training and 10% used for hyperparameter validation and calculating optimal thresholds. The performance measures AUC and G-mean were used. The sampling methods ROS and ROS-RUS hybrid gave the best results, outperforming baseline scores with average AUC scores of 0.8505 and 0.8509, respectively.

We found two paper that compares different sampling distribution ratios with only one type of sampling method and one type of learner [19]. In [19] Hasanin and Khoshgoftaar compare the effect of class distribution methods on the supervised learning algorithm Random Forest using the sampling method Random Under Sampling. The dataset Medicare Part B data set 2012–2016, namely the Medicare Provider Utilization and Payment Data: Physician and Other Supplier [20] along with the List of Excluded Individuals and Entities (LEIE) database [22] was used. The class distributions of 1:1, 10:1, 100:1, 1000:1, 10000:1, and 100,000:1 were used. The authors used four different datasets with three of them gathered from UCI Machine Learning Repository. Each dataset was evaluated using 5-fold cross validation. The 1:1 gave the same result as the 99:1 class distribution for Random Forest. The lowest result came from using 100,000:1 and the best result from using 90:10.

We found one paper that compares only different supervised learning algorithms without using any sampling. In [23], Herland et al. compares the three supervised learning algorithms: Logistic Regression (LR), Random Forest (RF), and Gradient Boosted Trees (GBT). No sampling method was used. The dataset used was the combination of three Medicare datasets, from the years 2013 to 2015, which are provided by the Centers for Medicare and Medicaid Services (CMS) [20], that include Part B, Part D, and DMEPOS [20] and for labeling used the List of Excluded Individuals/Entities (LEIE) dataset [22]. In these datasets there is a huge imbalance between fraudulent practitioners and non fraudulent practitioners. For the non-numeric data in this set, one-hot encoding was used. Since there is no sampling method, this large class imbalance is used when training the models. Five-fold cross-validation was used when evaluating each learner. For the performance measure AUC was used. The results show that Logistic Regression has the best performance while Random Forest has the worst performance.

III. Proposed solution: comparison of different machine learning algorithms and sampling methods for medical fraud detection

Initially the data is sampled into several class ratios. Different sampling methods are used in this section. 5-fold cross validation will be used to train and test the dataset, thus, 20% of the data will not be used in training. Then, the previously specified supervised learner methods are used on the dataset. The results will be validated using the 20% testing set, and the performance of each method will be evaluated using the Performance Metrics of AUC and F1 scores.

A. Data

Center for Medicare and Medicaid Services has been publicizing its Medicare datasets. Specifically, we've used *Medicare Provider Utilization and Payment Data: Physician and Other Supplier* datasets, released in 2016 and 2017 [20]. These datasets contain information about the services and procedures that healthcare professionals and providers give to the Medicare program.

This dataset, denoted as "Part B Dataset" contains information about healthcare services that physicians commit in a year [20]. Also, there are information about average payments and charges, amount of procedures that the physician has conducted. Healthcare Common Procedure Coding System (HCPCS) code is also included. This code is used to identify products, supplies and services.

The data does not contain information about fraudulent services. In order to label the data, List of Excluded Individuals and Entities (LEIE) dataset has been used. This dataset contains information about fraudulent healthcare providers. Both datasets contain an "NPI" column. NPI stands for National Provider Identifier. It is a unique value assigned to every provider. By cross referencing the NPI

values in the LEIE dataset with the entries in the Medicare dataset, we labeled our data. One thing to note is that we've considered any services prior to the exemption date of the provider as also fraudulent.

Below table represents some features of the dataset, along with their explanations.

