ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΑΝΙΧΝΕΥΣΗ ΑΠΑΤΗΣ ΣΤΙΣ ΑΣΦΑΛΕΙΕΣ ΑΥΤΟΚΙΝΗΤΟΥ ΜΕ ΧΡΗΣΗ ΜΗ ΕΠΟΠΤΕΥΟΜΕΝΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Χαράλαμπος-Παναγιώτης Μιχελάκης

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Ιούνιος 2024

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Μπερσίμης Σωτήριος, Καθηγητής (Επιβλέπων)
- Πολίτης Κωνσταντίνος, Αναπληρωτής Καθηγητής
- Πλαγιανάκος Βασίλειος, Καθηγητής

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

FRAUD DETECTION IN CAR INSURANCE USING UNSUPERVISED MACHINE LEARNING

By

Charalampos-Panagiotis Michelakis

M.Sc. Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
June 2024

Περίληψη

Η ανίχνευση απάτης στις ασφάλειες αυτοκινήτων αποτελεί ένα ζήτημα με σημαντικές οικονομικές και ηθικές επεκτάσεις. Μελέτες δείχνουν ότι οι δόλιες αξιώσεις αποζημίωσης στις ασφάλειες αυτοκινήτων αποτελούν το 10%-20% των συνολικών ασφαλιστικών αξιώσεων που υποβάλλονται στην Κεντρική και Ανατολική Ευρώπη. Για το λόγο αυτό, θα διερευνήσουμε τις δυνατότητες αξιοποίησης μεθόδων μη εποπτευόμενης μηχανικής μάθησης για την αντιμετώπιση αυτού του προβλήματος. Συγκεκριμένα, αυτό το προς έρευνα αντικείμενο παραμένει σχετικά ανεξερεύνητο στη βιβλιογραφία ανίχνευσης ασφαλιστικής απάτης, η οποία κατά κύριο λόγο επικεντρώνεται σε ένα περιορισμένο σύνολο μεθόδων μη εποπτευόμενης μηχανικής μάθησης. Η δουλειά μας υιοθετεί μια ευρύτερη προσέγγιση όσον αφορά τις μεθόδους που χρησιμοποιούνται, αντλώντας έμπνευση από τον γενικότερο και ταχέως εξελισσόμενο τομέα της ανίχνευσης ανώμαλων/εκτρόπων παρατηρήσεων. Όσον αφορά την αξιολόγηση αυτών των μεθόδων, αυτή θα διεξαχθεί μέσω μελέτης προσομοίωσης, καθώς η εύρεση δημόσια διαθέσιμων πραγματικών συνόλων δεδομένων, (λόγω του εμπιστευτικού χαρακτήρας τους), είναι εξαιρετικά δύσκολη και αποτελεί σημαντική πρόκληση στην έρευνα του της ανίχνευσης απάτης στις ασφάλειες αυτοκινήτων. Η επιλογή μιας μελέτης προσομοίωσης είναι ο τρόπος με τον οποίο θα «παρακάμψουμε» αυτό το εμπόδιο. Τα προσομοιωμένα σύνολα δεδομένων μας θα είναι το αποτέλεσμα μιας «συνθετικής ανακατασκευής» ενός συνόλου δεδομένων πραγματικού κόσμου, το οποίο χρησιμοποιείται ως "πηγή" για τη δημιουργία τυπικών/μη-δόλιων δειγμάτων δεδομένων, τα οποία στη συνέχεια αναμιγνύονται με πολλούς διαφορετικούς τύπους παραμετρικά δημιουργημένων συνθετικών εκτρόπων παρατηρήσεων. Η δουλειά μας, λοιπόν, θα ολοκληρωθεί με την σύγκριση της απόδοσης σχεδόν τριάντα διαφορετικών αλγορίθμων ανίχνευσης εκτρόπων παρατηρήσεων σε πέντε διαφορετικά (συνθετικά) σενάρια τέτοιων τιμών, η οποία θα μπορούσε να παράσχει νέες πληροφορίες για την καταπολέμηση της απάτης στις ασφάλειες αυτοκινήτων χρησιμοποιώντας μη εποπτευόμενη μηχανική μάθηση.

