**DATA-698 Research Project**

Title – to be provided later…

**Mid-Term Draft** Simon Ustoyev

Abstract

To be provided later…

Introduction

Fraud detection, and particularly healthcare insurance fraud detection, has shown to be a very important, highly demanded and a challenging problem to solve for Data Science and Analytics. This particular research study will focus on fraud detection in Medicare insurance.

The Federal Medicare spending is growing substantially due to medical expenses and demographic factors, such as an increase in elderly population receiving the coverage. It is not surprising then that there will be less tolerance for fraud and a more growing need for better detection. “*The Federal Bureau of Investigation (FBI) estimates that fraud accounts for 3–10% of healthcare costs, totaling between $19 billion and $65 billion in financial loss per year. Medicare accounts for 20% of all U.S. healthcare spending [8] with a total possible cost recovery (with the potential application of effective fraud detection methods) of $3.8 to $13 billion from Medicare alone.* **[[[1]](#endnote-1)]”**

Several studies have been done to tackle this problem using various data science methods with good and promising results. It is a goal of this research paper to try to use some of these prior studies as a starting point to build upon and then attempt to introduce some novelty into the process that may show some improvement in fraud detection results.

Related Research

This research project would mainly consider and follow two prior research publications from the “Journal of Big Data”. The first publication is from 2018, “[Big Data fraud detection using multiple Medicare data sources](https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-018-0138-3)”. This paper is mainly exploring the application and evaluation of traditional modelling techniques, such as Logistic Regression, Gradient Boosted Trees and Random Forest. The second paper is a later study in 2019, which appears to be an extension of the first (as both papers share at least one same author). In this paper, named, “[Medicare fraud detection using neural networks](https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-019-0225-0)”, the researchers explore the use of Deep Neural Networks (DNN), as well as using better data sampling choices to address the drastic class-imbalance issue in the data.

Both papers combine the (yearly) time-series data of payments to providers for various services obtained from “*Centers for Medicare and Medicaid Services”* ([CMS](https://www.cms.gov/)), with the data from “*Office of Inspector General*” ([OIG](https://oig.hhs.gov/)) that publishes a registry of providers who were convicted of fraud. The combined data becomes a labelled dataset to serve as the basis for supervised learning techniques for building various prediction models.

The 2018 paper, explored prediction outcomes based on a wider set of Medicare datasets listed below:

1. **Part B** - Medicare Provider Utilization and Payment Data: Physician and Other Supplier
2. **Part D** – Medicare Provider Utilization and Payment Data: Part D Prescriber
3. **DMEPOS** – Medicare Provider Utilization and Payment Data: Referring Durable Medical Equipment, Prosthetics, Orthotics and Supplies
4. **Combined** dataset – a combination of the three above datasets
5. **LEIE** - provides fraud labels (i.e. class labels on Providers), using the List of Excluded Individuals and Entities (LEIE) from the Office of the Inspector General (OIG)

Based on the Area under the Receiver Operating Characteristic (ROC) Curve performance metric, the results of this paper show that the Combined dataset with the Logistic Regression (LR) learner yielded the best overall score at 0.816.

In contract, the 2019 paper, used only Part B and LEIE datasets. However, it explored and compared six deep learning methods for addressing class imbalance with a minority class size of 0.03%. The six methods are comprised of three data-level and three algorithm-level methods. The data-level techniques include random over-sampling (ROS), random under-sampling (RUS), and a hybrid ROS–RUS. The algorithm-level techniques include a cost-sensitive loss function, the Focal Loss, and the Mean False Error Loss. Results of this study indicate that eliminating class imbalance from the training data through ROS or ROS–RUS produces significantly better AUC scores than all other methods, i.e. 0.8509 and 0.8505. While ROS methods perform best using the 50:50 class ratio, plain RUS outperforms baseline methods and achieves its highest AUC score with a 99:1 class ratio.

