LITERATURE SURVEY

CS226 : BIG DATA MANAGEMENT : FALL 2023

PROJECT TITLE

Healthcare Fraud Detection

Team Hector

Rohith Reddy Kancharakuntla
Student Id: 862466402
rkanc004@ucr.edu

Devaki Kalyan Chandra Yadav Podila
Student Id: 862468795
pyada015@ucr.edu

Shubham Mishra
Student Id: 862467767
smish040@ucr.edu

Omkar Kadam
Student Id: 862467471
okada001@ucr.edu

Yash Kathe
Student Id: 862464930
ykath001@ucr.edu

Contents

1. Big Data Fraud Detection Work

Paper 1: Prescription Fraud Detection through Statistic Modeling

Authors: Hongxiang Zhang, Lizhen Wang

Introduction

The paper "Prescription Fraud Detection through Statistic Modeling" delves into the analytical techniques for identifying fraudulent activities within prescription data.

Work and Methodology

The paper tries to address the identification of anomalous patterns that are often associated with fraudulent prescriptions. By using statistical models, it tells how data, when methodically analyzed, can reveal inconsistencies that would usually go undetected. It discusses various statistical methods that have been historically used to analyze prescription data, offering insights into the typical behaviors of fraudulent activities.

- Identify a typical prescribing behavior that may suggest fraudulent activity.
- Detect payment irregularities that could be indicative of corrupt practices.
- Establish baselines for normal provider activity, simplifying the identification of outliers that might represent fraudulent actions.

The paper focuses on identifying the anomalies and patterns. paper deals with the statistical side of fraud detection

Conclusion

The paper concludes by telling the power of statistical analysis in fraud detection and also the insights gained from those which play a crucial role.

Paper 2: Medicare Fraud Detection using Random Forest with Class Imbalanced Big Data

Authors: Richard A. Bauder, Taghi M. Khoshgoftaar

Introduction:

The paper tries to find Medicare fraud by using the Random Forest algorithm which has class imbalanced data.

Work and Methodology:

The paper focuses on balancing the data by employing Synthetic Minority Oversampling Technique before applying the machine learning model. Author highlights Random Forest's effectiveness in feature selection and fraud pattern recognition. Furthermore, the paper talks about techniques which can enhance fraud detection systems, which can provide valuable insights into feature importance for detecting fraud. And also showcases the potential for improved real-time fraud detection, which in fact helps in cost savings and increased efficiency in fraud detection.

Conclusion:

Paper demonstrates the viability of using advanced machine learning techniques, specifically Random Forest, it showcases how balanced data can improve and enhance the identification and prevention of fraud in healthcare systems.

Paper 3: Detecting medical prescriptions suspected of fraud using an unsupervised data mining algorithm

Authors: Mohammad Haddad Soleymani Mehdi Yaseri, Farshad Farzadfar, Adel Mohammadpour, Farshad Sharif, Mohammad Javad Kabir

Introduction:

This paper explores the application of an unsupervised data mining algorithm to identify potential fraudulent patterns within Medicare prescription records. Paper aims to find a viable solution for identifying irregularities without prior labeling of data.

Work and Methodology:

The paper focuses on identifying anomalies within prescription claims that may indicate fraudulent activity, rather than relying on pre-labeled instances of fraud, which are often scarce or unreliable. The algorithm examines data characteristics such as unusual prescription volumes or atypical patient-provider interactions. This method has a significant advantage in cases where confirmed instances of fraud are not readily available in the dataset. Furthermore, the paper emphasizes the importance of feature analysis, which includes factors such as geographical trends and prescription patterns,

Conclusion:

The paper concludes by giving us a model(unsupervised) for identifying fraudulent prescriptions, and provides a way that can be adapted to large-scale healthcare fraud detection. The insights offered help in using machine learning in recognizing and preventing fraudulent activities.

