Machine Learning for Insurance Fraud Detection

Maria Chousa Santos[1(✉)], Teresa Pereira[2], Isabel Mendes[3], and António Amaral[4]

[1] University of Minho, Braga, Portugal
a89166@alunos.uminho.pt
[2] School of Engineering (DIS), University of Minho, Guimarães, Portugal
[3] School of Technology and Management (ESTGA), University of Aveiro, Águeda, Portugal
[4] Polytechnic of Porto - School of Engineering (ISEP), Porto, Portugal

Abstract. Fraudulent activities are a complex problem, and still evolve in a continual basis in all company sectors. These activities are considered as one of the major difficulties the insurance companies have to deal with on a daily basis. Thus, insurers are looking for ways to effectively manage, control, and mitigate fraud. In addition, improving profits by minimizing fraud is the main goal. The exponential amount of information collected, and the technology evolvement has been a strategy to address frauds. The Internet of Everything enables organizations to access diverse information's resources through the interconnection of people-to-machines, which involves machines, data and people, contributing to increase their knowledge and intelligence. In the world of technology, Machine Learning has been widely implemented in multiple contexts. The insurers companies start using Machine Learning to support the detection of fraudulent complaints through the application of algorithms aimed to find patterns in a database, which are hidden through a large amount of data. This paper intends to present the use of Machine Learning technology to support the insurers companies to detect fraudulent activities and further analyze the impacts of technology in people and thus enable to achieve a more rapid and accurate information.

Keywords: Insurance · Fraud · Machine Learning-IoE · Artificial Intelligence

1 Introduction

The core of a company's mission revolves around its people, as they are essential to its growth. The Internet of Everything (IoE) refers to the interconnectedness of people, processes, data, and things, which enabling real-time data extraction and analysis, with a focus on automating processes for the benefit of individuals and organizations. Previously separate entities are now connected to the internet,

T. Pereira et al. (Eds.): IOECON 2023, LNICST 551, pp. 56–65, 2024.
https://doi.org/10.1007/978-3-031-51572-9_5

including machine-to-machine (M2M), person-to-machine (P2M), and person-to-person (P2P) systems. The convergence of these elements through the IoE allows businesses to leverage their benefits and accomplish their objectives [17].

Technology is clearly an important factor in the IoE. Langley et al. [13] divide the IoE into 4 areas:

- IoT - is a subset of IoE, which connects things to provide new possibilities for enhancing the level of intelligence of things;
- Data - is considered the key, enabling technological development, driving the growth of smart things;
- Artificial Intelligence - where "intelligent" things are understood as objects that are sensing, reasoning and performing actions based on input data to achieve a certain predefined goal;
- Semantic Interoperability - the ability of heterogeneous devices to understand each other.

Insurers companies encounter a significant challenge in identifying and detecting fraud committed by policyholders. Given the severity of this issue and the substantial financial losses it can incur for insurers, this work aims to present a solution to mitigate or minimize the impact of fraud in these organizations.

The aim of this study is to explore a solution that can aid in identifying fraudulent claims, thereby minimizing the harm caused to organizations. Additionally, the study presented seeks to address the main research question: how do models inherent to Machine Learning assist and enhance fraud detection measures in insurance companies?

This paper is organized as follows: Sect. 2 provides an overview of the insurance industry, with a specific focus on auto insurance. This section will also provide information on fraud and its detection; Sect. 3 deepens the study in a more specific manner, exploring the use of Artificial Intelligence and Machine Learning technologies in detecting fraud, and presenting possible solutions; lastly, in Sects. 4 and 5 are presented conclusions and outlines future work, respectively.

2 Insurance Business Area Overview and Fraud

In 2020, the insurance penetration rate in OECD (Organization for Economic Co-operation and Development) countries was 9.4 %. This suggests that the insurance industry plays a significant role as a key component of the global economy [12]. Why is insurance so important? Insurance plays a critical role in a company's growth by protecting its financial well-being. Insurance companies provide information about potential risks and the likelihood of loss, thereby minimizing investment risk. As a result, insurers can motivate companies to take a long-term view and increase their risk tolerance. Insurance companies also make a significant contribution to the growth of the capital market in the global economy by managing substantial assets, enabling them to mobilize national savings and bridge the investment gap in emerging markets [7].