Feature	Description	Type
Npi	Unique provider identification number	Categorical
Provider_type	Medical provider's specialty (or practice)	Categorical
Nppes_provider_gender	Provider's gender	Categorical
Line_srvc_cnt	Number of procedures/services the provider performed	Numerical
Bene_unique_cnt	Number of distinct Medicare beneficiaries receiving the service	Numerical
Bene_day_srvc_cnt	Number of distinct Medicare beneficiary/per day services	Numerical
Average_submitted_chrg_amt	Average of the charges that the provider submitted for the service	Numerical
Average_medicare_payment_amt	Average payment made to a provider per claim for the service	Numerical
Exclusion	Fraud labels from the LEIE dataset	Categorical

B. Sampling

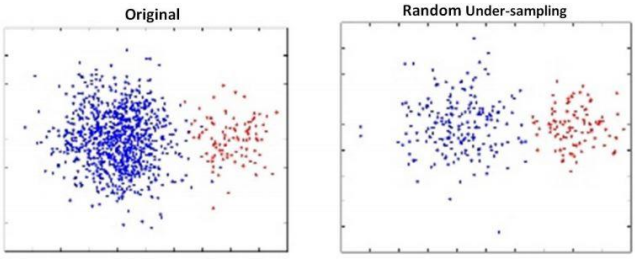
Class imbalance is a problem that occurs in many fraud datasets. The performance of classifiers are significantly weakened in presence of imbalanced class ratios. The imbalance causes the learner to be biased towards the majority class [16]. Additionally, if the class ratio is very high, the learner will tend to perform with high accuracy. However, this high accuracy will not reflect the truth, since our interest is on the minority class.

We have used sampling methods to mitigate the severe class imbalance. There are two main methods of sampling, oversampling and undersampling. The former, populates the minority class whereas the latter removes samples from the majority class. The manner in which samples are removed or populated, defines the properties of the sampler.

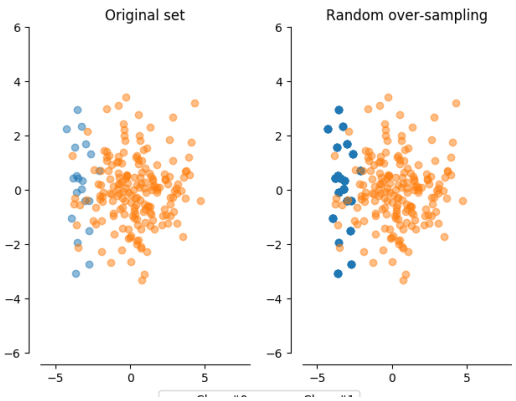
Finally, our intention on using different samplers was, to observe how these differences would affect the learners.

The rest of this section represents the samplers that we have used.

1) *Random Undersampling (RUS)*: This technique uniformly discards samples from the majority class. It is simple, but one drawback is the potential loss of important data [10]. As its name suggests, entries are removed randomly. Samples from majority class are excessively dropped. This leads to faster computations [15].



2) *Random Oversampling (ROS):* In this technique, entries from the minority class are populated. The duplication is done randomly. The count of minority class are increased. However, since more entries are added, computational costs increase [14]. Also, since synthetic data is generated, clustering may occur, which may lead to overfitting of the learner.



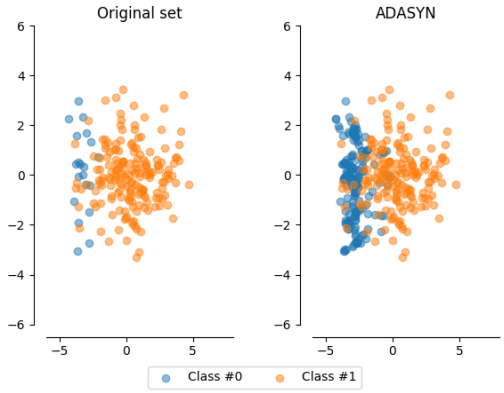
3) *Synthetic Minority Oversampling Technique* (SMOTE): SMOTE is an oversampling technique that uses existing minority class samples to generate new samples. In brief, for all samples of minority class, closest n neighbors of the same class are found. A line is drawn between each neighbor and the original sample, then, samples are generated in between. Below figure displays the pseudocode [10].

