A Hybrid Technique for Health Insurance Fraud Detection on Highly Imbalanced Dataset

Shamitha S K , V Ilango

Abstract: ***Health Insurance industry is producing a massive amount of heterogeneous data. Detecting fraud from these data is a challenging task. Highly imbalanced data causes huge challenge to the Insurance Data Analysis. Classification of imbalanced data is a critical issue faced by the fraud detection methodologies. Fraud only covers less than 10% of the whole data. In this study, we use highly imbalanced data and propose a hybrid method for fixing class imbalance problem by using a combination of SMOTE, Cross Validation, and Random Forest. We used Medicare data, which will be applied to various sampling techniques, and further a classification model was built. We observed that SMOTE with Random forest with cross validation produced excellent results. Our model should be capable of identifying all the relevant(fraud) instances, i.e., the model should have a high recall value. SMOTE with Random forest had average recall of 86% and an overall accuracy of 90%, which could be considered as good among the existing models.***

Keywords : ***Health Insurance Fraud, SMOTE, Cross Validation, Random Forest***

I. INTRODUCTION

Health Insurance systems have become one of the major concerns of modern life. These systems financially help people to pay high costs healthcare expenses. Healthcare services had been utilizing the conservationist strategy for determination and treatment, where most specialists relied upon their individual experience and aptitudes in diagnosing maladies in patients, bringing about a less exact and patient-driven. Digitization, progression in innovation, the requirement for evidence-based medication, powerlessness to outgrowth and get knowledge from consistently growing different medicinal information are a portion of the drivers for embracing health care analytics [1]. Insurance plays a central role in the healthcare field. More than 80 percent of healthcare expenditures are funded by insurance companies, either public or private. Insurance thus offers the money that motivates and cares for the health care system [2].

The health insurance industry generates a massive amount of data. These data involve details such as patient records, providers data, treatment information, drug details, etc. Detecting fraud from this enormous information is a challenging task. Insurance fraudulent claims setback a severe harm over billions for many insurance organizations. Fraudulent claims create a negative impact on the insurance firms; it also terribly hurt socio-economic structures. The effect of loss to an organization or an economy is very high. While implementing machine learning algorithms to these kinds of problem, working on unbalanced data has unavoidably become a major challenge to data analysts [3].

There are two main problem imposed by fraud detection data due to class imbalance. First one is a higher probability of instances belonging to non-fraudulent classes. This imbalance may lead the classifiers to classify new observations to the non-fraudulent class. If 99% of the cases in a dataset belongs to the same class, when a classification model is built from the above-said data the classifier will label the test cases with these majority of non-fraudulent classes, it may lead to an accuracy of 99%. This can be merely considered as “accuracy paradox,” i.e., a high level of accuracy is not considered as an indicator of higher classification performance [4]. Fraud detection is one of the examples for such models where the data is highly imbalanced, and accuracy cannot be taken in to account. The second problem in fraud detection the cost of False Negatives, which should be considered much seriously than False Positives. Classifiers penalize both at a similar weight. Cost of missing even a single case of fraud (say, False negatives) is high.

An appropriate model is necessary that classifies unbalanced nature of data in fraud detection problem. There are some methods introduced to address the issue of data imbalance; these methods can reduce the impact of skewed data on classification performance. These methods are classified into two levels, data level and algorithm level [5]. In the algorithmic approach, the algorithm itself will be modified to increase the predictive performance of the minority class. Data level approach consists of oversampling and under sampling the minority or majority classes in order to lessen the effect carried by class imbalance. In this paper, we have used data level approaches for balancing the data. Several sampling approaches are used here such as SMOTE, ADASYN, Random Over Sampling. In order to evaluate performance metrics on fraud detection four learning algorithms (Random Forest, XGBoost, Light GBM, (GradientBoostingClassifier) using Python framework. All the above said unbalanced data sampling techniques are applied to all four learning algorithms. Random Forest Classifier with SMOTE produced good overall results.