2.1 Fraud

Fraudulent activities, such as malicious acts, pose different financial and legal problems in different markets. Within the insurance industry, a significant number of fraud cases are identified each year when, according to the report on insurance fraud presented by FRISS in 2022, 18 % of claims in 2020 were fraudulent and in 2021 this figure rose to 20 % [8], and it is estimated that fraudulent claims account for 10 % of claims costs in Europe [19].

A study has identified a generic fraud control model that an insurer goes through when it receives a new claim, as shown in Fig. 1. This model consists of three distinct phases. The first phase, focuses on determining the nature of the complaint. It can be resolved quickly and at minimal costs if it is determined to be non-suspicious.

However, if the complaint is considered suspicious, it requires a human investigation, resulting in higher costs and resource utilization.

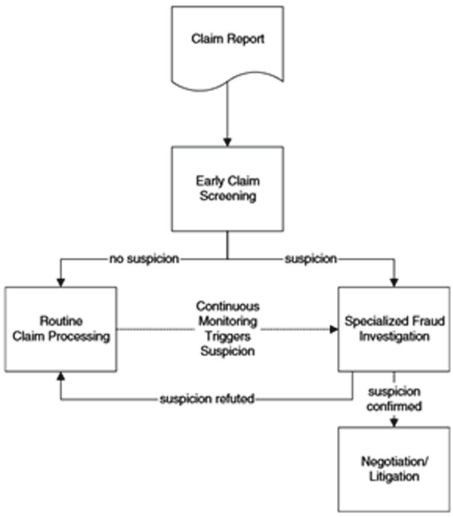

Fig. 1. Fraud control model, source [21]

The second phase, known as the investigation, involves the use of specialized investigators who conduct a rigorous and meticulous investigation to uncover the true nature of the complaint, whether it is fraudulent or not. Finally, in the last phase, the fraudulent claim is dealt with either through negotiation or litigation [21].

The costs associated with investigating and resolving fraudulent claims can be a significant financial burden for insurers, which ultimately impacts honest customers through higher premiums. Furthermore, such fraudulent activities create intense competitive pressure within the insurance market, compelling insurers to enhance the attractiveness of their products to customers [9].

While investigating suspicious claims can be a valuable exercise, it is important to note that it comes with an additional cost to the insurer. The costs of investigating fraudulent claims often exceeds the amount recovered, resulting in financial loss to the insurer. Consequently, the costs of fraudulent claims affects the profitability of the insurer and can lead to higher premiums for honest customers.

2.2 Car Insurance

Regarding the products that insurers provide is auto insurance, which has been found to be significantly impacted by fraudulent activity, according to several research studies [20]. Car insurance is mandatory in the European Union, and covers bodily injury and property damage incurred by third parties and passengers. When it comes to this type of insurance, there are many factors that affect customers and their insurance premiums. Some of these factors include the driver's experience, age, and vehicle specifications, as well as the expected repair costs in the event of an accident [4].

2.3 Fraud Detection

Although there are several approaches to mitigating fraud in the insurance industry. There is no definitive solution that is both completely effective and efficient. One possible measure is the sharing of relevant information among insurers to facilitate the detection of potential fraud. Border cooperation, the creation of a dedicated fraud investigation team, partnerships with law enforcement agencies in different countries, and training for insurers' employees can also help in reducing fraudulent activity [19].

The ways in which fraud is reduced and detected have been evolving with the emergence of new technology. Increasingly, researchers are exploring potential solutions that involve Machine Learning and Artificial Intelligence to develop predictive techniques that can improve the detection of fraudulent activity. The following section, introduces Artificial Intelligence approaches to reducing fraud.

3 Artificial Intelligence Approach in Insurance Fraud Detection

As technology evolves, insurances exploit the potential of those technologies developing applications to detect and mitigate fraud. Increasingly, potential solutions using Machine Learning and Artificial Intelligence are being explored to implement predictive fraud detection techniques.

Artificial Intelligence and Machine Learning can be a valuable starting point to address part of this problem: Artificial Intelligence automatically learns from data, enabling to make predictions and estimate actions. Additionally, Machine Learning techniques can further extend the potential for prediction-making by uncovering implicit or concealed correlations [6].

Benedek and László [4] suggested using data mining algorithms to address fraud detection. These are designed to search for patterns in a database that may be difficult to identify due to the large amount of data involved. This study aimed to determine the most effective of three methods: Decision Tree, Neural Networks, and Naive Bayes. The researchers identified key factors in analyzing the data, including vehicle make, accident location, policyholder age and marital status, policy type, and the vehicle category and price. After analysing the data, the researchers concluded that the Decision Tree algorithm outperformed the other methods and was therefore the best choice.