For each point p in S:
1. Compute its k nearest neighbors in S.
2. Randomly choose $r \leq k$ of the neighbors (with replacement).
3. Choose a random point along the lines joining p and each of the r selected neighbors.
4. Add these synthetic points to the dataset with class S.

We initially intended to use the following two sampling methods, however given that our data is big data, our machines' computation weren't enough to train the learners on them. Nevertheless, we opted to include the following two sampling methods as they have been used in related works.

4) *SMOTE + Tomek Link Removal:* If any two sample, of different classes, are each other's closest neighbor, these two samples are referred to as Tomek Link. In Tomek Link Removal, usually both data in the pair are deleted. However, in this study, we have merged SMOTE and Tomek sampling methods. First, we have generated synthetic data from the minority class. Then, we have calculated all Tomek Links in the oversampled dataset. In those Tomek Links, we have opted to remove the minority class instances. Research conducted by More [10] suggests that this sampling method yields the best results with supervised learners.

5) *Adaptive Synthetic (ADASYN):* ADASYN is an oversampling technique that is a modified version of SMOTE. In ADASYN, the likelihood of any particular sample to be chosen for being duplicated, is biased towards points that belong in irregular locations [17].



As can be seen from the figure, minority samples that are located in outlier regions have been populated.

C. Classifiers

In our experiments, we will use four different classifiers: Random Forest (RF), Logistic Regression (LR), Classification and Regression Trees (CART). We chose these classifiers based on results of previous papers. These classifiers were implemented using the Python library 'scikit-learn'.

Random Forest

Random Forest (RF) is an ensemble learning method. In this algorithm a large number of trees are generated which use sampling with replacement thereby making a number of randomized datasets used to train each tree. In these trees features are selected automatically. Each tree within this forest takes a random vector given by a random number generator that is independently sampled which is used seperately by each tree. This generation of random datasets in each tree minimizes overfitting. The class value is estimated most often from these trees is the class predicted as output from the model (majority voting). As an ensemble learning method, Random Forest is an aggregation of various tree predictors. This generation makes Random Forest to give good performance even on imbalanced datasets as it prevents overfitting.

The reason we chose Random Forest as one of our classifiers is that in some papers Random Forest gave very good results [14][19]. In [14] the authors compared three different learners (Logisitic Regression, Random Forest, and Gradient Boosted Trees). The best score out of these three learners was obtained by using the Random Forest algorithm along with the sampling method Random Under Sampling with a ratio of 90:10 which gave a AUC of 0.82793. In [19] the authors compared random forest with several different class distribution ratios (1:1, 10:1, 100:1, 1000:1, 10000:1, and 100,000:1) and found out that Random Forest gives the best results when the class distribution ratio is 90:10. The authors say that many classifications use a 50:50 ratio as default, which does not give the best performance for Random Forest. Therefore the Random Forest algorithm needs further analysis, which is why Random Forest is one of our classifiers.

Logistic Regression

Logistic Regression (LR) predicts probabilities for which class a categorical distributed dependent variable belongs to. This is done by employing a logistic function. Logistic Regression uses a logistic (sigmoidal) function to generate values that are between [0,1]. These values can be interpreted as the probability of an item belonging to that class. Logistic Regression is similar to linear regression but LR uses a different hypothesis class when predicting class membership.

The reason we chose the Logistic Regression classifier is that in many research on medical fraud detection Logistic Regression gave the best overall results and in most cases the highest score compared to other supervised learning algorithms[14][23]. In [14] the authors compared three different learners (Logisitic Regression, Random Forest, and Gradient Boosted Trees) with several oversampling methods (Random Undersampling, Random Oversampling, Synthetic Minority Oversampling Technique (SMOTE), Boderline-SMOTE type 1, Boderline-SMOTE type 2 and Adaptive Synthetic (ADASYN)) and sampling ratios (99:1, 90:10, 75:25, 65:35, 50:50). Logistic Regression gave the best scores in all but one combination of these three methods and ratios. In [23], the authors used three different classifiers (Logisitic Regression, Random Forest, and Gradient Boosted Trees) with no sampling method used. The result showed that Logistic Regression gave the best performance. Since no sampling was used there was a huge imbalance in the dataset, yet Logistic Regression still performed well. From these results we can see that Logistic Regression gives very good results in almost all cases when detecting medical fraud.