The work is organized as follows. Section 2 discusses

Revised Manuscript Received on July 22, 2019.

Shamitha S K, Research Scholar, Department of Master of Computer Applications**,** CMR Institute of Technology**,** AECS, Layout, ITPL Main Road, Bangalore-560037, Email: shamithashibu@gmail.com

V Ilango, Head - Centre of Excellence for Intelligent Human-Computer Interaction (c-IHCI)**,** CMR Institute of Technology**,** AECS, Layout, ITPL Main Road**,** Bangalore-560037 **,** Email: thirukkural69@gmail.com

Retrieval Number: F12100886S19/2019©BEIESP
DOI:10.35940/ijrte.F1210.0886S19

previous works related to class imbalance and fraud detection in health insurance.

Section 3 gives a brief introduction to the algorithms used throughout the study, which includes classification and data sampling algorithm. Section 4 provides an idea of the data used in the study. Section 5 discusses the results obtained from the study. Section 6 concludes the work.

II. Related Works

An excellent survey on fraud detection was conducted by Aisha et al. [6], which explains Issues and challenges associated with health insurance domain. Tahir Ekin et al. [7] used a Bayesian approach in healthcare fraud detection to identify providers or beneficiaries who exhibits fraudulent behavior. A novel hybrid approach was built in the paper [3] by Youjun Zhang et al., which was handling data imbalance problem using a combination of both K Reverse Nearest Neighborhood and One Class support vector machine (OCSVM). Several other learning algorithms were also employed in this dataset. The model was producing an overall accuracy of 91.89% on Insurance dataset, but could not see any balancing techniques applied in the model. Richard Bauder et al., in their paper [8] made an effort to asses' providers fraudulent activities by examining the payment done to them for the services they have rendered. The author proposed a novel method for detecting fraud, which deals with detecting outliers in payment data using multiple predictors as model input. Richard Bauder et al. had further extended their work combining Medicare databases and building a model with a Random Forest model using five cross-validations which yielded an accuracy of 87.3% [9].
The effects of class imbalance problem have been reported as a significant hindrance on machine learning performance [9], [10]. Chawla [11] presented a detailed review of the issues and challenges related to data unbalancing. The studies state that the problem of unbalanced data classification persists in the real-world scenario. Similar studies were carried out by Foster [12]. In his paper, he discussed sampling strategies. The author also believes that a proper understanding of the unbalanced data problem will create broader implications for ML and AI areas. Paper [5] states that four factors affect the performance of the classifier on unbalanced data sets. They are Imbalanced class distribution, data sample size, class separability and within-class concept. Kubat and Matwin [13] used data sampling methods in their paper. They applied one-sided selection techniques by keeping the minority class as fixed and under sampled the majority class. Then they categorized the minority classes into noise overlapping. There are many other studies related to a class imbalance in credit card fraud detection[14], [15]. There are only a few studies into class unbalancing in health insurance fraud detection. To name a few of them are [16], [17]. In this section, we describe related work from the perspective of data sampling techniques available to handle health insurance data and fraud detection in the Medicare database..

III. Experiment

3.1 Data Sampling Techniques Used

3.1.1 SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE is a systematic algorithm used to create synthetic samples. It is even called Synthetic Minority Over-Sampling as it creates synthetic samples from the minor classes rather than multiplying the copies. SMOTE Algorithm selects two or more similar instances and modifying the individual attributes in an instance with a random value which is in between the boundary of the neighboring instances attribute. In this study, we use SMOTE to mitigate the problem arising due to unbalanced classes. We applied various sampling techniques on the data. Among them, SMOTE was providing better results. SMOTE synthesizes new minority samples in between the existing classes based on its nature [18]. Figure 1 and 2 shows the data distribution before and after the sampling. Considering the quality of the data, the model tends to focus on oversampling minority classes (Fraud classes). For each fraud samples, SMOTE calculates the k nearest neighbors. By oversampling the minority classes, we are trying to create a balance between majority and minority class here, this helped us improving the model performance.