Alternatively, some studies approach fraud detection as a classification issue and propose methods such as Logistic Regression, Decision Tree, K-Nearest Neighbor, Bayesian Neural Networks, among others [9]. While there are numerous studies detailing strategies to reduce fraudulent activity, most of them rely on the implementation of Machine Learning techniques and their corresponding algorithms.

3.1 Machine Learning

Machine Learning is a type of Artificial Intelligence that enables machines to effectively interpret data through data learning. It uses mathematical models called algorithms to collect and analyze data to make a final decision that matches the user's request. This technology has the ability to perform tasks automatically and to learn from the data that is provided to it [15].

Machine Learning offers several types of analysis, two of which are classification and regression analysis. Classification analysis uses algorithms to make decisions or predict behavior, making it relevant to fraud detection. Regression analysis, on the other hand, attempts to predict the outcome of a continuous variable based on one or more values. These two analyses differ in their predictive content, with classification predicting variables from different classes and regression providing continuous prediction. Since the goal of this study is to identify suspected fraudulent claims, classification analysis is the most appropriate. Using algorithms such as the Decision Tree, Naive Bayes and other classification algorithms, suspicious requests are flagged based on the results of this analysis, since this type of algorithm tries to predict a certain behavior using variables from different classes.

3.2 Technologies Solutions

Today, there are solutions that can identify cases of fraud. FRISS is as an example of an organization that provides such solutions. Among the products offered, by this organization is one dedicated to detecting fraudulent activity. This specific product uses Artificial Intelligence and transparent predictive models to prevent fraud in real-time. This creates a scenario where the insurer can avoid paying the customer's claim. [10]. SAS provides cutting-edge technologies powered by Artificial Intelligence that enable faster, more efficient analysis. Among the software solutions that SAS offers, it focus on fraud detection and prevention.

SAS Detection and Investigation for Insurance is an advanced software program designed to identify, prevent and monitor fraudulent behavior associated with insurance claims. This software leverages a range of analytical techniques and Artificial Intelligence technologies, such as Machine Learning and deep learning, to improve its detection capabilities. [18].

SHIFT is a company that provides decision-making solutions based on Artificial Intelligence. The company is focused on providing solutions that enable insurance companies to automate and streamline their decision-making processes, resulting to increased operational efficiency and reduced costs. One of the many solutions provided by SHIFT is the Shift Claims Fraud Detection tool. This tool is capable of detecting fraud in claims in real-time and provides all the relevant information about the claim, along with a detailed explanation of the conclusion reached [11].

Linkurious is a company that specializes in graph intelligence solutions that help to easily connect complex data to enable faster more informed business decisions. Its anti-fraud solution combines Machine Learning with graph analytic to create a network of customers and their connections in a matter of seconds. By doing so, it provides the ability to see the full context of a customer or suspicious claim, filter the data, explore other relationships, and ultimately dismantle fraud networks [14].

The Table 1 provides a clearer overview of the characteristics of the solutions offered by the companies.

Table 1. Solution's Comparison.

	FRISS	SAS	Shift	Linkurious
Velocity of investigations	–	–	3x faster	Increase up to 10x
Number of insurers and countries involved	Over 200 insurers in more than 40 countries	Almost 150 countries	More than 25 countries	2000 companies
Implementation Time	4 months	–	4 months	–
Number of risk assessments per year	26 million	–	Hundreds of millions	–
Reduce case alert volume	–	40 %	–	–
Expose fraud	–	35 %	–	Up to 20 %

The data presented in the table indicates that the average implementation time for each solution is four months, a reasonable and typical duration that can yield significant benefits to the implementing company. The table shows several metrics that support the argument, such as the millions of annual assessments, the reduction in alert cases, and the expedited investigations. These indicators

ultimately lead to the primary issue addressed in the table, which is the mitigation of fraud exposure. Typically, increasing the speed of investigations allows companies to handle cases more efficiently, resulting in greater customer satisfaction with the promptness of the process. In addition, as noted above, minimizing the risk of fraud provides benefits to organizations, such as reduced costs and improved financial stability.

