Using a new data mining method for automobile insurance fraud detection: A case study by a real data from an Iranian insurance company

Maryam Esna-Ashari

Insurance Research Center, Tehran, Iran

Abstract

The car insurance fraud is one of the most important issues for insurance companies because it can result in a huge financial loss in an insurance company. Therefore, timely and early detection of a suspected case can greatly prevent this loss. Within the past decade, several studies have been conducted using data mining techniques for fraud detection. In this article, we first investigate the challenge of imbalanced data, and after resolving it, we apply a new algorithm proposed for fraud discovery called XGBoost, for a real data set. Finally, we compare this algorithm with the older one, Random Forest algorithm, and show the proposed algorithm functionality.

Keywords: Fraud detection, Imbalanced data, XGBoost algorithm, Random Forest algorithm.

AMS Subject Classification: 68T09

1. Introduction

The major issue faced by insurance companies is a fraud that causes immense loss to insurance companies that may not be recovered. The main concern is to avoid fraudulent activities at all costs since investigating fraud cases in insurance companies is very challenging. It's been reported that from 21% to 36% of cases of auto insurance claims are suspected to be fraudulent but only 3% of cases are prosecuted [1]. The first step to avoid fraud is to detect them which is quite difficult and cost-ineffective", and because of lengthy and cumbersome investigations, it may infuriate authentic customers [2]. High investigation cost is another barrier in detecting fraud cases. Therefore, companies may not be able to conduct an appropriate investigation leading to several potential pitfalls. Manual fraud detection is no longer employed due to high cost and low efficiency. Furthermore, the investigation needs to be initiated before finalizing payment for a claim.

To address these challenges, the application of data mining techniques has emerged as a pivotal strategy against automobile insurance fraud. Data mining, a subset of artificial intelligence (AI) and machine learning (ML), enables insurers to harness the power of vast datasets to uncover hidden patterns, anomalies, and correlations that indicate fraudulent activities. By analyzing historical claims data, customer behaviors, and external factors, data mining algorithms can identify suspicious claims, flag potential fraud cases, and significantly enhance the efficiency and accuracy of fraud detection processes. The evolution of data mining in insurance fraud detection represents a paradigm shift from traditional methods reliant on manual review and statistical analysis to proactive, data-driven approaches capable of detecting fraud in real-time. This transformation is driven by the exponential growth in digital data, advancements in computing power, and the development of sophisticated algorithms that can process and interpret complex datasets with speed and precision [3].

The two basic learning techniques are supervised and unsupervised. In supervised learning, we are provided with fully labeled data that means in the training data against each input we have the desired result as well. It is highly useful for solving problems of classification and regression. In classification, the aim is to predict a discrete value whereas regression deals with continuous data. On contrary, in an unsupervised learning paradigm, we are provided with unlabeled data where results are not known [4]. In a fraud detection scenario in a supervised learning method, we can find out fraud and legal cases from training data but in unsupervised learning, we cannot infer which one is a fraud case and which one is legal. However, our data set has the label and thus our method is supervised.

In Iranian studies, there is one research by unsupervised method in [5] in which they used isolation forest algorithm. For supervised method, [6,7,8] used traditional methods Naive Bayes, decision tree and logistic regression. Since data labeled fraud is really limited, [9] recently used another approach called Target replacement. In foreign studies, there are many references that used traditional methods mentioned above (see e.g., [10,11,12,13,14]). In recent years, a new method called XGBoost (eXtreme Gradient Boosting) has being used in foreign studies (see e.g., [15,16,17]) while, as the best of our knowledge, there is not any finding in Iranian research.

In this article, we first argue about the imbalanced data in Section 2. In Section 3, we then apply the XGBoost method in our real data. Finally, we compare it with another recently developed method, random forest method.