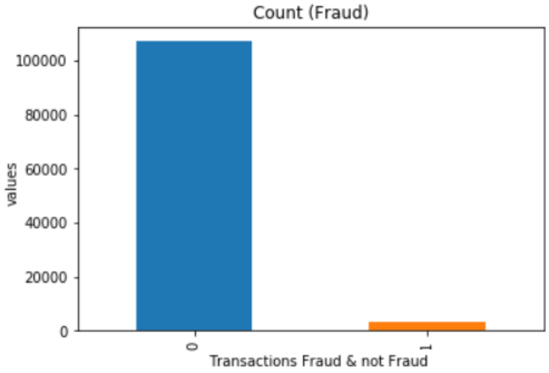

Figure 1 Class distribution before sampling

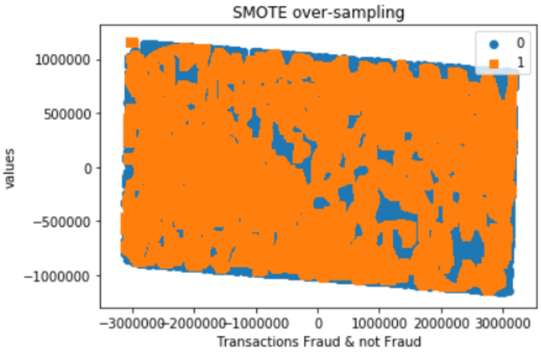

Figure 2 Class distribution after SMOTE

3.1.2 ADASYN

ADASYN is much similar to SMOTE but could be attributed as better off considering its accuracy. ADASYN is more fine-tuned and improved. Once you create the samples, the algorithm tries adding small random values to the points. This will make ADASYN more approximate towards reality. Unlike SMOTE where the samples are more related to its neighboring sample or the parent, ADASYN adds more variance and thus are more scattered across.

3.1.3 Random Oversampling (ROS)

Random Oversampling or ROS algorithm tries to add on to the samples with multiple copies of selected minority classes. You can repeat the algorithm to any number of times you prefer to run.

Also, here instead of blindly cloning every sample, some might be selected with a complete replacement.

3.1.4 Random Under-Sampling (RUS)

RUS is an algorithm where it tries to remove samples from the majority class, and this removal could be with or without a replacement. This is considered to be one of the most uncomplicated technique. It helps in to narrow down the high imbalance in the samples. Said this, it could also result in high variance amongst the records.

3.2 Classification techniques Used

This section provides a brief explanation of the classifiers used throughout the study.

3.2.1 Random Forest

Random Forest classifier is an ensemble method which is highly flexible and can be used in both classification and regression tasks. We have used Random forest classifier in our study to assess the performance of the dataset. Several classifiers were tested against the synthetic oversampled data; among them, Random Forest classifier yielded better results [19]. Random Forest can also be used in feature selection. It uses the Gini index, which assigns score and rank features based on the score. The most important feature will be assigned the highest score. Random Forest is also used in the paper for feature selection; out of 27 features, eight were selected using the classifier. Random forest classifier was used to find the variable importance. Figure 3 shows the results of the scoring and ranking of the top 8 variables. Random Forest classifier is an ensemble method in which decision trees will be generated based on the samples of data. At each node of the tree, the ultimate goal of RF classifier is to reduce the entropy and increase the information gain [20].

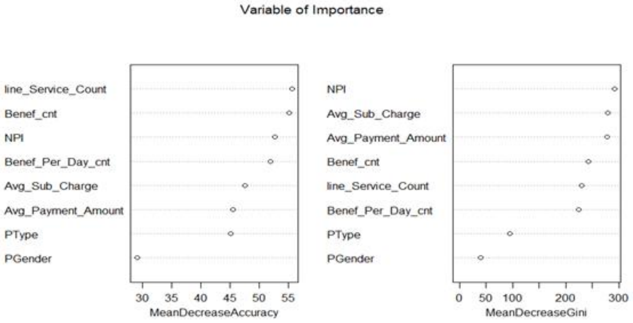

Figure 3 Variables of importance using Random Forest Algorithm

3.2.2 Light GBM

Light GBM(LGBM) is a tree-based learning algorithm. Tree-Based algorithm grows horizontally(level-wise), but LGBM grows vertically, ie., leaf-wise. Leaf wise algorithm is capable of reducing more loss, ie., it will lower the error faster than level wise. LGBM uses gradient boosting framework [21].

