Detecting Insurance Fraud with Isolation Forests

Extended Abstract

Information asymmetries are a key characteristic of many interactions in insurance services (Derrig, 2002; Dionne et al., 2009). Fraud in insurance claims is the exploitation of one of these information asymmetries by policyholders. This paper aims at detecting automobile insurance fraud using isolation forests, a new algorithm of unsupervised learning. Fraud is a significant issue to insurers. As an example, the German Insurance Association (2020) estimates that six percent of all insurance claims are based on fraudulent behavior which leads to estimated costs of approximately five billion Euro per year. These costs have to be reflected in insurance premia and are, therefore, also harmful to honest policyholders. From the perspective of insurers and in the interest of policyholders, the question arises of how insurance fraud can be detected as effectively and efficiently as possible. Machine learning algorithms offer solutions to address these problems in an automated and smart way.

Prior literature has started to study the use of machine learning methods to detect insurance fraud (e.g., Viaene et al., 2002; Caudill et al., 2005; Johnson and Nagarur, 2016; Van Vlasselaer et al., 2017). However, the vast majority of studies relies on methods based on supervised learning such as logistic regression and neural networks. These methods are, therefore, prone to two shortfalls. First, there are typically only few insurance claims identified as fraudulent. This is because the manual identification of insurance fraud is very time and cost intensive. As a consequence, the proposed methods have to rely on few labeled observations, which makes robust estimations difficult. Second, and equally important, many fraudulent claims remain unlabeled (Brockett et al., 2002). In this way, supervised algorithms are prone to only replicate the existing fraud identification mechanisms. New and unknown fraud patterns remain undetected in this way. The aforementioned methods could, therefore, be extended by state-of-the-art unsupervised learning for a practical implementation in daily insurance operations.

In order to address this problem, we use a new method of unsupervised learning, so-called isolation forests (Liu et al., 2008; Hariri et al., 2021). Isolation forests are an ensemble of trees that detect anomalies. We analyze a large proprietary data set of more than 10,000 real automobile insurance claims in 2020 from a German insurance company. We build four different Isolation Forests to capture the multidimensional character of insurance fraud: We model whether the claimed amount is anomalous, whether the accident description is anomalous, whether the way of reporting of the accident is anomalous, and whether the economic motivation for the policyholder is exceptional compared to other peers. Therefore, we obtain four anomaly scores for every insurance claim in the data.

In a validation analysis, we assess whether the anomaly scores are helpful in identifying fraud. Our results show that all four anomaly scores significantly predict fraud. The effect size is economically significant. This is also the case taking the risk score from an inhouse automated rule-based fraud detection system into account. Most interestingly, the anomaly scores are particularly helpful in detecting those fraudulent claims that were manually reported as suspicious (i.e., not by the automated rule-based fraud detection system). Thereby, we show that isolation forests can assist

methods based on already identified fraud cases in the process of detecting unknown patterns of insurance fraud.

As our main contribution, we fill a gap in the literature on insurance fraud detection by further addressing the use of unsupervised learning in fraud detection. In particular, we show the usefulness of isolation forests, a new ensemble method of unsupervised learning, to efficiently detect automobile insurance fraud. This has important implications for insurance companies since isolation forests are an easy-to-adapt algorithm that helps to detect and potentially deter insurance fraud, which is a main driver of costs for insurers.

References

Brockett, P. L., R. A. Derrig, L. L. Golden, A. Levine, and M. Alpert (2002). Fraud classification using principal component analysis of RIDITs. Journal of Risk and Insurance 69(3), 341–371.

Caudill, S. B., M. Ayuso, and M. Guillen (2005). Fraud detection using a multinomial logit model with missing information. Journal of Risk and Insurance 72(4), 539–550.

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

Dionne, G., F. Giuliano, and P. Picard (2009). Optimal auditing with scoring: Theory and application to insurance fraud. Management Science 55(1), 58–70.

German Insurance Association (2020). Insurers' concern: Corona gives boost to fraudsters.

Hariri, S., M. C. Kind, and R. J. Brunner (2021). Extended isolation forest. IEEE Transactions on Knowledge and Data Engineering 33(4), 1479–1489.

Johnson, M. E. and N. Nagarur (2016). Multi-stage methodology to detect health insurance claim fraud. Health Care Management Science 19(3), 249–260.

Liu, F. T., K. M. Ting, and Z.-H. Zhou (2008). Isolation forest. 2008 Eighth IEEE International Conference on Data Mining, 413–422.

Van Vlasselaer, V., T. Eliassi-Rad, L. Akoglu, M. Snoeck, and B. Baesens (2017). Gotcha! Network-based fraud detection for social security fraud. Management Science 63(9), 3090–3110.

Viaene, S., R. A. Derrig, B. Baesens, and G. Dedene (2002). A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. Journal of Risk and Insurance 69(3), 373–421.