4 Adapting Data and Using Machine Learning

Analyzing the insurance company's data is a crucial step in applying Machine Learning models. Studying the variables individually, taking into account the description of each column, the values it aggregates, the type of data, in order to help clean the data and make any necessary changes, are some steps required to clean the data. Once the data has been cleaned and the necessary changes have been made, the relationships between the various attributes are assessed using a correlation matrix, i.e., Pearson Correlation, Spearman, Kendall, among others. These matrix's have different objectives, Pearson's correlation focuses on the strength of the linear relationship between two vectors of data, while Spearman's correlation only describes the monotonic degree [22]. Kendall's correlation assesses the similarity between two sets of classifications on the same data [1]. Among these correlations, the one that is most appropriate and the one that will be applied is Pearson's correlation, since it obtains linear correlations of quantitative variables.

The Pearson correlation matrix provides values between 0 and 1 (absolute) where the higher the value, the greater the relationship between the attributes, known as coefficients. Correlation coefficients are generally used as a first tool to quantify, visualize and interpret relationships between different attributes. Once the correlations have been assessed, it is possible to differentiate between certain attributes and remove those with a very weak correlation, in order to obtain a dataset with a set of specific attributes.

Azure offers various services and tools for applying DevOps practices that focus on automating and optimizing processes using this technology, including continuous integration and continuous delivery (CI/CD) [5]. Azure Pipelines is one example. This allows CI/CD to be implemented to create, test and deploy continuously, on any platform and in any cloud [2]. Azure currently offers the ability to study the best Machine Learning model for a given data set and train it continuously. In this way, any insurance company can apply Machine Learning models to its data without requiring much effort or knowledge on the subject. Despite this Azure offering, each insurance company also has the opportunity to choose its own model and apply it.

However, it is necessary to mention the need to carry out tests in order to detect flaws and possible improvements, where the effectiveness of a Machine Learning model in identifying fraudulent cases as correct or incorrect can be related to various causes, including: the quality of the data, limitations of the algorithm and also the complexity of the problem.

5 Conclusion and Future Work

In conclusion, the impact of fraudulent claims on the insurance market cannot be overstated, as it can destabilize economies and affect people's cost of living [3]. These consequences highlight the critical need for organizations to invest in advanced technologies to prevent fraud and protect their reputations and financial stability, since traditional methods such as hiring companies focused on assessing claims can be rather time-consuming and may not even result in a reliable response. In this context, methods involving Machine Learning are increasingly considered the best alternative, as they allow insurers to analyze the relationships between various data automatically, leading to a shortening of investigation time.

Artificial Intelligence and Machine Learning, in particular, offer promising solutions to address this issue. The potential for these technologies to transform the insurance industry was recognized by Forbes Magazine, which named Artificial Intelligence as one of the top technology trends for 2022 [16]. As Artificial Intelligence continues to advance, it is likely to become an increasingly important investment for businesses to ensure their continued success.

With this in mind, Machine Learning is poised to become a prudent investment for companies in the near future, especially in the area of fraud detection. It is important the use of classification algorithms, such as Decision Tree, Bayer Naive and Neural Net to speed up fraud investigations and identify larger fraud schemes. These algorithms have shown promising results in generating positive financial outcomes for both companies and their customers.

It's also important to focus on some future work, to obtain some results based on the research presented in this paper. Is intended to put the analysis into practice by using a dataset related to auto insurance fraud. Then apply Machine Learning models, such as Decision Tree, Bayesian Classifiers, and Neural Networks. These models are well suited for pattern recognition in complex datasets and are effective in fraud detection.

However, the development of an application that can efficiently and effectively detect fraud claims using the Machine Learning is expected. Such an application would be a valuable tool for insurance companies, allowing them to detect and prevent fraud, potentially saving them millions of dollars.

Overall, companies must remain vigilant and stay aligned with the of technological advancements in order to protect themselves from fraudulent activities. By investing in emerging technologies, such as Artificial Intelligence and Machine Learning, companies can identify patterns in large data sets that are usually difficult to detect, be more efficient, provide a faster response, safeguard their reputation, protect their customers and above all preventing fraud and maintain their financial stability in the market.