Abstract

The detection of fraud in automobile insurance holds significant economic and ethical implications. Studies suggest that fraudulent automobile insurance claims account for 10%-20% of total claims submitted in Central and Eastern Europe. We will explore the possibilities of leveraging unsupervised machine learning methods in tackling this problem. Notably, this research area remains relatively unexplored within the insurance fraud detection literature, which predominantly focuses on a limited set of unsupervised machine learning methods. Our work takes a much broader approach regarding the methods used, drawing inspiration from the more general and rapidly evolving domain of anomaly/outlier detection. Regarding the evaluation of these methods, it is conducted by means of a simulation study, as the scarcity of publicly available real-world data sets, due to their confidential nature, poses a significant challenge in researching automobile insurance fraud. The choice of a simulation study is our way of circumventing this “roadblock”. Our simulated data sets are the outcome of a “synthetic reconstruction” of a real world data set, which is used as a “seed” for the generation of typical/non-fraudulent data samples which are then augmented by several different types of parametrically created synthetic outliers. The culmination of our work is the performance comparison of almost thirty different outlier detection algorithms across five different synthetic outlier scenarios, which could provide new insights for combating fraud in automobile insurance using unsupervised machine learning.

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

In this study we will tackle the problem of automobile insurance fraud detection by utilizing unsupervised machine learning methods. In the first chapter we will give an overview of insurance fraud, the motivation for its detection and challenges faced when attempting to do so; in this introductory chapter we will not limit ourselves to automobile insurance but we will make special mention of issues that are particularly relevant to automobile insurance. The following chapter will give an overview of the literature on automobile insurance fraud with a special emphasis placed on parts of the literature that utilize unsupervised machine learning methods. The subsequent chapter will present the unsupervised machine learning methods we will be using in our empirical application, while also providing a brief comparison between supervised and unsupervised machine learning along with some remarks on the particular properties of the latter. Our empirical application will constitute the next chapter. We will be conducting a simulation study for the comparison of the performance of unsupervised anomaly detection methods which can be applied in insurance fraud detection. The simulations will be based on a real automobile insurance claims dataset; we will obtain a parametric representation of it which we will use in order to generate synthetic typical (i.e. non-fraudulent) observations. We will be generating outliers which stand for the atypical (i.e. fraudulent) observations by the use of five different parametric techniques. The various outlier detection methods will be evaluated across these five different cases. The results of our simulations will be presented in the final section of our empirical application, along with any observation gleaned from these results. The final chapter will include concluding remarks on our work and comments on areas of further possible research

2 The Nature Of Insurance Fraud, Issues, Challenges And The Motivation For Its Detection

2.1 Insurance and Insurance Fraud

Since antiquity people made arrangements among themselves in order to mitigate risk by transferring and distributing it. In our times, this is mainly done through the insurance industry which has become a staple of modern societies. There is huge variety of different forms insurance takes in order to cover an extremely diverse number of potential losses. We will mention indicatively some of the more prominent forms of insurance: health insurance, life insurance, property and casualty (P&C) insurance, automobile insurance, liability insurance and credit insurance.

To paraphrase Stijn Viaene and Guido Dedene (2004), insurance is a contractual relationship between two parties: the insurer (also called underwriter) and the insured party (also called policyholder). The insurer agrees to make monetary provision on behalf of the insured party to cover the loss of an insurable interest due to one or more future, well-defined, but uncertain events. The insured party agrees to provide a relatively small payment to the insurer (called a premium) in exchange for the contractual obligation assumed by the insurer.

Any monetary compensation by the insurer may only be provided after a (first- or third-) party, the claimant party, files a formal claim for a loss covered by the contract. As Stijn Viaene and Guido Dedene (2004) note, all parties transacting in the context of this contract are required by law, to act with the utmost good faith toward one another at all times. This in turn obliges them to reciprocally disclose all material information known to them.