C5.0 Decision Tree

Decision Trees are a non-parametric supervised learning method that can be used for classification. The algorithm makes a model that predicts the value of a target by learning simple decision rules which are inferred from the dataset features during training.

We were originally going to use a C4.5 Decision Tree, however the machine learning library that we used (scikit-learn) only had C5.0 decision trees. However C5.0 in an upgraded version on C4.5 that uses less memory and builds smaller rule-sets while being more accurate [25]. In [1], the authors compared three different algorithms (C4.5 Decision Tree (C4.5), Support vector machine (SVM) and Logistic Regression (LR)) on the CMS 2017 [20] dataset. and found

that the C4.5 Decision Tree algorithm had the best performance out of all three algorithms.

D. Performance Metrics

In order to provide a comprehensive comparison for our medical fraud detection learners we use several different performance metrics to assess our models performance. We use Area Under the Receiver Operating Characteristic Curve (AUC), F1-Score, Precision, and Recall to assess model performance, with Area Under Curve (AUC) and F1-score metrics being the most important metrics when making general comparisons of model performance.

Area Under the Receiver Operating Characteristic Curve (AUC)

We use the AUC as one of our evaluation metrics for our fraud detection models to demonstrate the performance (how well it detects fraud) of each model. AUC has been found to be an effective metric for quantifying results for studies that include datasets with
class imbalances [21]. AUC shows the performance over all decision thresholds as a single value that ranges from 0 to 1, where a perfect classifier will receive an AUC of 1, while an AUC of 0.5 is equivalent to random guessing and less than 0.5 can indicate that there is a bias towards a certain class. The ROC curve shows a the learners ability to be able to distinguish between both classes and is a comparison between false positive (FP) and true positive (TP) rates.

F1-score

F1-Score (also known as F-measure) is a calculation made using both Precision and Recall (2*((precision*recall)/(precision+recall))). F1 score is used show the model performance results as a single metric for performance comparisons which is a number between 0 and 1. Values closer to one indicate better performance. F1-Score is robust when dealing with imbalanced data and is primarily interested in the prediction of true positives (correct prediction of actual fraudulent behaviors).

Precision

Precision shows how well a classifier has predicted a class. This is done by calculating the ratio of True positives (TP) from all positive guesses (including false positives (FP)). This means is showing the proportion that a given transaction is marked correctly against the amount of total fraudulent transactions from all of the transactions calculated as fraudulent.
Precision is calculated as: TP/(TP/FP)

Recall

Recall measures the ability of a classifier to determine the rate of positively estimated instances (in our case transactions that are estimated as fraudulent) are actually positive. For our dataset Recall is showing the proportion that fraudulent transactions are labeled correctly.

Recall =TP/(TP+FN)

IV. RESULTS AND ANALYSIS

Below tables display our results.

ROC AUC:

TABLE I: ROC AUC Results for the Combined Dataset

Learner	Sampling Method	90:10	75:25	65:35	50:50
C5.0	SMOTE	0.668	0.678	0.669	0.666
	RUS	0.676	0.709	0.710	0.713
	ROS	0.657	0.655	0.653	0.646
RF	SMOTE	0.609	0.664	0.684	0.695
	RUS	0.620	0.699	0.724	**0.746**
	ROS	0.611	0.627	0.630	0.629
LR	SMOTE	0.555	0.635	0.687	0.731
	RUS	0.554	0.624	0.670	0.732
	ROS	0.547	0.633	0.667	0.733