Paper 4: Multivariate outlier detection in medicare claims payments applying probabilistic programming methods

Authors: Richard A. Bauder, Taghi M. Khoshgoftaar

Introduction:

The paper introduces a probabilistic model for detecting Medicare fraud, which mainly focuses on multivariate outlier detection. This method is built to overcome the limitations of conventional univariate techniques, providing a more robust analysis of anomalies in healthcare billing data.

Work and Methodology:

Paper employs Multivariate Adaptive Regression Splines (MARS) to model Medicare claims and use Bayesian probability models to find the outliers. Unlike standard practices, it doesn't rely on preset thresholds or distance measures but calculates the probability of a claim being fraudulent.

This model enhances its credibility by incorporating reliable intervals, which offer a range within which an observation likely falls.

The author shows this method can be applied across seven different Medicare specialties, showcasing its flexibility and effectiveness in detecting fraudulent claims with varying degrees of data variability. It showcases that it outperforms methods like the Mahalanobis distance and k-means clustering in robustness.

Conclusion:

The paper concludes by telling how the outlier detection method offers a significant improvement in identifying and investigating potential Medicare fraud. The model's flexibility and meaningful probability distributions for claims make it a powerful tool

Paper 5: A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile

Authors: Pedro A. Ortega, Cristian J. Figueroa, Gonzalo A. Ruz

Introduction:

The paper explains the application of data mining techniques to detect medical fraud or abuse. Focusing on a case study specific to Chile, the paper seeks to develop a robust system to enhance the precision and efficiency of fraud detection in the healthcare domain.

Work and Methodology:

The goal of the paper is to build a system that applies data mining algorithms and to analyze medical claims data. The developed models use clustering, anomaly detection, and predictive modeling, which help in identifying irregularities like fraud or abuse. The paper explores the incorporation of contextual features, including geographical trends and prescription patterns, to further refine fraud detection processes. Paper also emphasizes on Feature selection and engineering which help in isolating pertinent attributes, enabling the accurate identification of fraudulent claims.

Conclusion:

In conclusion, the paper showcases the efficiency of data mining techniques to fight medical claim fraud and abuse. Also demonstrates the viability of the early fraud detection being done by the proposed system.

Paper 6: A prescription fraud detection model

Authors: Karca Duru Aral, Halil Altay Güvenir, Ihsan Sabuncuoglu, Ahmet Ruchan Akar

Introduction:

The paper focuses on the development of an offline batch screening/auditing system and an online real-time transaction control tool, taking into account the interactions between different features in the prescription data.

Work and Methodology:

The authors developed a framework, enabling users to set thresholds for risk matrices, providing a balance between true positive rate and human expert screening time. The offline system processes a database prescribed drug, creating incidence and risk matrices for different feature domains. These matrices are then used to calculate risk assessments. The online tool provides a user-friendly interface for inserting and auditing new prescriptions without re-running the offline code.

The authors conduct extensive testing on the system, achieving sensitivity levels tailored to each domain. The thresholds are refined with the assistance of a medical doctor to ensure meaningful outputs.

Conclusion:

The proposed system demonstrates its effectiveness in detecting prescription fraud, providing a user-friendly interface (i.e., online system) and efficient processing. The system's flexibility in handling both categorical, ordered features, and adaptable to different healthcare practices.

2. Data Selection/Sources

Paper 7: Healthcare Fraud Data Mining Methods: A Look Back and Look Ahead

Authors: Nishamathi Kumaraswamy, Mia K. Markey, Tahir Ekin, Jamie C. Barner, and Karen Rascati

Introduction:

This paper addresses the pressing issue of healthcare fraud in the United States and explores methods for its detection, emphasizing the need for digital advancements to combat this growing societal threat.

Work and Methodology:

This paper discusses healthcare fraud detection which relies on three primary data sources: practitioner data, administrative claims data, and clinical data, forming a comprehensive dataset to detect fraud. Administrative claims data are widely used in research due to their accessibility and standardization.These claims are typically structured based on the CMS 1500 form, which is a standardized template used for processing professional healthcare claims. Moreover, many healthcare fraud detection models make use of synthetic or de-identified data, available from sources like Veterans Affairs TRICARE, Health and Human Services etc. Although these aggregated data extracts offer

valuable insights, implementing fraud detection models based on them can be complex, involving challenges like linking results to specific claim-line level data and identifying responsible providers.