New Hypotheses

The above-mentioned studies are instructive, achieving good prediction results and can serve to be good baseline and guidance for further research and enhancements. The data used in these studies ranged from years 2012 to 2016 and combined datasets, published by CMS for each year on payments and service aggregated for that year. Simply put, the observations can be assigned a “Year” label and can be regarded as a time-series data. This time label, however, was not a factor in the studies.

Furthermore, the studies label all of the payment history (for a given provider) either with 0 (non-fraudulent) or 1 (fraudulent). In other words, there is an assumption that all of provider’s historical activity, prior to a fraud conviction, is deemed as fraudulent activity.

The focal point of this project is to challenge that assumption and offer an alternative one by making use of this time component and see if this will result in any [statistically] significant improvements for prediction rates. The claim made here is that, for a fraudulent provider, not all of activity should be regarded as fraudulent or at least equally fraudulent across all of the years of practice under consideration. It may be fair to say that most, if not all, providers start off with being compliant and that it takes time to build up practice and eventually end up to commit and perpetuate fraud. Therefore, one would ***hypothesize*** that in the case of a fraudulent provider, one can improve prediction rates by regarding more recent activity with greater weight in model training rather than the more distant activity. To put in other words, for fraudulent cases, more attention (or suspicion) should be given to more recent activity rather than to more distant activity.

Datasets and Variables

The datasets used in this study are similar to the referenced 2018 study, however covering a more recent 5-year period from years 2013 to 2017. The below provides a listing of these datasets, along with a brief description:

* [CMS](https://www.cms.gov/) datasets (Part B, Part D and, DMEPOS).
  + **Part B** dataset provides claims information for each procedure a physician performs withing a given year. [[[2]](#endnote-2)]
  + **Part D** dataset provides information pertaining to the prescription drugs that providers administer within a given year. [[[3]](#endnote-3)]
  + **DMEPOS** dataset provides claims information about Medical Equipment, Prosthetics, Orthotics and Supplies that providers referred patients to either purchase or rent from a supplier within a given year. [[[4]](#endnote-4)]
* [OIG](https://oig.hhs.gov/) *“List of Excluded Individuals and Entities”* (LEIE)
  + **LEIE** dataset contains the following information: reason for exclusion, date of exclusion and reinstate/waiver date for all current physicians found unsuited to practice medicine and thus excluded from practicing in the United States for a given period of time. [[[5]](#endnote-5)]
* Combined dataset
  + The Combined dataset is created after processing Part B, Part D, and the DMEPOS datasets, containing all the attributes from each, along with the fraud labels derived from the LEIE. The combining process involves a join operation on NPI, provider type, and year. Due to there not being a gender variable present in the Part D data, this variable was not included in the combined set.

The tables below detail the features chosen from the datasets, including a description and feature type (numerical or categorical) along with the exclusion attribute (fraud label) derived from the LEIE dataset. The three CMS datasets in their unaltered/original form provide aggregated data at a more granular level – by NPI and HCPCS(service)/drug level. However, in order to meet the need of mapping fraud labels using the LEIE and properly combining the data, these datasets were further grouped and aggregated at NPI (provider) level.

**Variables from Part B dataset**

|  |  |  |
| --- | --- | --- |
| **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 |

**Variables from Part D dataset**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Description** | **Type** |
| Npi | Unique provider identification number | Categorical |
| Specialty\_description | Medical provider’s specialty (or practice) | Categorical |
| Bene\_count | Number of distinct Medicare beneficiaries receiving the drug | Numerical |
| Total\_claim\_count | Number of drugs the provider administered | Numerical |
| Total\_30\_day\_fill\_count | Number of standardized 30-day fills | Numerical |
| Total\_day\_supply | Number of day’s supply | Numerical |
| Total\_drug\_cost | Cost paid for all associated claims | Numerical |
| Exclusion | Fraud labels from the LEIE dataset | Categorical |