Learners generally performed better on undersampling algorithms. Also Random Forest has earned the highest score.
C5.0 had consistent scores, and results of logistic regression has changed drastically on different split ratios.
The results for RF and LR were unstable but RF has received highest score, with random undersampling across the ratios of 75:25, 65:35 and 50:50.
As for samplers, evidently RUS performs better on more balanced datasets, and performs worser on highly imbalanced datasets. This may be because of overfitting.
The results for SMOTE were fairly stable for C5.0 algorithm, but highly unstable for RF and LR. This is a unique pattern that we have observed.
Regarding sampling ratios, 50:50 seems to have the best results across all learners and all samplers. Worst results were obtained in highly skewed datasets, with ratio of 90:10.
Looking at these scores, it can be said that RF with RUS sampler performs the best when the dataset is not very skewed.

F1:

TABLE II: Average F1 Results for the Combined Dataset

Learner	Sampling Method	90:10	75:25	65:35	50:50
C5.0	SMOTE	0.65	0.64	0.63	0.61
	RUS	0.65	0.60	0.58	0.53
	ROS	0.65	0.65	0.66	0.65
RF	SMOTE	0.65	**0.67**	**0.67**	0.66
	RUS	0.66	0.65	0.62	0.56
	ROS	0.65	0.66	0.66	**0.67**
LR	SMOTE	0.58	0.63	0.61	0.52
	RUS	0.58	0.62	0.62	0.52
	ROS	0.57	0.62	0.61	0.52

Random forest has obtained the highest scores. C5.0 has performed better on more skewed datasets than it did on

more balanced datasets. As for RF, the values were stable for all samplers and all ratios, and it acquired highest score of 0.67.
LR has stable, yet weak results. Smote seems to be the most optimal option regarding this score, for LR.
Regarding sampling ratios, 75:25 seems to have the best results across all learners. Worst results were obtained on the most balanced ratio, 50:50.
As for samplers, SMOTE seems to have performed the best, on more skewed datasets. For more balanced datasets, ROS seems to have worked better.
Finally, LR seems to be unphased by different ratios and even different classifiers.
In general, F1 score has yielded the most stable results across all performance metrics. However, stability visibly came in the expense of lower scores. And similar to precision scores, RF has obtained highest scores on all ratios.

Recall:

TABLE III: Average Recall Results for the Combined Dataset

Learner	Sampling Method	90:10	75:25	65:35	50:50
C5.0	SMOTE	0.67	0.68	0.67	0.67
	RUS	0.68	0.71	0.71	0.71
	ROS	0.66	0.66	0.65	0.65
RF	SMOTE	0.61	0.66	0.68	0.70
	RUS	0.62	0.70	0.72	**0.75**
	ROS	0.61	0.63	0.63	0.63
LR	SMOTE	0.55	0.64	0.69	0.73
	RUS	0.55	0.62	0.67	0.73
	ROS	0.55	0.63	0.67	0.73

Looking at the Recall table we see that for the C5.0 Decision Tree (DTC) learner we have the best recall using Random Under Sampling (RUS) while having the worst recall with Random Over Sampling (ROS). For Random Forest Classifier (RFC) we again have the highest recall using Random Under Sampling (RUS) while having the worst recall with Random Over Sampling (ROS). In Logistic Regression classifier (LR) recall is very similar across all sampling methods, with the best recall, by a small margin, being with SMOTE at a 65:35 class distribution ratio.