The existence of insurance is accompanied by the existence of insurance fraud. For example Ken Dornstein (1996) provides an account of different forms and instances of such fraud as early as the beginning of the twentieth century. Insurance fraud may take many different forms. As such we will not restrict ourselves to the automobile sector, as that would hinder the presentation of a quite complex and multifaceted phenomenon. We will however go into further depth and mention specifics regarding the automobile sector where appropriate. In this way we will illuminate the main motivation behind the detection of insurance fraud, while at the same time presenting the main challenges which accompany it[1].

According to Duffield and Grabosky (2001) "in its broadest terms, fraud means obtaining something of value or avoiding an obligation by means of deception". Insurance fraud, in particular, is a subject fraught with issues. We quickly become aware of that by the fact that there is no common and generally accepted definition of insurance fraud (Benedek, Ciumas, and Nagy 2022).

According to the same authors, one may use its legal definition, the most common worldwide being the Massachusetts Regulation (211 CMR 93. 03) which defines fraudulent claims as "claims submitted with the intent of receiving a larger payment from the insurer than the amount, if any, to which the claimant is entitled under the policy, including claims for: nonexistent losses; amounts in excess of actual losses; or incidents which the claimant has arranged in an effort to receive an insurance payment" (Massachusetts Regulation, 1993). They believe that this definition fails to cover some types of fraud such as misrepresentation or intentional recklessness due to the insurance coverage[2].

In any case, the meaning of fraud in legislation varies from place to place and is insufficient in describing what is considered fraud in practice: the term is often used broadly to encompass abuse of insurance and may also frequently be used without an implication of direct legal consequences according to Stijn Viaene and Guido Dedene (2004). The go on to point out that "the concept of insurance fraud is most often associated with, and sometimes reduced to, the case of deliberately inflated, false or fictitious claims (claim fraud)."

2.2 Insurance Fraud Typology

Despite the fact that it is common for one to refer to claim fraud when speaking of insurance fraud, we will devote some time in going over some of the different types of insurance fraud. Stijn Viaene and Guido Dedene (2004) classify insurance fraud based on three mutually exclusive and opposite characteristics: 1) internal vs. external 2) underwriting vs. claim, 3) soft vs. hard.

[1]As each insurance sector is characterized by its own idiosyncrasies, beyond any challenges that are common across sectors, we will mainly focus our efforts here on the presentation of the challenges that are endemic to the automobile insurance sector.

[2]This behavior constitutes a textbook example of the concept of moral hazard.

Internal fraud is perpetrated by insiders of the industry, while external by outsiders. Internal insurance fraud would fall under the "jurisdiction" of operational risk management and as such our work will solely focus on external fraud. Underwriting fraud concerns fraudulent acts which happen at the time of underwriting or renewal of an insurance contract. While the detection of fraud at underwriting time would be an interesting research field, it is outside the scope of this work; we will target our efforts on fraud which happens at claim time. We suggest that the interested reader looks at the work of Nagrecha, Johnson, and Chawla (2018) for research in this direction

Finally, arguably the most important distinction between types of fraud is that of soft versus hard fraud. We could also frame this distinction as opportunistic versus planned fraud as the terms soft/opportunistic and hard/planned are interchangeable. So called soft fraud refers to the phenomenon of typically honest people acting in an opportunistic unwanted manner. The typical example is that of a policyholder who has a legitimate reason for submitting a claim but opportunistically inflates the damages submitted in the claim. This behavior is typically called *claim padding* or *build-up*. On the other hand, hard fraud typically describes criminal offenses (Richard A. Derrig 2002). It involves a premeditated attempt of dishonestly making monetary gains at the expense of the insurance industry. No legitimate claim exists at any point in time in the case of hard fraud. The claims are completely fictitious and while they may often be the work of a single individual, they are also frequently perpetrated by well-organized fraud rings. For example, in automobile insurance, our sector of interest, a case of hard fraud may be a conspiracy involving the claimants in conjunction with medical professionals and/or automotive repair shops and others. It is easy to assume that such schemes pose extreme danger to automobile insurance companies.

2.3 Motivation For The Detection And Deterrence Of (Automobile) Insurance Fraud

In this section we will present the motivation behind our work. Insurance fraud is a critical multifaceted problem which presents serious not only for the insurance industry but also for the wider public. These concerns are mainly of a financial and legal nature[3].

