Fraud Detection with a Limited Number of Known Fraudulent Medicare Providers

Richard A. Bauder, Taghi M. Khoshgoftaar, Amri Napolitano
College of Engineering & Computer Science
Florida Atlantic University
rbauder2014@fau.edu, khoshgof@fau.edu, amrifau@gmail.com

Abstract

Medical claims fraud is a major contributor to increased healthcare costs, but the negative impact can be lessened through effective fraud detection. In this paper, we combine Medicare provider utilization and payment data from 2012 to 2015 with corresponding fraud labels from the List of Excluded Individuals/Entities (LEIE) database. We demonstrate the effectiveness of detecting Medicare fraud with a limited number of known perpetrators, leading to severe class imbalance. For each of the three selected specialties, we use random undersampling to create four class distributions. Random Forest and Logistic Regression learners are built and evaluated based on fraud detection performance. Good fraud detection is demonstrated through the use of random undersampling, across three selected medical specialties. Statistically significant results are seen across the class distributions, with the 80:20 distribution having the best results. Overall, Random Forest (with either 100 or 500 trees), for each class distribution across all specialties, significantly outperforms Logistic Regression, with average AUC scores of 0.881 and 0.88, respectively.

Keywords: Medicare fraud detection, class imbalance, random undersampling, ensemble learning

Introduction

Healthcare in the U.S. is a vastly complex and costly system significantly impacting the well-being of the general populace, with increasing importance to the health and welfare of the elderly population. With better medical practices contributing to longer overall life expectancy, the elderly population rose 28% from 2004 to 2015, versus an increase of just 6.5% for those under 65 years of age (ACL 2015; Kane 2013). This increase in the elderly population continues to burden the current U.S. healthcare system. In particular, the U.S. Medicare program, with its focus on providing medical insurance for people over the age of 65, is affected by these changing demographics and currently has over 54.3 million beneficiaries (CMS 2017b). Medicare contributes to 20% of the overall U.S. healthcare spending, at $646 billion in 2015 (CMS 2015). With these increasing financial resources and program complexity, Medicare is replete with potential fraudulent activities. The Federal Bureau of Investigation (FBI) estimates that fraud accounts for 3-10% of all medical claims (Morris 2009), which is $19 to $65 billion per year in potential losses. More information on healthcare fraud can be found in (Bauder, Khoshgoftaar, and Seliya 2017).

In order to better combat Medicare fraud, our research focuses on the detection of Medicare fraud using a limited number of known fraudulent providers. To demonstrate the detection of fraudulent provider claims, we combine the 2012 to 2015 Medicare Provider Utilization and Payment Data: Physician and Other Supplier (CMS 2017a) data, which includes medical claims information by physicians and other provider types, or specialties, such as nurses. This particular Medicare dataset is released approximately every two years, which creates a lag of two to four years in research that can utilize the entire dataset. Fraud labels are obtained by incorporating exclusions from the Office of Inspector General’s (OIG) List of Excluded Individuals/Entities (LEIE) database (LEIE 2017). We merge the LEIE exclusion labels with the Medicare data by using each provider’s unique National Provider Identification (NPI) number, taking into account exclusion start and end periods to avoid overlap and potential double counting of fraud labels. 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 to practice medicine and 21% were not suspended from medical practice despite their convictions (Pande and Maas 2013).

Even with the addition of these labels, the number of excluded providers is very small creating severe class imbalance between non-fraud and fraud instances. Because of this class imbalance, random undersampling (RUS) is performed on three provider types, or specialties, for 50:50, 65:35, 70:30, and 80:20 (majority:minority) class distributions, and repeated 10 times to reduce sampling bias and better represent the non-fraudulent (majority) class. RUS was chosen over oversampling methods because oversampling would increase the data substantially and make it more difficult to use. Fraud detection is evaluated on an ensemble learner, Random Forest, and Logistic Regression for each of the four class distributions. We chose Random Forest since it is an ensemble learner and known to perform well with
classification tasks (Fernández-Delgado et al. 2014). Logistic regression was chosen as a comparative learner due to its popularity, simplicity, and effectiveness. To assess performance, we use the Area Under the receiver operating characteristic Curve (AUC) followed by hypothesis testing to determine statistical significance of our results. We show that the Random Forest ensemble learner (with 100 and 500 trees) performs the best with average AUC values of 0.881 and 0.88, respectively, across the three Medicare provider types. To the best of our knowledge, this is the first work to provide such a study focusing on the problem of class imbalance as well as demonstrating the benefits of ensemble learning for the detection of Medicare fraud.