2. Imbalanced data

Our data set has 13200 records for the years between 2017 and 2023. Also, it includes one response random called fraud and 14 final features as follows (after data preprocessing):

1. The difference between the date of the accident and the declaration of damage;

2. Location of the accident (north, west, south, east and center of Iran);

3. The cause of the accident (not paying attention to the front, inability to control the vehicle, sudden change of direction, etc.);

4. Type of accident (accident, on-the-spot theft, broken glass, etc.);

5. Type of report (none, report of police authorities, non-compromising croqui, conciliatory croqui and others);

6. Who is responsible for the accident (whether or not the insured person was in fault);

7. Percentage of fault;

8. Used at the time of the accident (personal, administrative, cargo, administrative affairs, etc.);

9. The value of the vehicle;

10. Insurance policy premium;

11. Covering requested parts (does not have, has);

12. Vehicle group (rider, truck, car, motorcycle and spare parts only);

13. Change of owner (no, yes);

14. Initial damage assessment.

However, it is found that the number of malicious claims (11 records) is much less than the total claims submitted. This uneven distribution (data imbalance) leads to more burdensome fraud detection. Furthermore, most of the supervised classifiers generate inefficient classification models with unbalanced data [18], since they prefer to categorize all the data points as genuine class (major class samples) and ignore the fraudulent points (minority class samples).

[19] proposed Random Over-Sampling Examples (ROSE), which generates new minority samples based on the kernel density estimate around real minority cases (see Figure 1).

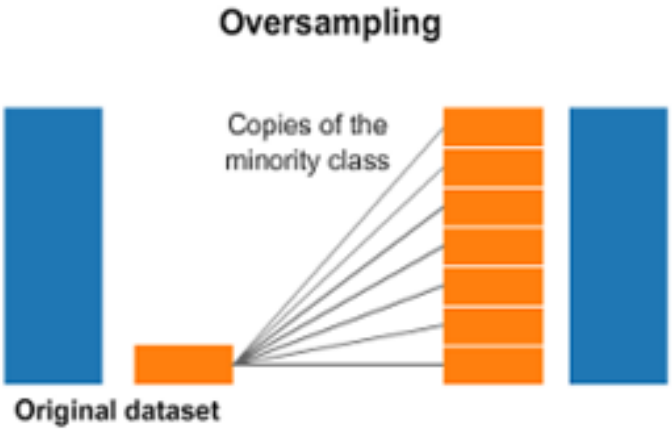

Figure 1. Random Over-Sampling Examples

So, we first apply this method to our data set to balance the data as the following:

```
    0     1
13189 13211
```

where 1 is for fraud and 0 is for non-fraud.

3. Main results

XGBoost is a popular model that optimizes gradient tree boosting and learns from tabular data. High scalability makes XGBoost run ten times faster than other conventional models and robust to a high dimensional dataset. This high scalability is empowered by implementing a tree-learning algorithm optimized for sparse data, a weighted quantile algorithm for more efficient computation, and a cache-aware block structure for parallelizing the tree-learning process using all processor cores [20].

Now, we apply this method to our data set explained in the previous section. The XGBoost algorithm has some user-friendly parameters in which 3 of most important ones are:

• max.depth: The maximum depth of a tree. Increasing this value makes the model more complex and increases the probability of overfitting. Zero indicates no limit in depth. Caution should be exercised here because XGBoost is memory intensive when training a deep tree. The exact tree method requires a non-zero value, and the default value of the algorithm is 2. Its range is between zero and infinity;

• eta: Shrinking step size is used in the update to avoid overfitting and ranges between zero and one;

• objective: here the work is binary:logistic.

These parameters can be set to different numbers and the results can be compared with each other. It should be noted that, the smaller the logloss values, the better performance. Finally, by using the XGBoost algorithm for different parameter values, the outputs were cross-compared, and the optimal relative mode was selected for the parameters. Some of the values of the model parameters along with the logloss results are given in the table below. The first line shows the results for the default values of the algorithm.

eta	**max.dept**	**train-logloss**
0.3	2	0.314227
0.1	2	0.527900
0.6	2	0.157292
1	2	0.080683
1	1	0.124420
1	3	0.070512
1	9	0.047558
1	10	0.047747
1	11	0.047747

Table 1: The logloss for different values of the model parameters

According to the Table 1, the best values obtained are as follows: max.depth=10 and eta=1.