Conclusion:

In conclusion, this paper underscores the pivotal role of administrative claims data in healthcare fraud detection, emphasizing its structured and industry-wide significance. Despite challenges, its use is indispensable for effectively combating healthcare fraud.

Paper 8: Big Data fraud detection using multiple medicare data sources

Author: Matthew Herland, Taghi M. Khoshgoftaar & Richard A. Bauder

Title:

This paper addresses the pressing issue of healthcare fraud, particularly in the context of Medicare, and explores the utilization of advanced statistical methods to detect fraudulent activities by analyzing a combination of comprehensive datasets provided by the Centers for Medicare and Medicaid Services (CMS).

Work and Methodology:

The Medicare datasets selected for the 'Identifying frauds and anomalies in Medicare-B dataset' project are not originally designed for fraud detection; they serve as primary data sources for the CMS to manage healthcare programs and process claims. However, their wealth of healthcare information, covering procedures, drug prescriptions, and medical equipment usage, makes them indispensable for this project. Moreover, these datasets are publicly accessible through the CMS website, regularly updated, and interconnected, providing a comprehensive view of healthcare activities that aids in detecting irregularities. These datasets offer a unique perspective on healthcare fraud by revealing patterns and anomalies in patient journeys. They are more than just chosen; they are essential partners in the critical mission to safeguard the healthcare system from fraud and abuse.

Conclusion:

In conclusion, while the Medicare datasets present challenges in complexity, privacy, and data size, their richness and real-world relevance make them the cornerstone for detecting healthcare fraud and preserving the integrity of the Medicare program.

Paper 9: A survey on statistical methods for health care fraud detection

Author: Gonzalo A Ruz, Cristián J. Figueroa, Pedro A. Ortega

Title:

This paper explores the complexity of healthcare fraud detection, highlighting its financial impact, evolving nature, and the need for robust detection mechanisms in the ever-changing healthcare landscape.

Work and Methodology:

This paper discusses how data selection forms the cornerstone of healthcare fraud detection projects, focusing on the choice of insurance carriers' data sources. These sources encompass governmental health departments and private insurance companies, such as the US Health Care Financing Administration (HCFA), the Bureau of National Health Insurance (NHI) in Taiwan, and the Health Insurance Commission (HIC) in Australia. Insurance claims, at the core of healthcare fraud data, involve a dynamic interplay between insurance subscribers and service providers. The rationale behind selecting insurance carriers' data is simple but profound: they offer a comprehensive view of healthcare transactions, providing the ideal vantage point to detect irregularities. Insurance carriers possess a treasure trove of information concerning healthcare services, costs, and providers. This rich tapestry of data empowers researchers to identify fraudulent patterns and safeguard the integrity of healthcare systems. The selection of insurance carriers' data is far from arbitrary; it's rooted in their data richness and accessibility, enabling the exploration of the intricate landscape of healthcare claims and the unveiling of instances of fraud.

Conclusion:

In conclusion, the selection of insurance carriers' data for healthcare fraud detection is well-founded due to their comprehensive nature, though challenges in data privacy and integration must be addressed.

3. Data cleaning

Paper 10: Approaches for identifying U.S. medicare fraud in provider claims data.

Author: Matthew Herland, Richard A. Bauder, Taghi M. Khoshgoftaar.

Work and Methodology:

The problem discussed in this research paper is the identification of fraud in U.S. Medicare provider claims data. Specifically, the paper focuses on the challenges and approaches for detecting fraudulent activities within the claims submitted by healthcare providers to the U.S. Medicare program. The key issue addressed is the need to develop effective methods and techniques for distinguishing between legitimate claims and those that are fraudulent, aiming to reduce financial losses and maintain the integrity of the Medicare system.

To overcome it uses Spark to efficiently clean and prepare large Medicare claims data. Spark's scalability will allow for parallel processing and handling of extensive datasets, ensuring effective data preprocessing and fraud detection.