The remainder of the paper is organized as follows. The Related Works section discusses works related to the current research. In the next section, Methodology, we detail our research methodology to include the Medicare and LEIE datasets, class imbalance, learners, performance metrics, and experimental design and evaluation. In the Discussion and Results section, we explain the research results. Finally, we summarize our conclusions and future work.

Related Works

Many studies can be found applying machine learning techniques for healthcare-related concerns, but for our purposes, we limit the related works to Medicare fraud detection and class imbalance. The interested reader can find additional information on general healthcare fraud detection in (Joudaki et al. 2015).

Khurjekar et al. (Khurjekar, Chou, and Khasawneh 2015) propose a two-step unsupervised approach to detecting fraud using the 2012 Medicare data. The authors first use the residuals from a multivariate regression model, with average payment as the dependent variable, to identify suspicious claims based on a residual threshold of $500. The second part incorporates these residuals using clustering to find fraudulent observations based on average cluster distances. Chandola et al. (Chandola, Sukumar, and Schryver 2013), in a preliminary study, employ several different data mining techniques to detect fraud in Medicare claims data and private provider enrollment data. The authors employ several different data mining techniques. Using features derived from the temporal analysis, the authors build a logistic regression model to detect known fraudulent cases using labeled data from the Texas Office of Inspector General’s exclusion database only. The details of data integration are limited with regards to mapping of fraud labels to the Medicare data. Several papers from our research group explore Medicare fraud detection using machine learning techniques. In one study (Bauder and Khoshgoftaar 2017), we incorporate a two-step approach in detecting Medicare fraud, per provider type. Their method uses a multivariate regression model from which the residuals are passed into a Bayesian probability model. In another study by our group (Bauder and Khoshgoftaar 2016), we use multivariate regression to establish a baseline for expected Medicare payments, per provider type, to compare against actual payments to flag possible fraud.

A study by our research group is one of two known studies addressing class imbalance with LEIE database labels to detect Medicare fraud. Our work, by Herland et al. (Herland, Bauder, and Khoshgoftaar 2017), continue the research efforts of previous studies (Bauder et al. 2016) to detect fraud through the prediction of a physician’s medical specialty. We use the 2013 Florida-only Medicare dataset with labels from the LEIE database. The effects of class imbalance were evaluated using both random undersampling and Synthetic Minority Over-sampling Technique (SMOTE). In another study by Branting et al. (Branting et al. 2016), the authors create a graph of providers, drug prescriptions, and procedures. They use two algorithms where one calculates the similarity to known fraudulent and non-fraudulent providers, and the other estimates fraud risk via shared practice locations. Medicare data from 2012 to 2014 was used with 12,153 excluded providers from the LEIE database. To address class imbalance, the authors used a 50:50 class distribution by retaining 12,000 known excluded providers and randomly selecting 12,000 non-excluded providers. Fraud detection was assessed using a J48 decision tree build with 11 graph-based features and 10-fold cross-validation.

Our study differs from the related works in its comprehensive nature. We use the entire Medicare dataset from 2012-2015 and do not filter by specific locations and combine the Medicare data with the LEIE excluded providers. We take into account exclusion limits and overlapping time frames to avoid over labeling. Two different learners are assessed in our experiments for different provider types and class distributions, focusing on the benefits of ensemble learning. We use four different ratios, rather than a single 50:50 ratio, performing 5-fold cross-validation with 10 repeats to avoid bias due to random selections.

Methodology

In this section, we outline the data used and discuss the integration of the Medicare and LEIE data sources. We also discuss the two learners, performance metric, and class imbalance. Lastly, we summarize our study’s experimental design and evaluation approach.