In this step, we compare this model with the Random Forest model (that has not been used in Iranian studies, to the best of our knowledge). It has some user-friendly parameters in which 2 of most important ones are:

• ntree: number of trees. We want enough trees to stabilize the error but using too many trees is unnecessarily inefficient, especially when using large data sets.

• mtry: the number of variables to randomly sample as candidates at each split.

After implementing the Random Forest algorithm several times with different parameter values, finally these two parameters were selected with 500 and 2, respectively, and increasing these parameters does not change the accuracy of the work and only prolongs the execution time of the program.

For comparing the models, we first arranged the data in two categories: training set (about 70% of the data) and test set (about 30% of the data) and then using the same criteria obtained from the Confusion Matrix.

		Actually	
		Positive	**Negative**
Predicted	**Positive**	True Positives (TPs)	False Positives (FPs)
	Negative	False Negatives (FNs)	True Negatives (TNs)

Figure 2: The Confusion Matrix

The performance of the algorithm is computed by a confusion matrix shown in Figure 2. The positive group indicates the fraud case, and the negative group represents the no-fraud case. True positives (TP) indicate the cases in which we predict fraud, and it actually has fraud. Likewise, true negatives (TN) are the cases

in which we predict no fraud, and it has no fraud. False positives (FP) specify the cases in which we predict fraud, but actually has no fraud. False negatives (FN) are the cases in which we predict no fraud, but it actually has fraud.

The performance of the algorithm is measured using accuracy, sensitivity (also known as recall), specificity, precision, and the F-score, which is the harmonic mean of precision and sensitivity. The associated formulas are listed below. The greater value, the greater performance:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$sensitivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{FP + TN}$$

$$precision = \frac{TP}{TP + FP}$$

$$F - score = \frac{2 \cdot precision \cdot sensitivity}{precision + sensitivity}$$

The desired results are compiled in Table 2, and the two algorithms are compared with one another.

criteria	XGBoost	Random Forest
accuracy	0.9997	0.9998
sensitivity	0.9995	09997
specificity	0.9999	1
precision	0.9999	1
F-score	0.9996	0.9998

Table 2: Comparison of XGBoost and Random Forest algorithms

As per values in Table 2, the Random Forest algorithm demonstrate slightly better performance, however, the differences with the XGBoost algorithm are minimal.

4. Conclusion

Insurance fraud prediction is a key step for protecting against fraud-related losses in an insurance company. Since claims proposed to the insurance company usually consist of many non-fraud/genuine cases and only a small percentage of fraud cases, imbalance class problems arise during fitting of machine learning models. Predictions may be biased toward majority of classes (?) or non-fraud cases in this research. Hence, Random Over-Sampling Examples (ROSE) is proposed as a solution to the imbalance dataset. Then, the results showed that the XGBoost algorithm, as a novel method, has a very good, comparative performance. It is worth mentioning that although the random forest algorithm has offered a relatively better performance, the time of implementation/run of the program is a critical and determining factor. With the size of the samples of this study, the implementation/run of the random forest algorithm takes several hours with a powerful CPU, while the XGBoost algorithm only takes a few seconds. This is crucial to the industry and in practice because the insurance company employee must quickly determine whether a case is suspected of fraud to submit the case to the relevant section if it is approved and to make sure that it won't take long at this stage. Therefore, according to this study, the XGBoost method is very useful for discovering the vehicle insurance fraud.

References

1. Nian, Ke, Haofan Zhang, Aditya Tayal, Thomas Coleman, and Yuying Li. "Auto insurance fraud detection using unsupervised spectral ranking for anomaly." The Journal of Finance and Data Science 2, no. 1 (2016): 58-75.
2. Kirlidog, Melih, and Cuneyt Asuk. "A fraud detection approach with data mining in health insurance." Procedia-Social and Behavioral Sciences 62 (2012): 989-994.
3. Bhowmik, Rekha. "Detecting auto insurance fraud by data mining techniques." Journal of Emerging Trends in Computing and Information Sciences 2, no. 4 (2011): 156-162.
4. Hastie, Trevor, Robert Tibshirani, Jerome H. Friedman, and Jerome H. Friedman. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: springer, 2009.