Precision:

TABLE IV: Average Precision Results for the Combined Dataset

Learner	Sampling Method	90:10	75:25	65:35	50:50
C5.0	SMOTE	0.64	0.62	0.61	0.59
	RUS	0.63	0.59	0.57	0.56
	ROS	0.65	0.65	0.66	0.65
RF	SMOTE	0.75	0.68	0.67	0.63
	RUS	0.75	0.63	0.60	0.58
	ROS	0.78	0.73	0.74	0.74
LR	SMOTE	**0.79**	0.62	0.59	0.56
	RUS	0.77	0.61	0.60	0.56
	ROS	0.77	0.61	0.59	0.56

Looking at the Precision table we see that for both C5.0 Decision Tree and Random Forest Classifier, Random Over Sampling (ROS) gives the best precision. This is expected since Random Over Sampling increases the amount of fraudulent classified data by duplicating the previous smaller data classified as fraudulent. Therefore our learners will predict the duplicated as having the same value. For Logistic Regression we see that SMOTE gives the best precision as the data becomes imbalanced. However as the ratio of fraud and not fraud come closer together, the method used for sampling has a less effect on precision for Linear Regression.

V. CONCLUSION

Healthcare industry is a vital and yet a vulnerable system. Healthcare affects everyone, especially the elderly population, and its importance and value grows everyday. Unfortuanently, this system has been attractive to fraudsters, who commit illegal services and actions for their own financial benefits. It can be said that fraud detection methods should be developed in this field. Fortunately, CMS has publicized various Medicare claims datasets. In this research, we have specifically chosen one particular dataset, Physician and Supplier Dataset and we have compared several different machine learning models in order to detect fraud.

Given that it is a fraud dataset, we have faced many problems that are unique to fraud. Class imbalance, Big Data, the dynamic techniques of fraudsters are several of these problems. In our study, we have discussed in depth, these problems and their potential solutions. Also, skewness in the dataset has tempted us to use sampling methods. As a result, this research is unique in a way that it merges different sampling methods with different learner methods. It is our intuition that in situations with class imbalance, comparing learner algorithms are not sufficient. Sampling methods should also be evaluated in order to obtain better results. During our research, and in our demonstrations, we have seen that the performance of a learner is affected by sampling methods, and sampling ratio.
In our research we obtained the best AUC score using Random Forest classifier with a Random Under Sampler. When taking account F1-score as our performance measure we obtained the best result with Random Forest Classifier using SMOTE. An interesting observation is that Random Forest got a better performance as it approached a class distribution ratio of 50:50, despite what Bauder and Khoshgoftaar [19] fouund in their research saying that the best performance for RF is obtained with a 90:10 distribution ratio.

Remarks

We aimed this paper to be an informative and detailed paper. Thus, we did not refrain from explaining our learners, samplers, datasets, and other information regarding our research in depth. Also, we aimed to include a diverse set of learners and samplers. This was a challengenging research for our group, but we have learned a lot and gained experience about machine learning concepts.

Future Works

Continued research on this topic includes obtaining better datasets for fraudulent transactions, using neural networks as classifiers and training separate classifiers for each physician profession. In [7] Herland shows that combining and using all three Centers for Medicare and Medicaid Services (CMS), which include Part B, Part D, and DMEPOS. In our research we only used Part B data set, namely the Medicare Provider Utilization and Payment Data: Physician and Other Supplier. So by using a more comprehensive dataset better results can be obtained. In [15], Johnson and Khoshgoftaar show that using deep neural networks to classify medical fraud, gives better performance (AUC score) compared to using conventional machine learning algorithms. In [7] Herland shows that by making a separate classifier for each profession, the performance of detecting fraud increases by a significant amount for some professions. By using separate neural networks for different professions for the purposes of identifying medical fraud, there might be a significant performance increase in fraud detection. Further could be done on this topic.

REFERENCES

[1] "National Health Expenditures 2016 Highlights", Centers for Medicare & Medicaid Services, 2018.

[2] L. Morris, "Combating Fraud In Health Care: An Essential Component Of Any Cost Containment Strategy," 2009. [Online]. Available: http://content.healthaffairs.org/content/28/5/1351.