Data

The Medicare Provider Utilization and Payment Data: Physician and Other Supplier, from 2012 to 2015, describes payment and utilization claims data, with information on services and procedures provided to Medicare beneficiaries, recorded after completed claims payments. The data were compiled and aggregated by CMS, grouping claims information by National Provider Identifier (NPI), Healthcare Common Procedure Coding System (HCPCS) code, and place of service. The NPI is a unique number for each provider (CMS Office of Enterprise Data and Analytics 2017). The HCPCS code identifies the procedure/service performed (e.g., office visit) and the place of service is either an office or facility, such as a hospital. The combined Medicare dataset has 37,147,213 instances and 30 features, covering 89 provider types or specialties, for 1,080,115 distinct providers. Across the four years of Medicare data, there
are a few notable differences that needed to be addressed in order to create the combined dataset. The standardized payment variables are not included since they only appear in the 2014 and 2015. Similarly, the standard deviation variables were excluded, since they pertain to 2012 and 2013 only. The use of the remaining variables is left as future work. In our study, we use two categorical and five numerical features. The feature exclusion is the class variable that contains the fraud or non-fraud labels. NPI and provider_type are used for identification only and are not part of the model. Table 1 gives descriptions for each of these features.

The excluded providers come from the List of Excluded Individuals/Entities (LEIE) database, which are used to create the fraud labels to build and test the learners. The exclusions are categorized by different exclusion authorities per the Social Security Act, which indicate the severity and exclusion periods (OIG 2017). For our study, we selected the mandatory exclusions indicating more severe punitive actions for fraudulent providers. Specifically, we included the following exclusion authorities’ numbers: 1128(a)(1), 1128(a)(2), 1128(a)(3), 1128(a)(4), 1128(c)(3)(g)(i), and 1128(c)(3)(g)(ii). We use only NPI numbers when matching known provider exclusions. Due to limitations in the LEIE database being at the NPI- or provider-level only, we assume excluded providers (NPI) include all of the corresponding procedures (HCPCS) performed for the exclusion period. Currently, there is no known publicly available dataset which includes fraud labels for both provider and procedure performed. Furthermore, we label providers as fraudulent only during the exclusion period within the available Medicare years. These labels indicate a submission of claims for services by an excluded provider, thus considered fraud under the federal False Claims Act (United States Code 2006).

To map fraud labels to the combined Medicare dataset, we cross-referenced NPI numbers in the Medicare data and LEIE database, matching any providers with past or current exclusions. For periods covering the 2012 to 2015 Medicare dataset, only the 1128(a) rules were found in and used from the LEIE database. This particular rule has a 5-year minimum exclusion period. Note that only annual records are available in the Medicare data. Because of this limitation, we assume an exclusion within any given year indicates all instances in that year are considered fraud. After combining the Medicare dataset and the LEIE exclusion labels, we selected provider types based on the number of exclusions (or fraud labels). Due to space limitations, we elected to use three of the 89 possible provider types with the highest number of fraud labels. Table 2 lists the selected provider types and the number of exclusion and non-exclusion instances (by provider and procedure code), which shows the extent of the class imbalance.

Class Imbalance

The Medicare dataset is a challenging dataset due to the skewed nature of the provider exclusions. With class imbalance (Haixiang et al. 2017), the learner will tend to focus on the majority class (i.e. the class with the majority of instances), which is usually not the class of interest. In our case, the non-fraud labels are the majority class, whereas the fraud labels are the minority class and class of interest. An effective way to compensate for some of the harmful effects of severe class imbalance is by changing the class distribution in the training data to increase the representation of the minority class which can help improve model performance. One sampling method is oversampling, in which classes are balanced by increasing the minority class, whereas undersampling removes samples from the majority class. Oversampling can significantly increase the size of the dataset and potentially overfit the data by making identical copies of the minority class instances. On the contrary, with undersampling, we retain the original fraud instances and randomly sample without replacement from the remaining majority class instances.

<table>
<thead>
<tr>
<th>Provider Type</th>
<th># of Instances</th>
<th># of Exclusions</th>
<th>% of Exclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal Medicine</td>
<td>4,726,499</td>
<td>1,187</td>
<td>0.03%</td>
</tr>
<tr>
<td>Hematology/Oncology</td>
<td>692,155</td>
<td>411</td>
<td>0.06%</td>
</tr>
<tr>
<td>Podiatry</td>
<td>770,259</td>
<td>319</td>
<td>0.04%</td>
</tr>
</tbody>
</table>