Relevancy with Project:

This paper tackles the issue of spotting fraud in U.S. Medicare claims to reduce financial losses. It uses Spark for efficient data cleaning and fraud detection, vital for healthcare fraud prevention.

Conclusion:

Medicare fraud is a persistent issue that costs money and impacts service quality. Our study uses data analysis and machine learning to successfully detect fraud in different medical specialties. By predicting a physician's specialty and comparing it to their actual Medicare data, we can spot potential fraud signs, like billing discrepancies. This highlights the role of data cleaning and integration in fraud detection.

Paper 11: The effects of class rarity on the evaluation of supervised healthcare fraud detection models.

Author: Matthew Herland, Richard A. Bauder, Taghi M. Khoshgoftaar.

Introduction:

Detecting healthcare fraud is important to maintain honest healthcare systems and provide better care to patients. We will look at the problems of dealing with imbalanced data and how researchers are trying to find new ways to solve them.

Work and Methodology:

This paper's significance lies in its exploration of class grouping and removal strategies as potential enhancements for healthcare fraud detection. Carried out the thorough monitoring process with reference to the LEIE dataset. However, the results show mixed effectiveness. The primary challenges in this field are handling extensive, imbalanced datasets, where fraudulent cases are a minority among legitimate ones, and dealing with the class rarity issue, making it challenging to build accurate fraud detection models.

In healthcare fraud detection, it groups providers or claims with similar traits to find irregularities more easily. To tackle fraud, we use Apache Spark to create useful data features like billing patterns and provider behavior. It also analyzes data in real time to catch fraud as it happens and quickly update our detection methods, making healthcare systems more secure.

Relevancy with Project:

This paper explores strategies to improve healthcare fraud detection, though the results vary. It uses Apache Spark to create useful data features and detect fraud in real-time, enhancing healthcare system security.

Conclusion:

Here, the healthcare fraud detection is tough due to the large data volume and imbalanced fraud cases. This creates rarity issues. The study focused on real-world datasets and experimented with severe class imbalance and rarity, along with data sampling techniques. Detecting fraud is vital for improving healthcare funding programs like Medicare in the United States.

4. Data transformation

Paper 12: The use of Big Data Analytics in healthcare.

Author: Kornelia Batko, Andrzej Ślęzak**.**

Introduction:

This paper explores how using Big Data Analytics is changing the way healthcare works. It shows how analyzing large amounts of data is improving patient care and helping healthcare providers make better decisions.

Work and Methodology:

The use of Big Data Analytics in healthcare for fraud detection is how to effectively employ Big Data Analytics to detect and prevent healthcare fraud. The paper addresses challenges related to managing and analyzing vast datasets to enhance the accuracy and efficiency of fraud detection methods, ultimately improving the integrity of healthcare systems.

The paper suggests using techniques like spotting unusual patterns, using smart computer programs, recognizing common fraud signs, and predicting potential fraud to detect healthcare fraud effectively. These methods make use of Big Data Analytics to enhance fraud detection in the healthcare sector.

Relevancy with Project:

This paper emphasizes using Big Data Analytics to improve healthcare fraud detection by managing large datasets. It suggests techniques like recognizing patterns and using smart programs to enhance accuracy in identifying fraud, highlighting the relevance of Big Data Analytics in healthcare fraud prevention.

Conclusion:

The quantitative analysis of this research helped determine the use of Big Data Analytics in fraud detection. It was found that in fraud detection, structured and unstructured data from various sources are used, including databases, transactions, emails, documents, devices, and sensors. Facilities apply analytics in administrative, business, and clinical aspects, making data-driven decisions.

Paper 13: Fraud claim detection using Spark.

Author: KARTHIKA,K. P. PORKODI.

Introduction:

This research paper, titled 'Fraud Claim Detection Using Spark,' explores the application of Apache Spark in the detection of fraudulent claims. It delves into methods for efficiently identifying and mitigating fraud within various domains.