3.2.3 XGBoost (eXtreme Gradient Boosting)

XGBoost is an implementation of gradient boosted decision trees. It is an ensemble algorithm because it offers solutions by combining multiple classification algorithms. It is prevalent because of its performance and time optimization.

3.2.4 Gradient Boosting

Gradient boosting is a boosting the algorithm, which converts weak learners into strong learners. Gradient boosting algorithm applies three elements to train the models they are additive, gradual and sequential. Gradient boosting contains a loss function; It uses gradients in the loss function to optimize the loss. This algorithm allows the user to optimize a specified cost function, which makes it more ideal for real-world applications.

3.3 Cross-Validation

We have used K fold cross-validation in our work; It helps in further resampling the data. The value of K chosen was 5. Cross-validation involves dividing a group of data samples into k folds of similar size. In that 1st group or fold is considered as a training data set and this set will be further fitted in to the remaining folds

IV. Data

In this section, we describe the dataset used in the study. Two CMS Medicare datasets are used for the study, Part B and LEIE [22][23]. The initial dataset provides details of treatment or procedure a physician performs for a year, basically the claims information. The above-said database is available at the CMS website for the year
2012-2016 calendar years. Part B data set contains variables such as NPI (Provider Identifier Number) of a physician, which is a unique number. Healthcare Common Procedure Coding System (HCPCS) code, Claims Information including average payment and charges, number of procedures, number of beneficiaries, facility or non-facility and more.

LEIE (List of Excluded Individuals and Entities) is a list of data which contains the details of providers excluded from Medicare facility. This database includes information such as, the reason for exclusion, date of exclusion and waiver date of physicians who are found unsuited to practice. This dataset is maintained by the Office of Inspector General (OIG). OIG has the authority to waive off providers from these kinds of funded programs if they are found guilty.

As the last stage, a new data set was formed, which is a labelled dataset. LEIE database which contains the excluded providers are matched with their NPIs and matching NPI is considered as fraud else non fraud. There was 18 provider type in total; only eight providers are used in the study.

V. Results and discussions

The study proposes a hybrid method for fraud detection, which effectively deals with the highly imbalanced dataset. Figure 4 describes the model. The model uses a combination of SMOTE, Random Forest with cross-validation in the study. The classes were highly unbalanced with the distribution of 99:1(majority: minority). 99% of the data set covers legitimate transactions, and 1% includes illegitimate transactions. To start of Random forest classifier was applied to the data without any data sampling techniques, the model produced an accuracy of 99%. Though the accuracy and precision were high, the recall was very low. Where precision(P) is calculated by P=FP/(FP+TN) and Recall(R) is calculated by R=TP/(TP+FN). AUC curve is considered as a right performance metrics while dealing with fraud detection, from the above results average AUC was less.

As a next step four sampling techniques were applied to random forest algorithm. SMOTE was performing well compared to other data sampling techniques. The combination of the above-said methods produced 90% AUC and 86% average precision-recall values. Figure 5 shows the AUC and precision-recall curve. The model outperforms the existing models in terms of AUC for 99:1 class distribution.