[3] Pande and W. Maas, "Physician Medicare fraud: characteristics and consequences, "International Journal of Pharmaceutical and Healthcare Marketing, vol. 7, no. 1, pp. 8–33, 2013

[4] Sohony, I., Pratap, R., & Nambiar, U. (2018). Ensemble learning for credit card fraud detection. doi: 10.1145/3152494.3156815

[5] Junqué, E., Stankova, M., Moeyersoms, J., Minnaert, B., Provost, F., & Martens, D. (2014). Corporate residence fraud detection. KDD '14 Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1650–1659. doi: 10.1145/2623330.2623333

[6] Behdad, M., French, T., Barone, L., & Bennamoun, M. (2010). On the problems of using learning classifier systems for fraud detection. GECCO '10 Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation , 1067–1068.

[7] Piotr Juszczak, Niall M. Adams, David J. Hand, Christopher Whitrow, and David J.Weston. Off-the-peg and bespoke classifiers for fraud detection. Computational Statistics and Data Analysis, 52(9): pp: 4521 – 4532, 2008

[8] C. Phua, V. C. S. Lee, K. Smith-Miles, and R. W. Gayler. A comprehensive survey of data mining-based fraud detection research. CoRR, abs/1009.6119, 2010.

[9] Bauder, R. A., & Khoshgoftaar, T. M. (2018). The Detection of Medicare Fraud Using Machine Learning Methods with Excluded Provider Labels. The Thirty-First International Florida Artificial Intelligence Research Society Conference (FLAIRS-31) 405.

[10] More, A. (2016). Survey of resampling techniques for improving classification performance in unbalanced datasets.

[11] Bauder, R., Rosa, R. da, & Khoshgoftaar, T. (2018). Identifying Medicare Provider Fraud with Unsupervised Machine Learning. 2018 IEEE International Conference on Information Reuse and Integration (IRI). doi: 10.1109/IRI.2018.00051

[12] Sadiq, S.; Tao, Y.; Yan, Y.; and Shyu, M.-L. 2017. Mining anomalies in Medicare big data using patient rule induction method. In Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on, 185–192. IEEE.

[13] J. S. Ko, H. Chalfin, B. J. Trock, Z. Feng, E. Humphreys, S.-W. Park, H. B. Carter, K. D. Frick, M. Han, "Variability in Medicare utilization and payment among urologists", Urology, vol. 85, no. 5, pp. 1045-1051, 2015.

[14] Bauder, R. A., Khoshgoftaar, T. M., & Hasanin, T. (2018). Data Sampling Approaches with Severely Imbalanced Big Data for Medicare Fraud Detection. 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI). doi: 10.1109/ICTAI.2018.00030

[15] Johnson, J., & Khoshgoftaar, T. (2019). Deep Learning and Data Sampling with Imbalanced Big Data. 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI). doi: 10.1109/IRI.2019.00038

[16] Herland, M., Bauder, R., & Khoshgoftaar , T. (2019). The effects of class rarity on the evaluation of supervised healthcare fraud detection models.

[17] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. IEEE, 2008, pp. 1322–1328

[18] Branting, K., Reeder, F., Gold, J., & Champney, T. (2016). Graph analytics for healthcare fraud risk estimation. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). doi: 10.1109/ASONAM.2016.7752336

[19] Bauder, R., & Khoshgoftaar, T. (2018). Medicare Fraud Detection Using Random Forest with Class Imbalanced Big Data. 2018 IEEE International Conference on Information Reuse and Integration (IRI). doi: 10.1109/IRI.2018.00019

[20] CMS Office of Enterprise Data and Analytics. 2017. Medicare Provider Utilization and Payment Data: Physician and Other Supplier. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html

[21] Herland, Matthew Andrew. (2019) Big Data Analytics and Engineering for Medical Fraud Detection

[22] OIG. Office of Inspector General Exclusion Authorities.

[23] Herland, M., Khoshgoftaar K, T. M., & Bauder, R. A. (2018). Big Data fraud detection using multiple Medicare data sources. Journal of Big Data.

[24] 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).

[25] Python Sci-kit Learn library. https://scikit-learn.org/stable/modules/tree.html