Work and Methodology:

This paper addresses the challenges of detecting fraud in various domains, focusing on issues like accurate identification, efficient scalability, real-time detection, algorithm selection, data preprocessing, and model deployment. It proposes solutions using Apache Spark to enhance fraud detection across domains.

Additionally, the paper highlights a significant improvement in data processing and fraud detection. While the existing system takes 22 hours to process daily data, the proposed system powered by Apache Spark

processes real-time data in just 20 minutes. It can also detect global fraud claims using doctor ID information displayed on a dashboard.Furthermore, the proposed system not only processes real-time data faster but also significantly enhances data transformation capabilities, allowing for more efficient fraud detection.

Relevancy with Project:

This paper discusses fraud detection challenges, including healthcare fraud. It presents solutions using Apache Spark, significantly improving data processing, which is crucial for faster and more efficient healthcare fraud detection.

Conclusion:

Apache Spark will be used for handling large amounts of healthcare claims data. It efficiently gets the data ready for analysis and helps identify fraudulent claims accurately. It can also be used to create computer programs that can tell the difference between real and fake healthcare claims. By using Spark, healthcare organizations can find fraud faster, save money, and keep their systems reliable.

5. Feature Engineering

Paper 14: Feature Selection for Classification []

Authors: M. Dash, H. Liu

Introduction:

This paper discusses the importance of feature selection in classification problems and presents a categorization of existing feature selection methods. This paper also benchmarks the datasets with different characteristics used for comparative study.

Work & Methodology:

In this paper four steps of the feature selection process are recognized: generation procedure, evaluation function, stopping criterion, and validation procedure. The generation procedures are grouped into three categories: complete, heuristic, and random, and the evaluation functions into five categories: distance, information, dependence, consistency, and classifier error rate measures. This paper emphasizes the significance of reducing irrelevant and redundant features to improve the efficiency of classifiers. This paper covers the representative methods that are chosen from each category for detailed explanation and discussion via example. These insights help in selecting and implementing the most suitable feature selection techniques for detecting healthcare fraud, ensuring that only the most relevant and valuable data is utilized in the process.

Conclusion:

In conclusion this paper provides the guidelines for applying feature selection methods based on data types and domain characteristics. This paper also paves the way for practitioners who search for suitable methods for solving domain-specific real-world applications. Though most of the paper is very generic in nature, it helps in understanding basic nuances of feature selection for the three different types of datasets we are using for the project.

Paper 15: A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile

Authors: Pedro A. Ortega, Cristi´an J. Figueroa, Gonzalo A. Ruz

Introduction:

This paper discusses an effective fraud detection system based on data mining implemented by a private health insurance company in Chile. The system aims to identify fraudulent and abusive behavior related to medical claims, significantly impacting revenue losses. The proactive use of data mining and neural network-based detection has proven out to be essential in identifying and preventing various forms of fraud linked to medical claims, affiliates, medical professionals, and employers.

Work & Methodology:

The approach mentioned in the paper involves detailed analysis and division of data sets, ensuring bias reduction and thus creating a proactive anti-fraud culture within the health insurance company. The system mentioned in this paper considers historical data related to affiliates, medical professionals, and employers before the actual medical claim review, focusing on predictive detection of potential fraudulent and abusive claims. The paper highlights the significant use of neural network classifiers in predictive detection. The techniques mentioned in this paper such as first domain feature classification and then deleting redundant second domain features and dealing with discriminative features helps in performing better feature engineering on CMS govt data and draw valuable insights.

Conclusion:

The approach mentioned in this paper led to rejecting a significant percentage of medical claims, contributing to a notable reduction in overall costs. The knowledge gained from this system facilitated the identification of various fraud and abuse types, creating a structured fraud taxonomy. There is a significant advantage of early detection in identifying fraudulent and abusive behavior associated with medical claims.