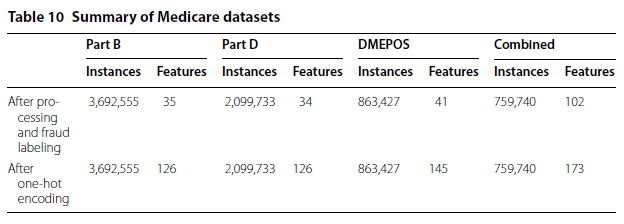
**Variables from DMEPOS dataset**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Description** | **Type** |
| Referring\_npi | Unique provider identification number | Categorical |
| Referring\_provider\_type | Medical provider’s specialty (or practice) | Categorical |
| Referring\_provider\_gender | Provider’s gender | Categorical |
| Number\_of\_suppliers | Number of suppliers used by provider | Numerical |
| Number\_of\_supplier\_beneficiaries | Number of beneficiaries associated by the supplier | Numerical |
| Number\_of\_supplier\_claims | Number of claims submitted by a supplier from a referring order | Numerical |
| Number\_of\_supplier\_services | Number of services/products rendered by a supplier | Numerical |
| Avg\_supplier\_submitted\_charge | Average payment submitted by a supplier | Numerical |
| Avg\_supplier\_medicare\_pmt\_amt | Average payment awarded to suppliers | Numerical |
| Exclusion | Fraud labels from the LEIE dataset | Categorical |

**One-hot encoding**

As was done in the prior studies, in order to build models with numerical and categorical features in the data, on-hot encoding technique is used to transform categorical features into “*dummy*” variables where a value of 1 will be assigned if a categorical feature is present in a given observation and 0 if it is not. This transformation process is also necessary in order to employ ROS/RUS techniques for addressing class imbalance in the data. Also, note that NPI is (obviously) not used for building models and is removed from each dataset after this transformational step.

Note: the below table is to be replaced with results of this study.

****

Analysis Methods

**Learners**

Just as with data selection and preparation approaches, this study will also employ the same analytical methods as the ones in the 2018 study. Namely, the following learners are used here as well:

* Logistic Regression (LR)
* Gradient Boosted Trees (GBT)
* Random Forest (RF)

The goal of the first phase of this research is to establish a baseline against which the offered hypothesis will be tested against. Specifically, the analysis starts off with preparing the data and then training the above-mentioned learners against the portion of the data allocated for training. The training portion will consist of 80% of randomly selected data, leaving 20% for the final model testing and performance results.

**Class imbalance**

Unlike the 2018 study (but more like the 2019 one), this study employs the ROS/RUS random sampling techniques in order to address the drastic class imbalance in the data. As was shown in prior study, ROS results in more accurate results as it doesn’t remove any existing data, but rather increases the instances of the minority class. RUS and combined ROS-RUS techniques are effective and useful in trying to speed up the model execution time. Unlike the prior study, however, this study will simply settle on a 60%-40% split so that the minority (fraud label) class will be expanded to represent the 40% portion, while the majority class left in the remaining 60% still signifies to some degree the imbalance inherent in the data.