5. Khanizadeh, Farbod, Farzan Khamesian, and Maryam Esna-Ashari. "Employing unsupervised learning to detect fraudulent claims in auto insurance (isolation forest)." *Journal of Management Accounting* 15, no. 53, (2022): 141-153. (In Persian)
6. Firoozi, Mahdi, Shakoori, Morteza, Kazemi, Leila and Zahedi, Sahar. "Detecting fraud in car insurance using data mining methods." *Iranian Journal of Insurance Research* no. 3, (2011): 103-128. (In Persian)
7. Goodarzi, Atoosa and Jannatbabaei, Sajad. "Evaluation of decision tree, Naive Bayes and logistic regression algorithms in detecting car insurance frauds." *Insurance Research* no. 2, (2017): 61-80. (In Persian)
8. Goleiji, Leila, and M. Tarokh. "Identification of influential features and fraud detection in the Insurance Industry using the data mining techniques (Case study: automobile's body insurance)." *Majlesi J Multimed Process* 4 (2015): 1-5.
9. Khanizadeh, Farbod, Maryam Esna-Ashari, Farzan Khamesian, and Azadeh Bahador. "Target replacement, a new approach to increase the performance of fraud detection system in auto insurance utilizing supervising learning." *Journal of Quality Engineering and Management* 11, no. 4 (2022): 413-428. (In Persian)
10. Gepp, Adrian, J. Holton Wilson, Kuldeep Kumar, and Sukanto Bhattacharya. "A comparative analysis of decision trees vis-a-vis other computational data mining techniques in automotive insurance fraud detection." *Journal of data science* 10, no. 3 (2012): 537-561.
11. Prasasti, Iffa Maula Nur, Arian Dhini, and Enrico Laoh. "Automobile insurance fraud detection using supervised classifiers." In *2020 International Workshop on Big Data and Information Security (IWBIS)*, pp. 47-52. IEEE, 2020.
12. Na Bangchang, Kannat, Sangdao Wongsai, and Teerawat Simmachan. "Application of Data Mining Techniques in Automobile Insurance Fraud Detection." In *Proceedings of the 2023 6th International Conference on Mathematics and Statistics*, pp. 48-55. 2023.
13. Simmachan, Teerawat, Weerapong Manopa, Pailin Neamhom, Achiraya Poothong, and Wikanda Phaphan. "Detecting fraudulent claims in automobile insurance policies by data mining techniques." *Thailand Statistician* 21, no. 3 (2023): 552-568.
14. Salmi, Mabrouka, and Dalia Atif. "Using a data mining approach to detect automobile insurance fraud." In *International Conference on Soft Computing and Pattern Recognition*, pp. 55-66. Cham: Springer International Publishing, 2021.
15. Hanafy, Mohamed, and Ruixing Ming. "Machine learning approaches for auto insurance big data." *Risks* 9, no. 2 (2021): 42.

16. Averro, Nathanael Theovanny, Hendri Murfi, and Gianinna Ardaneswari. "The Imbalance Data Handling of XGBoost in Insurance Fraud Detection." In *DATA*, pp. 460-467. 2023.
17. Okagbue, Hilary I., and O. Oyewole. "Prediction of automobile insurance fraud claims using machine learning." *The Scientific Temper* 14, no. 03 (2023): 756-762.
18. Abdallah, Aisha, Mohd Aizaini Maarof, and Anazida Zainal. "Fraud detection system: A survey." *Journal of Network and Computer Applications* 68 (2016): 90-113.
19. Menardi, Giovanna, and Nicola Torelli. "Training and assessing classification rules with imbalanced data." *Data mining and knowledge discovery* 28 (2014): 92-122.
20. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794. 2016.