Table 1: Description of selected Medicare features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Feature Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>npi</td>
<td>Identification</td>
<td>Unique provider identification number</td>
</tr>
<tr>
<td>provider_type</td>
<td>Identification</td>
<td>Medical provider’s specialty (or practice)</td>
</tr>
<tr>
<td>nppe__provider__gender</td>
<td>Categorical</td>
<td>Provider gender</td>
</tr>
<tr>
<td>hpcps_code</td>
<td>Categorical</td>
<td>Procedure or service performed by the provider</td>
</tr>
<tr>
<td>line__srvc__cnt</td>
<td>Numerical</td>
<td>Number of procedures/services the provider performed</td>
</tr>
<tr>
<td>bene__unique__cnt</td>
<td>Numerical</td>
<td>Number of distinct Medicare beneficiaries receiving the service</td>
</tr>
<tr>
<td>bene__day__srvc__cnt</td>
<td>Numerical</td>
<td>Number of distinct Medicare beneficiary / per day services performed</td>
</tr>
<tr>
<td>average_submitted_chrg_amt</td>
<td>Numerical</td>
<td>Average of the charges that the provider submitted for the service</td>
</tr>
<tr>
<td>average_medicare_payment_amt</td>
<td>Numerical</td>
<td>Average payment made to a provider per claim for the service</td>
</tr>
<tr>
<td>exclusion</td>
<td>Class label</td>
<td>Mapped fraud labels from the LEIE database</td>
</tr>
</tbody>
</table>

Table 2: Exclusions by provider type
In our study, we use random undersampling with the following class distributions (majority:minority): 50:50, 65:35, 75:25, and 80:20. These class ratios retain a reasonable amount of the majority class and reduce loss of information relative to the minority class. We repeat the RUS process 10 times per provider type, for each class distribution, to reduce bias due to poor random draws, and to better represent the majority class through the use of multiple random samples.

Learners

We build two different learners to classify Medicare claims fraud (note that we count Random Forest as one learner with the application of 100 and 500 trees). We employ the Weka software tool (Witten et al. 2016) to build and test each learner, with default parameters unless otherwise stated. Logistic Regression (LR) is a common machine learning method for classification. LR is like linear regression except it employs a sigmoidal function to generate class probabilities (Le Cessie and Van Houwelingen 1992). Random Forest (RF) is an ensemble method that constructs multiple unpruned decision trees which make up the so-called forest (Breiman 2001). The final classification is made by combining the results from the individual trees via majority voting. The algorithm creates random datasets using sampling with replacement to train each of the decision trees and chooses the most discriminating feature between the classes using information entropy. Additionally, RF performs random feature subspace selection, at each node of a tree, where a subset of $m$ features is considered for the decision at that node. In our study, we use 100 and 500 trees, referred to as RF100 and RF500, respectively.

Performance Metric

The classification models are evaluated using the Area Under receiver operating characteristic Curve (AUC) performance metric (Bekkar, Djemaa, and Alitouche 2013). AUC is used to assess the capabilities of binary classification methods. The ROC curve is used to characterize the trade-off between true positive rate and false positive rate, depicting a learner’s performance across all decision thresholds. Perfect classification is indicated by an AUC value of 1, with a range from 0 to 1. Due to the severe class imbalance of the testing data, AUC is used as the performance measure for our experiment (Jeni, Cohn, and De La Torre 2013).

Design and Evaluation

We use 5-fold cross-validation to evaluate the performance of each of the models, per provider type. The reason for 5-fold cross-validation is that since some of the specialties have such a low number of fraud labels (positive class), we wanted to reduce the likelihood that a fold would have too few positive class instances possibly hindering fair performance evaluations. In order to reduce bias due to unlucky random draws, cross-validation is repeated 10 times with the final performance values being the average over the 10 repeats.

Discussion and Results

We build two types of learners, specifically RF with 500 and 100 trees and LR, to detect known fraudulent Medicare providers. For this study, we are interested in the overall detection of fraud in the context of ensemble versus non-ensemble learning and different class distributions. Figure 1 shows the average AUC results per learner. Moreover, the trending of AUC scores across class distributions is depicted indicating how class distributions influence a learner’s fraud detection performance. From the AUC results, the RF500 and RF100 ensemble learners are shown to be the best performers. The reasons for the high performance of the RF500 and RF100 learners may be due to the RF’s ensemble nature and the automatic features selection process, which is not performed implicitly with the default LR learner. Additionally, the 80:20 class ratio has the best overall performance for RF and LR learners, followed by the 75:25, 65:35, and 50:50 class ratios. This indicates that retaining more non-fraud instances, while retaining all of the fraud instances, provides the most information for correct classification.

