A new machine learning approach to detecting insurance fraud

Liang Hong*[1] and Haopeng Yang †[2]

[1]University of Texas at Dallas.
[2]University of Texas at Dallas.

Abstract

Detecting insurance fraud is one of the most important problems for the insurance industry. While different insurers have different fraud claim detecting systems, the general process can be described as follows. When a new claim arrives, it will first go through an initial screening process which is often an automated system based on a statistical method. If a claim is flagged, it will be singled out for further investigation; otherwise, it will be paid immediately. The process of evaluating a potentially fraudulent claim can be complicated and costly. It usually involves many human components, such as adjusters, special investigators, prosecutors, lawyers, and judges; see, Derrig (2002), for a detailed review. Since each insurer faces a large amount of claims every year, it is practically impossible for the insurer to thoroughly investigate every single incoming claim. In addition, if a fraudulent claim passes the initial screening, it will be paid as a valid claim. Therefore, it is critical that the initial screening can detect as many fraudulent claims as possible.

Researchers have been investigating this problem for decades. As a result, several statistical and machine learning methods have been proposed for detecting insurance fraud; see Ai et al (2009), Ai et al. (2013), Gomes et al (2021), Tumminello et al. (2023) and references therein. Though the existing insurance fraud-detecting methods make great contributions to the insurance field, none of them is perfect and false positive cases occur frequently. If the false positive rate is relatively low, then more fraudulent claims will pass the initial screening. As a result, fewer claims will be flagged and subject to further investigation. This reduces the cost of investigation, but more fraudulent claims might slip through the initial screening. On the other hand, if the false positive probability is relatively high, then more claims will be flagged. This leads to higher cost of investigation, but fewer fraudulent claims will be able to slip through the initial screening. In practice, an insurer would like to control the false positive rate for many reasons, such as operational budget and risk management procedure. To our knowledge, no extant fraud-detecting methods provide the insurer with such an option. The purpose of this article is to propose such a method for detecting insurance fraud. The proposed method is based on *conformal prediction*—a powerful machine learning method. For

*E-mail address: liang.hong@utdallas.edu
†E-mail address: haopeng.yang@utdallas.edu

a general discussion of conformal prediction, see Shafer and Vovk (2008) and Vock et al. (2005); for applications of conformal prediction to insurance, see Hong and Martin (2021) and Hong (2023). Our method is distribution-free and is applicable regardless of whether the predictors are numerical or categorical. It will also allow the user to control the false positive rate. This is practically important because further evaluation of flagged claims can be costly and an insurer needs to choose the appropriate false positive rate according to its own resources.

Keywords: fraud detection; guaranteed false positive rate; predictive analytics.

References

[1] Ai, J., Brockett, P. L., Golden, L. L. and Montserrat, G. (2013). A robust unsupervised method for fraud rate estimation. *Journal of Risk and Insurance* 80(1), 121–143.

[2] Ai, J., Brockett, P. L. and Golden, L. L. (2009). Assessing consumer fraud risk in insurance claims: an unsupervised learning technique using discrete and continuous predictor variables. *North American Actuarial Journal* 13(4), 438–458.

[3] Derrig, R.A. (2002). *Journal of Risk and Insurance* 69(3), 271–287.

[4] Gomes, C., Jin, Z., and Yang, H. (2021). Insurance fraud detection with unsupervised deep learning. *Journal of Risk and Insurance* 88(3), 591–624.

[5] Hong, L. and Martin, R. (2021). Valid model-free prediction of future insurance claims. *North American Actuarial Journal* 25(4), 473–483.

[6] Hong, L. (2023). Conformal prediction credibility intervals. *North American Actuarial Journal,* to appear, https://doi.org/10.1080/10920277.2022.2123364.

[7] Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning* 9, 371–421.

[8] Tumminello, M., Consiglio, A., Vassallo, P., Cesari, R. and Farabullini, F. (2023). Insurance Fraud detection: A statistically validated network approach. *Journal of Risk and Insurance* https://doi.org/10.1111/jori.12415.

[9] Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World.* New York: Springer.