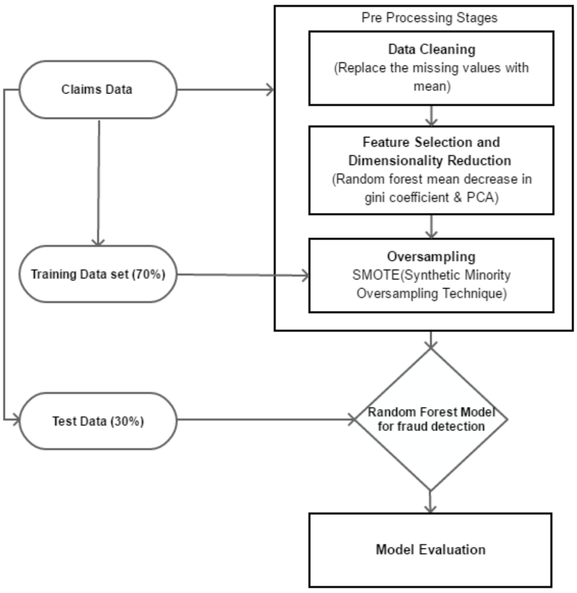
Pre Processing Stages
Data Cleaning
(Replace the missing values with mean)
Feature Selection and Dimensionality Reduction
(Random forest mean decrease in gini coefficient & PCA)
Oversampling
SMOTE(Synthetic Minority Oversampling Technique)
Claims Data
Training Data set (70%)
Test Data (30%)
Random Forest Model for fraud detection
Model Evaluation

Figure 4. Hybrid Model for fraud detection on imbalanced data

Algorithms	**ROC_AUC score**	**Precision**		**Recall**	
		Class 0	**Class 1**	**Class 0**	**Class 1**
Random Forest	0.9	0.83	0.83	0.83	0.83
XGB	0.5	1	0	1	0
LGBM	0.52	1	0.08	1	0.06
Gradient Boosting	0.82	0.84	0.8	0.8	0.82

Table 1. Comparison Performance of different classifiers on SMOTE

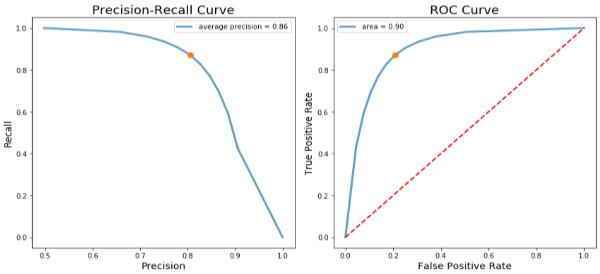

Figure 5. AUC and Average Precision Recall Graph

VI. Conclusion

Health Insurance is a significant contribution to the overall expenditure of an individual. This makes fraud detection in health insurance a necessity to smooth processing of funds for the health care expenditure. Data imbalance is found as a major problem while processing health insurance data. In our study, we presented an efficient method for fraud detection, which also deals with class imbalance. We have used a hybrid combination of SMOTE, Random Forest with Cross-Validation. This hybrid combination is exhibiting good data balancing capabilities. We obtained AUC 0.90 for a 99:1 class distribution, which is considered as a highly imbalanced dataset. Future work will be focused on extending the database with more related data from the Medicare database and additionally performing Real-time fraud detection which effectively handles class imbalance problem promptly.