Paper 16: Impact of the composition of feature extraction and class sampling in medicare fraud detection

Authors: Akrity Kumari, Sanjay Kumar Sonbhadra, Narinder Singh Punn, Sonali Agarwal

Introduction:

This paper addresses the challenges due to class imbalances and high dimensionality. The study focuses on feature extraction followed by data sampling, utilizing techniques like autoencoders for feature extraction and SMOTE for addressing class imbalance. The study aims to enhance fraud detection by applying various gradient boosted decision tree-based classifiers.

Work & Methodology:

The work mentioned in this paper identifies that traditional fraud detection techniques, based on pre-established patterns, fall short due to the volume and diversity of fraudulent activities.To rectify the issue machine learning models and dimensionality reduction techniques such as autoencoders are explored to enhance fraud detection capabilities. The study finds that the combination of autoencoders and SMOTE(Synthetic Minority Oversampling Technique) yields the most effective results, particularly when used with LightGBM classifiers.

Conclusion:

The paper's extensive research which includes L1-regularization and stacked autoencoder layers, successfully achieved the goal of finding optimal solutions for the problem at hand. It also underlines the importance of employing modern techniques like autoencoders and SMOTE to effectively combat fraudulent activities within healthcare insurance claims.

6. Machine Learning Algorithms

Paper 2: Medicare Fraud Detection using Random Forest with Class Imbalanced Big Data

Author: Richard A. Bauder, Taghi M. Khoshgoftaar

Introduction

This paper explores Medicare fraud detection using machine learning by addressing challenges like class imbalance and evaluates the performance of different machine learning models.

Work & Methodology

This paper uses Random Forest classification algorithm to detect Medicare fraud, utilizing Medicare provider claims data from 2012 to 2015 and incorporating fraud labels from the List of Excluded Individuals/Entities (LEIE) database. To address class imbalance, it applies random undersampling, creating training sets with varying fraud-to-non-fraud ratios, ranging from 99.9:0.1 to 50:50. Random Forest models are constructed for each undersampled dataset, and performance is evaluated using 5-fold stratified cross-validation repeated ten times, with the area under the ROC curve (AUC) as the key metric. The study reveals that a 90:10 fraud-to-non-fraud ratio yields the best performance with an AUC of 0.87, outperforming other ratios, including the commonly used 50:50 balanced ratio. This research highlights the significance of addressing class imbalance in Medicare fraud detection models, challenging the notion of balanced class ratios as a universal standard.

Relevance to Project

In the context of Medicare fraud detection, this algorithm implementation holds significant relevance. By utilizing real Medicare claims data linked to fraud labels, it provides a valuable benchmark dataset for the healthcare industry. Moreover, it underscores the necessity of sampling and balancing techniques when dealing with imbalanced Medicare data to enhance the effectiveness of fraud detection systems. The paper's findings also challenge the prevailing belief that balanced class ratios are always optimal, as the study demonstrates that a 90:10 ratio outperforms the traditional 50:50 ratio. This research highlights the importance of algorithmic approaches in improving Medicare fraud detection, ultimately contributing to better healthcare fraud prevention and more efficient use of resources.

Conclusion

The research highlights how machine learning can effectively address Medicare fraud, with a specific focus on managing class imbalance. By employing the Random Forest model, the study recommends using a 90:10 class distribution, which improves fraud detection while retaining non-fraudulent cases.

Paper 17: Medicare Fraud Detection using Machine Learning Methods
Author: Richard A. Bauder, Taghi M. Khoshgoftaar

Introduction

This paper investigates machine learning models to identify Medicare providers engaged in fraud. Using data from the 2015 Medicare PUF dataset and the LEIE database, the study evaluates various machine learning techniques to address this Medicare fraud