In the United States only, the insurance industry consists of more than 7000 companies that collect over one trillion U.S. Dollars (Federal Bureau of Investigation 2023). As Stijn Viaene and Guido Dedene (2004) point out "insurance, by its very nature, is especially prone to fraud. Information asymmetries leave the players with no option other than to trust each other at transaction time. Due to the absence of perfect information, many opportunities naturally arise in which one or more of the parties involved have a clear economic incentive to commit fraud, either premeditated or opportunistic".

The combined effect of the economic size of the industry and its susceptibility to fraud results in huge financial losses due to fraudulent activity. Despite the aforementioned fact, until the late 1980s there were no attempts either at the industry or firm level to systematically quantify the extent of the cost

[3] We should not however dismiss the importance of more nuanced concerns like consumer protection or ethical considerations.

of insurance fraud. However, since then, several sources began efforts in this direction (Stijn Viaene and Guido Dedene 2004).

According to the FBI the total cost of insurance fraud excluding health insurance is more than $40 billion per year. Turning our attention to Europe, the European insurance and reinsurance federation[4] estimated that during 2017 total fraudulent claims in Europe were approximately worth 13 billion Euros (Insurance Europe 2019). In a document released in 1996 by the same organization it was claimed with utmost conviction that the insurance fraud that is discovered is only a limited subset of the fraud that takes place. There is a considerable gap between them (Comité Européen des Assurances 1996). This view is supported by the work of Coalition Against Insurance Fraud (CAIF). In 2022 the CAIF released a report claiming that the situation in the United States is much worse than what the FBI stated. According to their estimates the yearly cost to consumers is $308.6 billion[5].

Our sector of interest is particularly prone to fraudulent activity. The automobile insurance sector is widely believed to be among the most affected by fraud (Weisberg and Richard A Derrig 1998; Georges Dionne and Laberge-Nadeau 1999; E.-B. Belhadji and Georges Dionne 1998; Stijn Viaene and Guido Dedene 2004). In the USA and Western Europe 7%-10% of the policies are believed to be affected by fraud. This figure is even greater in the Central and Eastern European regions where it is estimated to be in the range of 10%-20%. The most extreme example commonly quoted in the literature is China where the minimum estimate is 18% of the policies with the highest being 20%.(Benedek, Ciumas, and Nagy 2022; Insurance Information Institute 2023)

The cost of fraud burdens not only the insurance industry but also the consumer. In order to ensure their viability and profitability despite the financial losses due to insurance fraud the insurance companies pass (at least part of) this cost to their customers. Any insurance taker is either directly (e.g. through lost savings) or indirectly (e.g. through higher premiums) negatively affected by insurance fraud[6] (Stijn Viaene and Guido Dedene 2004). Consequently the cost of life as a whole is increased for the average citizen (Stijn Viaene, Stijn Viaene, et al. 2007; Stijn Viaene and Guido Dedene 2004). Stijn Viaene and Guido Dedene (2004) believe that insurance fraud "may be extremely detrimental to established social and economic structures".

Because of this reality, it is in the interest of both insurance companies and honest policyholders to combat insurance fraud. A reduction in insurance fraud will help both the economic viability of companies in the insurance sector but also lead to lower, more affordable and more fair insurance premiums for honest costumers. As a result the main motivation behind the detection of insurance fraud becomes clear. This is especially true in the automobile insurance sector, which will be our focus, due to the high prevalence of fraudulent activity.

The gains accrued from the detection of insurance fraud are not limited to those mentioned previously. We will now mention (not exhaustively) some of the added benefits. Effective identification of insurance fraud can act as a

[4]The federation is now known as Insurance Europe. For most of its life (until March 2012) it was known by its founding name, Comité Européen des Assurances (CEA).

[5]In contrast to the FBI they include Life Insurance, estimating that its cost is $74.7 billion.