Because the Medicare dataset includes both categorical and numerical features, how each learner handles this mixture is important in assessing model performance. In our case, Random Forest naturally handles categorical variables and can manage a large number of categorical variables due to the random sampling of features and instances. On the other hand, LR does not naturally handle categorical variables and expects continuous variables, thus these variables must be converted prior to building a model. A common method for mapping categorical to numeric variables is known as one-hot encoding. This translates each of the original categorical values into distinct binary features. In Weka, the method to perform this is the NominalToBinary filter, which is run prior to creating any model with categorical inputs. Even though LR requires the categorical variables to be mapped, these binary features are easily processed because regression works well with Bernoulli distributions.

We provide additional rigor around our experiment by evaluating the statistical significance of our results with a three-factor ANalysis Of Variance (ANOVA) (Gelman 2005) and Tukey’s Honest Significant Difference (HSD) (Tukey 1949) tests, at a 95% confidence level. The ANOVA and corresponding Tukey’s HSD tests used learner, provider type, and class distribution as factors. Table 3 summarizes the results of the ANOVA test, where all factors are shown to be significant. To provide more fidelity on the significance of our results, each of the factors is subject to post hoc analysis via Tukey’s HSD. Table 4 shows average performance for each of the factors. As discussed previously, the RF learners perform best overall being significantly better than LR. Additionally, the 80:20 class distribution shows significantly better performance than the other class distributions. Because RF500 and RF100 show no statistically significant difference, RF100 is recommended since less time needed to train each model.

Furthermore, based on the average AUC across learners and class distributions, Hematology/Oncology is significantly easier to learn than the other Medicare specialties, whereas Podiatry is the most difficult to learn. Part of the...
reason for these differences in performance among special-

<table>
<thead>
<tr>
<th>Groups</th>
<th>Learner</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>RF500</td>
<td>0.881</td>
</tr>
<tr>
<td>a</td>
<td>RF100</td>
<td>0.880</td>
</tr>
<tr>
<td>b</td>
<td>LR</td>
<td>0.782</td>
</tr>
</tbody>
</table>

(a) Learners

<table>
<thead>
<tr>
<th>Groups</th>
<th>Class Distribution</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>80:20</td>
<td>0.863</td>
</tr>
<tr>
<td>b</td>
<td>75:25</td>
<td>0.857</td>
</tr>
<tr>
<td>c</td>
<td>65:35</td>
<td>0.846</td>
</tr>
<tr>
<td>d</td>
<td>50:50</td>
<td>0.824</td>
</tr>
</tbody>
</table>

(b) Class Distributions

<table>
<thead>
<tr>
<th>Groups</th>
<th>Provider Type</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Hematology/Oncology</td>
<td>0.860</td>
</tr>
<tr>
<td>b</td>
<td>Internal Medicine</td>
<td>0.848</td>
</tr>
<tr>
<td>c</td>
<td>Podiatry</td>
<td>0.835</td>
</tr>
</tbody>
</table>

(c) Provider Types

ties can be attributed to the distribution of HCPCS codes. The HCPCS codes, as one-hot encoded features, may not provide enough information to a model to help discriminate one specialty from another, and thus contribute to model noise and poor performance. For future work, we will explore feature selection to determine the importance of the Medicare features, to include the HCPCS codes.

Conclusion

Programs such as Medicare, already limited in their ability to provide effective and affordable care, are adversely impacted by fraudulent activities. In this study, we use machine learning to detect fraudulent Medicare providers. The Medicare data is labeled with fraud labels from the excluded providers found in the LEIE database, and we evaluate fraud detection performance across three diverse specialties. We explore two different machine learning methods and compare performance results, with statistical significance, across four different class distributions. Our results indicate that the RF ensemble learner is the best performer with AUC scores significantly higher than LR. Further, the 80:20 class distribution provides the best results across all specialties, for both learners. Even with the extremely limited number of known fraudulent providers, we find that the detection of Medicare fraud is promising. Future work will involve building and testing additional learners and adding more
Medicare specialties. We also intend to aggregate the Medicare data, with LEIE labels, to the NPI-level to assess model fraud detection performance.

Acknowledgment
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.

References
CMS Office of Enterprise Data and Analytics. 2017. Medicare Fee-For-Service Provider Utilization & Payment Data Physician and Other Supplier.
CMS. 2017b. US Medicare Program.