Work & Methodology

This study investigates the application of machine learning techniques for detecting fraudulent Medicare providers, utilizing Medicare claims data from 2015 and incorporating fraud labels from the List of Excluded Individuals/Entities (LEIE) database. The research conducts a thorough comparison of various machine learning methods, including supervised techniques like Gradient Boosted Machine (GBM), Random Forest (RF), Deep Neural Network (DNN), and Naive Bayes, as well as unsupervised techniques such as Autoencoder, Mahalanobis distance, k-Nearest Neighbors (kNN), and Local Outlier Factor. To address the class imbalance between the limited fraud cases and the larger set of normal cases, two sampling methods are employed: oversampling of the fraud class and 80-20 undersampling that retains all fraud cases. The evaluation metrics used, including balanced accuracy, F-measure, G-measure, and Matthew's correlation coefficient (MCC), help assess the performance of these methods. The study reveals that supervised methods, particularly DNN, GBM, and RF, outperform unsupervised or hybrid methods, especially when using the 80-20 sampling technique. However, the effectiveness of fraud detection varies across provider types, with more specialized fields being easier to detect fraud in than general ones, such as Family Practice.

Relevance to Project

The relevance of this research to Medicare fraud detection is substantial, as it offers an in-depth analysis of machine learning methods applied to healthcare claims data. The findings can inform and guide projects aimed at healthcare claims fraud detection, including your own. By comparing different sampling techniques and model evaluation metrics, this study provides valuable insights into optimizing the approach for detecting fraudulent Medicare providers. It underscores the importance of tailored machine learning models and sampling strategies to address class imbalance effectively, which can ultimately lead to more accurate and efficient healthcare fraud detection systems.

Conclusion

Data sampling methods significantly affect performance, with the 80-20 method outperforming oversampling.

Paper 18: Identifying Medicare Provider Fraud with Unsupervised Machine Learning
Author: Richard A. Bauder, Raquel C. da Rosa, Taghi M. Khoshgoftaar

Introduction

This paper uses machine learning techniques to address this issue, making use of publicly available data and the LEIE database. Its primary aim is to evaluate the effectiveness of various machine learning methods in addressing Medicare fraud.

Work & Methodology

This research paper investigates the application of unsupervised machine learning methods for Medicare fraud detection, evaluating five different approaches, including Isolation Forest, Local Outlier Factor (LOF), Unsupervised Random Forest (URF), autoencoder (AE), and k-Nearest Neighbors (KNN). The study uses Medicare Part B data from 2012-2015, comprising over 37 million claims, with fraud labels obtained from the List of Excluded Individuals/Entities (LEIE) database. The performance evaluation metrics include AUC, ROC curves, sensitivity, and specificity. The study reveals that LOF with 40 neighbors (LOF40) performs the best overall, achieving an AUC of 0.629, making it a promising method for identifying potential Medicare fraud. The research underscores the potential of unsupervised machine learning for flagging suspicious providers and reducing investigation time and resources in Medicare fraud detection.

Relevance to Project

The relevance of this research to Medicare fraud detection is significant. It offers a comprehensive comparison of various unsupervised learning techniques on a substantial Medicare claims dataset with real fraud labels. The paper's data preprocessing and model evaluation methods can serve as valuable insights for similar projects in the field. Furthermore, the study highlights LOF as a top-performing algorithm, which can be instrumental in enhancing the accuracy and efficiency of Medicare fraud detection models. In summary, this research provides a valuable framework for implementing and assessing unsupervised learning algorithms in the context of Medicare fraud detection, facilitating improved fraud prevention and resource allocation.

Conclusion

This study demonstrates the potential of unsupervised machine learning models in detecting Medicare provider fraud. While models like LOF and URF show promise, they can be further improved, especially in terms of specificity and overall AUC scores.

7. References

[1] Zhang, H., & Wang, L. (2018). Prescription fraud detection through statistic modeling. *Proceedings of 2018 International Conference on Mathematics and Artificial Intelligence*.

[2] Bauder, R.A., & Khoshgoftaar, T.M. (2018). Medicare Fraud Detection Using Random Forest with Class Imbalanced Big Data. *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 80-87.

[3] Haddad Soleymani M, Yaseri M, Farzadfar F, Mohammadpour A, Sharifi F, Kabir MJ. Detecting medical prescriptions suspected of fraud using an unsupervised data mining algorithm. Daru. 2018 Dec;26(2):209-214. doi: 10.1007/s40199-018-0227-z. Epub 2018 Nov 20. PMID: 30460618; PMCID: PMC6279664.