[6]Some segments of the population are more vulnerable to some types of fraud. It is particularly troubling that, according to the authors, some of the more negatively affected are vulnerable segments of our society like the elderly and certain immigrant groups

potent deterrent that contributes greatly to the operating robustness of the insurance industry (Tennyson and Salsas-Forn 2002; Picard 1996). Furthermore, the detection of hard fraud may help in preventing organized crime: insurance companies that detect potential criminal activity may cooperate with law enforcement agencies in order to prosecute criminal rings. This would contribute to public safety. The detection and reduction of insurance fraud would also aid in improving the public image of the insurance industry and solidifying consumer trust. This could potentially result in a complementary indirect "source" of fraud reduction: Stijn Viaene and Guido Dedene (2004) claim that a significant amount of the cases of soft fraud are guided by a "widespread public feeling of unfairness with regard to insurers".Hence, by improving the public image and the reputation of the sector due to the reduction of fraud, there may be a secondary indirect reduction of cases of soft fraud. Furthermore, by leveraging new technologies certain aspects of the fraud control process could be automated; it would "enable proactivity" and "reduce the investigative process lead time and allow for more optimal allocation of scarce investigative resources" according to Stijn Viaene and Guido Dedene (2004). This would result in overall efficiency gains, streamlining processes and ensuring that honest customers receive timely compensation.

2.4 Issues and Challenges in Detecting (Automobile) Insurance Fraud

In this section we will present some of the challenges one may face when tackling the problem of detecting insurance fraud, with special mentions to the automobile sector.

The most important challenge is the fact that insurance fraud is by its very nature "not self-revealing" (Stijn Viaene and Guido Dedene 2004). While this may seem as an obvious or trivial observation it has some profound consequences. When dealing with insurance fraud, one is not simply trying to detect a phenomenon in the midst of noisy data; the express purpose of the phenomenon under investigation is to "blend in" with legitimate claims and go unnoticed. As such the detection is made harder since any attempts at hiding or obfuscating fraudulent activity must be overcome. Time is also of the essence. "Fraud control is subject to the constraints of speedy detection and minimal investigative lead time." (Stijn Viaene and Guido Dedene 2004). Unless a timely detection of (potential) fraud is made, it is impossible to realize any material benefits as the fraudulent act cannot be effectively prosecuted after any payment has been made and the claim has been settled. Moreover, after some time has passed it may also be impossible to verify if the suspicious claim is truly fraudulent, since any investigation by the Special Investigation Units must take place during the processing of the claim.

Furthermore, fraud is a dynamic phenomenon. Sophisticated criminal actors adjust their schemes in step with any changes in the business environment and are extremely benefited by its complexities as they provide cover for their activities (Stijn Viaene and Guido Dedene 2004). Likewise, as detection methods improve criminal activity evolves to bypass them. A constant "tug of war" between the industry and the criminals takes place.

Fraud also varies by region (Benedek, Ciumas, and Nagy 2022). A multitude of factors contribute to this: most importantly, there may be significant differences in legislation between different regions/countries. Criminal activity may also take different forms in different environments. Various idiosyncratic characteristics of each different region may be responsible for this. For example, when dealing with automobile insurance, one notices that fraudulent bodily injury claims are much more common in the United States and less relevant in Europe. This could be attributed to the lack of universal health care in the USA (Artís, Ayuso, and Guillén 1999).

Stijn Viaene and Guido Dedene (2004) go on to point out a number of additional concerns. According to them, transaction-level monitoring is not enough. "Successful detection of sophisticated fraud schemes generally relies on cross-sectional and longitudinal analysis of context enriched transaction data and rigorous external validation of the veracity of the submitted transaction data". They also suggest that any fraud detection method must not interfere negatively in the processing time of claims. Insurance companies are under heavy competition so claim processing efficiency is required.