References

1. Abdi, H.: The kendall rank correlation coefficient (2007)
2. Azure Pipelines (2023). https://azure.microsoft.com/en-us/products/devops/pipelines. Accessed 19 June 2023
3. Benedek, B., Ciumas, C., Nagy, B.Z.: Automobile insurance fraud detection in the age of big data - a systematic and comprehensive literature review. J. Finan. Regul. Compliance **30**(4), 503–523 (2022). ISSN 1358–1988. https://doi.org/10.1108/JFRC-11-2021-0102. https://www.emerald.com/insight/content/doi/10.1108/JFRC-11-2021-0102/full/html
4. Benedek, B., László, E.: Identifying key fraud indicators in the automobile insurance industry using SQL server analysis services. Studia Universitatis Babes-Bolyai Oeconomica **64**(2), 53–71 (2019). ISSN 2065–9644. https://doi.org/10.2478/subboec-2019-0009. https://www.sciendo.com/article/10.2478/subboec-2019-0009
5. DevOps (2023). https://azure.microsoft.com/pt-pt/resources/cloud-computing-dictionary/what-is-devops. Accessed 19 June 2023
6. Eckert, C., Neunsinger, C., Osterrieder, K.: Managing customer satisfaction: digital applications for insurance companies. Geneva Pap. Risk Insur. - Issues Pract. **47**(3):569–602 (2022). ISSN 1468–0440. https://doi.org/10.1057/s41288-021-00257-z
7. Baur, E., Birkmaier, U., Rüstmann, M: The economic importance of insurance in Central and Eastern Europe and the impact of globalisation and e-business (2021)
8. FRISS - Insurance fraud report 2022 (2023). https://www.friss.com/insight/insurance-fraud-report-2022/. Accessed 20 Jan 2023
9. Galeotti, M., Rabitti, G., Vannucci, E.: An evolutionary approach to fraud management. Eur. J. Oper. Res. **284**(3), 1167–1177 (2020). ISSN 03772217. https://doi.org/10.1016/j.ejor.2020.01.017. https://linkinghub.elsevier.com/retrieve/pii/S0377221720300382
10. Home. FRISS (2023). https://www.friss.com/. Accessed 23 Jan 2023
11. Home (2023). https://www.shift-technology.com/?hsLang=en. Accessed 24 Jan 2023
12. Insurance indicators : Penetration. https://stats.oecd.org/Index.aspx?QueryId=25444. Accessed 20 Jan 2023
13. Langley, D.J., et al.: The internet of everything: smart things and their impact on business models. J. Bus. Res. 122, 853–863 (2021). ISSN 01482963. https://doi.org/10.1016/j.jbusres.2019.12.035. https://linkinghub.elsevier.com/retrieve/pii/S014829631930801X
14. Linkurious — graph intelligence solutions for the enterprise — let us light the way in your connected data (2023). https://linkurious.com/. Accessed 30 Jan 2023
15. Mahesh, B.: Machine learning algorithms -a review. https://doi.org/10.21275/ART20203995
16. Marr, B.: The 5 biggest technology trends in 2022. Forbes. Section: Enterprise Tech. (2021). https://www.forbes.com/sites/bernardmarr/2021/09/27/the-5-biggest-technology-trends-in-2022/ (visited on 01/25/2023)
17. Miraz, M.H., et al.: A review on internet of things (IoT), internet of everything (IoE) and internet of Nano things (IoNT). In: 2015 Internet Technologies and Applications (ITA), pp. 219–224 (2015). https://doi.org/10.1109/ITechA.2015.7317398

18. SAS: analytics, artificial intelligence and data management (2023). https://www.sas.com/pt_pt/home.html. Accessed 24 Jan 2023
19. The impact of insurance fraud (2013)
20. Viaene, S., Dedene, G.: Insurance fraud: issues and challenges. Geneva Pap. Risk Insur. - Issues Pract. **29**(2), 313–333 (2019). ISSN 1018–5895, 1468–0440. https://doi.org/10.1111/j.1468-0440.2004.00290.x. http://link.springer.com/10.1111/j.1468-0440.2004.00290.x. Accessed 20 Jan 2023
21. Viaene, S., Van Gheel, D., Ayuso, M., Guillén, M.: Cost-sensitive design of claim fraud screens. In: Perner, P. (ed.) ICDM 2004. LNCS (LNAI), vol. 3275, pp. 78–87. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30185-1_9
22. De Winter, J.C., Gosling, S.D., Potter, J.: Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: a tutorial using simulations and empirical data. Psychol. Methods **21**(3), 273–290 (2016). ISSN 1939–1463, 1082–989X. https://doi.org/10.1037/met0000079. http://doi.apa.org/getdoi.cfm?doi=10.1037/met0000079. Accessed 19 June 2023