[4] Bauder, R.A., Khoshgoftaar, T.M. Multivariate outlier detection in medicare claims payments applying probabilistic programming methods. *Health Serv Outcomes Res Method* **17**, 256–289 (2017). https://doi.org/10.1007/s10742-017-0172-1

[5] Ortega, Pedro & Figueroa, Cristián & Ruz, Gonzalo. (2006). A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile. DMIN. 6. 224-231.

[6] Karca Duru Aral, Halil Altay Güvenir, İhsan Sabuncuoğlu, Ahmet Ruchan Akar,
A prescription fraud detection model, Computer Methods and Programs in Biomedicine,
Volume 106, Issue 1, 2012, Pages 37-46, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2011.09.003.

[7] Kumaraswamy N, Markey MK, Ekin T, Barner JC, Rascati K. Healthcare Fraud Data Mining Methods: A Look Back and Look Ahead. Perspect Health Inf Manag. 2022 Jan 1;19(1):1i. PMID: 35440932; PMCID: PMC9013219.

[8] Herland, M., Khoshgoftaar, T.M. & Bauder, R.A. Big Data fraud detection using multiple medicare data sources. *J Big Data* **5**, 29 (2018). https://doi.org/10.1186/s40537-018-0138-3

[9] Li J, Huang KY, Jin J, Shi J. A survey on statistical methods for health care fraud detection. Health Care Manag Sci. 2008 Sep;11(3):275-87. doi: 10.1007/s10729-007-9045-4. PMID: 18826005.

[10] Herland, M., Bauder, R. A., & Khoshgoftaar, T. M. (2020). Approaches for identifying U.S. medicare fraud in provider claims data. *Health care management science*, *23*(1), 2–19. https://doi.org/10.1007/s10729-018-9460-8

[11] Herland, M., Bauder, R.A. & Khoshgoftaar, T.M. The effects of class rarity on the evaluation of supervised healthcare fraud detection models. *J Big Data* **6**, 21 (2019). https://doi.org/10.1186/s40537-019-0181-8

[12] Batko, K., Ślęzak, A. The use of Big Data Analytics in healthcare. *J Big Data* **9**, 3 (2022). https://doi.org/10.1186/s40537-021-00553-4

[13] Karthika, I., and K. P. Porkodi. "Fraud Claim Detection Using Spark." *International Journal of Innovations in Engineering Research and Technology*, vol. 4, no. 2, 2017, pp. 1-4

[14] M. Dash, H. Liu, Feature selection for classification, Intelligent Data Analysis, Volume 1, Issues 1–4, 1997, Pages 131-156, ISSN 1088-467X, https://doi.org/10.1016/S1088-467X(97)00008-5

[15] Ortega, Pedro & Figueroa, Cristián & Ruz, Gonzalo. (2006). A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile. DMIN. 6. 224-231.

[16] Kumari, A., Punn, N.S., Sonbhadra, S.K., Agarwal, S. (2023). Impact of the Composition of Feature Extraction and Class Sampling in Medicare Fraud Detection. In: Tanveer, M., Agarwal, S., Ozawa, S., Ekbal, A., Jatowt, A. (eds) Neural Information Processing. ICONIP 2022. Lecture Notes in Computer Science, vol 13625. Springer, Cham. https://doi.org/10.1007/978-3-031-30111-7_54

[17] R. Bauder and T. Khoshgoftaar, "Medicare Fraud Detection Using Random Forest with Class Imbalanced Big Data," 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, USA, 2018, pp. 80-87, doi: 10.1109/IRI.2018.00019.

[18] R. A. Bauder and T. M. Khoshgoftaar, "Medicare Fraud Detection Using Machine Learning Methods," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 2017, pp. 858-865, doi: 10.1109/ICMLA.2017.00-48.

[19] R. Bauder, R. da Rosa and T. Khoshgoftaar, "Identifying Medicare Provider Fraud with Unsupervised Machine Learning," 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, USA, 2018, pp. 285-292, doi: 10.1109/IRI.2018.00051.