In the same vain, they turn our attention to a number of economic considerations of the fraud detection process. The return on any resources spent on the fraud control process is hard to quantify and as a result also difficult to justify to the company's upper management. One should also consider that, at the firm level, there always exists the concern of *"freeriding"*, that is other companies benefiting passively from the fraud control processes of others. This may very well be one of the reasons that since the 1980s a number of organizations with the express purpose of fighting fraud have been established like the Coalition Against Insurance Fraud (CAIF) and the International Assosiation of Special Investigation Units (IASIU)

Based on somewhat similar economic concerns, Benedek and Nagy (2023) point out the there is a lack of systematic comparison and research on the cost-effectiveness of fraud identification. They believe that the performance of any fraud detection system should be judged in terms of its cost-effectiveness

Finally, the raw data itself is associated with a number of issues. We may take the data presented in Debener, Heinke, and Kriebel (2023) as an example. When the available dataset contains labels regarding fraud, (which is not always the case), usually the only claims that are marked as fraudulent are those that have been proven as such. Since proving fraud legally is difficult (see Stijn Viaene and Guido Dedene 2004), these cases are but a small subset of the total fraud taking place. In the same data set mentioned above, we see that an additional label is included, marking highly suspicious claims. However, as that is a subjective judgment made by the company's fraud control staff, we should expect that these data contains cases which are not actually fraudulent (false positives) and also fails to include all actual cases of fraud (i.e. false negatives also exist in the data).

The aforementioned data set related problem is particularly prevalent in the automobile insurance sector. Unlike other cases of fraud[7], in automobile insurance the dependent variable (i.e. whether a claim was fraudulent or not) can not ultimately be verified in the real world in most cases, because it is too costly and/or time intensive. It would require all suspicious cases to be legally prosecuted and court decisions to be rendered. However Brockett, Xia, and Richard A. Derrig (1998) note that insurance companies, especially in the case

[7] For example credit card fraud

of (suspected) soft fraud, avoid resolving the claims in this manner as it is not only costly but also risky. Richard A. Derrig and Ostaszewski (1995)Weisberg and Richard A. Derrig (1991) concur that data sets used in fraud detection in the context of automobile insurance contain in most cases "subjective indicators and classifications". Additionally, we do not really know the extent of the problem as there is a lack of reliable statistics regarding the actual size of this kind of fraudulent activity (Benedek, Ciumas, and Nagy 2022).

Another problem we have to overcome in automobile insurance fraud data sets is that they are highly unbalanced. The same authors argue that "in general, 5%-20% of the claims are fraudulent, which means that a fraud detection model, which classifies all the claims in the legitimate classes, has an overall classification accuracy between 80% and 95%." This fact has to be taken into account when we try to find appropriate methods for dealing with these data sets[8].

Finally, dataset availability is another concern. Insurance companies are hesitant when publishing proprietary information, even more so when it concerns the number of fraudulent claims (Benedek, Ciumas, and Nagy 2022). Oftentimes empirical studies in the field of automobile insurance fraud detection are conducted using the same data sets (Debener, Heinke, and Kriebel 2023).

3 Literature Review

The purpose of our thesis is to explore the potential of unsupervised machine learning methods for accomplishing the task of detecting automobile insurance fraud. However, our literature review will cover any research about insurance fraud, regardless of the sector. Besides unsupervised learning, we will also cover supervised machine learning as well as more "traditional" statistical methodologies in order to be as comprehensive as possible and give a holistic view of the research on this subject. Since many studies combine methodologies and moreover since there is not always a clear delineation between the three aforementioned fields, we will not try to separate the literature into different groups or sections. Our approach will be similar to that of Benedek, Ciumas, and Nagy (2022), and as such our presentation will follow mostly a chronological order.

As we mentioned in the previous chapter till the late 1980s the problem of detecting insurance fraud had not been researched. The first steps in approaching the subject took place during the 1990s. Early research focused mainly on identifying a list of indicators for fraud (see Weisberg and Richard A. Derrig 1991; Richard A. Derrig and Ostaszewski 1995; Weisberg and Richard A Derrig 1998; E. B. Belhadji, George Dionne, and Tarkhani 2000).