REFERENCES

1. W. W. Malek, K. Mayes, and K. Markantonakis, “Fraud Detection and Prevention in Smart card Based Environments Using Artificial Intelligence,”, vol. 5186, pp. 118-132, 2008.
2. M. Kirlidog and C. Asuk, “A Fraud Detection Approach with Data Mining in Health Insurance,” vol. 62, pp. 989–994, 2012.
3. G. G. Sundarkumar and V. Ravi, “A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance,” Eng. Appl. Artif. Intell., vol. 37, pp. 368–377, 2015.
4. F. J. Valverde-Albacete, J. Carrillo-de-Albornoz, and C. Peláez-Moreno, “A proposal for new evaluation metrics and result visualization technique for sentiment analysis tasks,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8138 LNCS, pp. 41–52, 2013.
5. A. Wong and M. S. Kamel, “Classification of imbalanced data : a review CLASSIFICATION OF IMBALANCED DATA : A REVIEW,” no. November, 2011.
6. A. Abdallah, M. A. Maarof, and A. Zainal, “Journal of Network and Computer Applications Fraud detection system : A survey,” vol. 68, pp. 90–113, 2016.
7. T. Ekin, F. Leva, F. Ruggeri, and R. Soyer, “Application of Bayesian Methods in Detection of Healthcare Fraud Application of Bayesian Methods in Detection of Healthcare Fraud,” no. February, vol 33, pp. 151-156, 2013.
8. R. A. Bauder and T. M. Khoshgoftaar, Multivariate outlier detection in medicare claims payments applying probabilistic programming methods. Vol. 17, pp. 256-289, Springer US, 2017.
9. R. A. Bauder and T. M. Khoshgoftaar, “Medicare fraud detection using random forest with class imbalanced big data,” Proc. - 2018 IEEE 19th Int. Conf. Inf. Reuse Integr. Data Sci. IRI 2018, pp. 80–87, 2018.
10. C. Phua, D. Alahakoon, and V. Lee, “Minority report in fraud detection,” ACM SIGKDD Explor. Newsl., vol. 6, pp. 50, 2007.
11. N. V Chawla, N. Japkowicz, and P. Drive, “Editorial : Special Issue on Learning from Imbalanced Data Sets Aleksander Ko l cz,” vol. 6, no. 1, pp. 2000–2004, 2004.
12. Foster, P. Machine learning from imbalanced data sets 101. In Proceedings of the AAAI’2000 Workshop on Imbalanced Data

Sets; New York University: New York, NY, USA, 2000.

13. Kubat M and Matwin S. Addressing the curse of imbalanced training sets: one-sided selection. In: Proceedings of the 14th inter-national conference on machine learning, Nashville, TN, 8–12 July 1997, pp. 179–186. San Francisco, CA: Morgan Kaufmann
14. Chen R.C., Chen, T.S., & Lin C.C. " Detecting Credit Card Fraud by using Questionnaire – Responded Transaction Model based on support vector machines",Springer-Verlag Berlin Heidelberg. pp. 800-806.," no. February, 2006.
15. T. M. Padmaja, N. Dhulipalla, P. R. Krishna, R. S. Bapi, and A. Laha, "An Unbalanced Data Classification Model Using Hybrid Sampling Technique for Fraud Detection," Pattern Recognit. Mach. Intell., pp. 341–348, 2007.
16. R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin, "Data sampling approaches with severely imbalanced big data for medicare fraud detection," Proc. - Int. Conf. Tools with Artif. Intell. ICTAI, vol. 2018-Novem, pp. 137–142, 2018.
17. M. Herland, T. M. Khoshgoftaar, and R. A. Bauder, "Big Data fraud detection using multiple medicare data sources," J. Big Data, pp. 1–21, 2018.
18. N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Over-sampling Technique," vol. 16, pp. 321–357, 2002.
19. J. M. Pérez, J. Muguerza, O. Arbelaitz, I. Gurrutxaga, and J. I. Martín, "Consolidated Tree Classifier Learning in a Car Insurance Fraud Detection Domain with Class Imbalance," pp. 381–389, 2005.
20. I. C. Society, L. I. Kuncheva, and C. J. Alonso, "Rotation Forest : A New Classifier Ensemble Method," vol. 28, no. 10, pp. 1619–1630, 2006.
21. G. Ke et al., "LightGBM : A Highly Efficient Gradient Boosting Decision Tree," no. Nips, pp. 1–9, 2017.
22. "Part B National Summary Data File (Previously known as BESS)," Data base _ Medicare CMS, 2018. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Part-B-National-Summary-Data-File.
23. Office of Inspector General, "Exclusions - Office of Inspector General," Exclusions database U.S. Department of Health and Human Services, 2018. [Online]. Available: https://oig.hhs.gov/exclusions/background.asp.
24. Vashistha R, Kumar P, Dangi AK, Sharma N, Chhabra D, Shukla P. Quest for cardiovascular interventions: precise modeling and 3D printing of heart valves. Journal of biological engineering. 2019 Dec;13(1):12.