**Cross-validation**

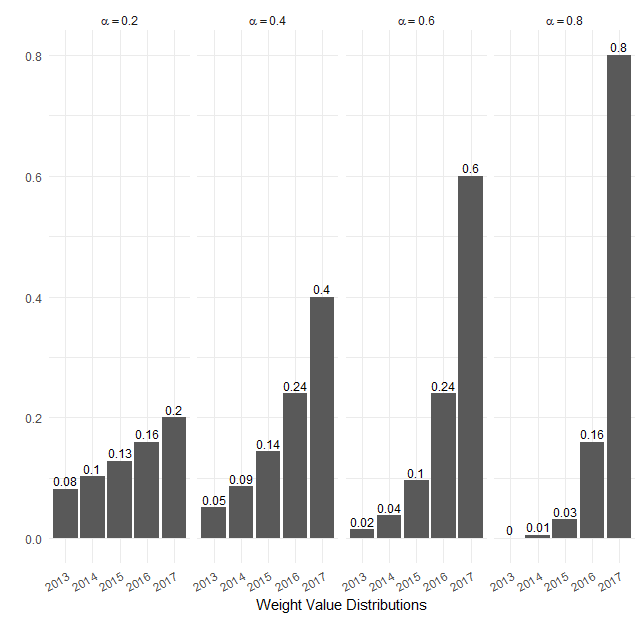
To train and evaluate the models, this study (as was done in the prior ones as well), employs k-fold cross-validation technique with k = 5. This means that the training data is evenly divided into five folds where four folds will be used for training the model and the remaining fold tests the model. This process is repeated 5 times allowing each fold an opportunity as the test fold, ensuring the entire dataset is fully leveraged being used in training and validation. In addition to that, this modeling process will be repeated 10 times helping to reduce bias due to possible bad random draws when creating the folds.

**Hypothesis testing**

After establishing the “*baseline*” metrics in the first phase of the analysis, the second phase introduces the hypothesis proposition and repeats the modeling process by introducing weights vector to the models. As mentioned before, each observation is labeled with the year on which the observation was for. The weights vector essentially replaces the year label with a numerical value so that past years have smaller values compared to more resent years. The distribution of these weight values follows Geometric distribution (or decay) curve across the time period, so that most recent cases will be labelled with a value closer to 1 and more distant cases with values closer to 0. The tables and graphics below demonstrate the weight value assignments for each year label. Selecting different alpha () parameter for the geometric distribution can result in generating different set of weight values to be tried.

**Table of weight vectors for various alpha ()**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **∝=0.2** | **∝=0.4** | **∝=0.6** | **∝=0.8** |
| 2017 | 0.2 | 0.4 | 0.6 | 0.8 |
| 2016 | 0.16 | 0.24 | 0.24 | 0.16 |
| 2015 | 0.128 | 0.144 | 0.096 | 0.032 |
| 2014 | 0.1024 | 0.0864 | 0.0384 | 0.0064 |
| 2013 | 0.08192 | 0.05184 | 0.01536 | 0.00128 |



One of the goals of this modeling and evaluation process is repeat the process for different set of weight vectors and to therefore determine an optimal value the alpha () parameter.

**Performance metric**

In assessing fraud, we are presented with a two-class classification problem where, in this case, a physician is either fraudulent or non-fraudulent. The positive class, or class of interest, is fraud and the negative class is non-fraud. Confusion matrix for each model is commonly used to assess the performance of learners. Confusion matrices provide counts comparing actual counts against predicted counts. From the resultant matrices, AUC score is calculated and used to measure fraud detection performance. AUC is the Area under the Receiver Operating Characteristic (ROC) curve, where ROC is the comparison between false positive and true positive rates. The AUC results in a single value ranging from 0 to 1, where a perfect classifier results in an AUC of 1, an AUC of 0.5 is equivalent to random guessing and less than 0.5 demonstrates bias towards a given class.

**Significance testing**

ANOVA statistical tests will be performed on AUC performance results in order to test the hypothesis and attempt to show statistical significance for any improvements in precision.

It is important to note that for an insurance company it is important to minimize false alarms, as it implies the cost and effort for getting a team of investigators to get involved for each positive prediction. In other words, an insurance company in a fraud detection analysis cares more for precision of predicted results.

Results and discussion

To be provided later…

Conclusion

To be provided later…

1. <https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-018-0138-3> [↑](#endnote-ref-1)
2. <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier> [↑](#endnote-ref-2)
3. <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber> [↑](#endnote-ref-3)
4. <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/DME> [↑](#endnote-ref-4)
5. <https://oig.hhs.gov/exclusions/authorities.asp> [↑](#endnote-ref-5)