To be more specific, in Weisberg and Richard A Derrig (1998) the authors chose the 25 most important fraud indicators out of a list of 65, then tried all linear models that had a subset of 10 of these indicators. Their 5 best performing models had an R^2 of 0.65, and they all contained 10 out of the following 13 fraud indicators. We present those indicators in Table 1. E. B. Belhadji, George Dionne, and Tarkhani (2000) focused on probit models instead of the linear models of Weisberg and Richard A Derrig (1998), since those models could provide probabilities of fraud given each indicator. The indicators they

[8]In Benedek, Ciumas, and Nagy (2022) we see that quite frequently the technique of oversampling cases belonging to the minority class (fraud) is used to overcome this problem

Table 1: Significant Fraud Indicators according to Weisberg and Richard A Derrig (1998)

	Fraud Indicators
1	No report by police officer at scene
2	No witnesses to accident
3	No plausible explanation for accident
4	Claimant in an old, low-value vehicle
5	Property damage was inconsistent with accident
6	Insured felt set up, denied fault
7	Appeared to be "claims-wise"
8	Was difficult to contact/uncooperative
9	No objective evidence of injury
10	Injuries were inconsistent with police report
11	Large number of visits to a chiropractor
12	DC provided 3 or modalities on most visits
13	Long disability for a minor injury

found significant are presented in Table 2. However most of these approaches enjoyed limited success. Richard A. Derrig and Ostaszewski (1995) found that even among experts there is disagreement about which claims are fraudulent. To combat this they employed *fuzzy* classification techniques.

Towards the end of the decade, E.-B. Belhadji and Georges Dionne (1998) proposed methods that belong in the class of "expert systems"[9]. Sternberg and Reynolds (Nov./1997) combined an expert system with cultural algorithms which theoretically would enable the expert system to adjust dynamically to changes in its "environment". They applied this technique to a dataset of automobile insurance claims but the number of observations this data set contained was only 40 so any results were questionable. Much later, a new take on expert systems was published: Šubelj, Furlan, and Bajec (2011) proposed an expert system that made use of social network analysis. The objective of their work was to identify criminal networks of fraudsters

Around the same time, we saw some great strides in research which involved the automobile sector and used (at least partially) unsupervised machine learning methods. Brockett, Xia, and Richard A. Derrig (1998) proposed the use of Kohonen's self organizing feature maps for the detection of fraudulent claims in automobile insurance. Self-organizing maps are a (complex) method of clustering. Based on that method, they classified the claims by degree of suspicion. Through comparative experiments they show that their "technique performs better than both an insurance adjuster's fraud assessment and an insurance investigator's fraud assessment with respect to consistency and reliability".

Artís, Ayuso, and Guillén (1999) and Artís, Ayuso, and Guillén (2002) presented research on the automobile insurance sector. Their work was important because of a multitude of factors: they were the first to apply the methodology of discrete choice models for this application; they used data where the minority class (fraudulent claims) was oversampled and provided corrections for choice-

[9]Expert systems were at the forefront of Artificial Intelligence research during the 1980s but their usefulness and subsequently their popularity waned as time passed

Table 2: Significant Fraud Indicators according to E. B. Belhadji, George Dionne, and Tarkhani (2000)

	Fraud Indicators
1	No police report when there should have been one
2	A minor collision has led to excessive costs.
3	Existence of damage not related to the loss or inconsistent with the facts reported about the accident
4	The vehicle is reported stolen and found shortly after with heavy damage.
5	The vehicle is not attractive to thieves (i.e. ordinary old car)
6	Shortly before the loss, the insured checked the extent of coverage with his or her agent
7	The insured is having personal and business-related financial difficulties
8	The insured is extraordinarily familiar with the insurance and vehicle repair jargon
9	The insured (or claimant) is too eager or too frank to accept blame for the accident.
10	The accident (or loss) took place shortly after the vehicle was registered and insured or in the months preceding the end of the policy (or of coverage)
11	Numerous taxi receipts or bills for rental of vehicle from a body shop.
12	Bills or proofs of payment which seen phony or forged
13	Documentation of the estimate and repairs is not available
14	Contradictory witness reports concerning the circumstances of the loss
15	Accident involving a single vehicle
16	Accident involving an unidentified third party
17	Vehicle purchased with cash
18	Claimant is very aggressive (threatens to call a lawyer, contact the government, etc)
19	During the investigation, insured is nervous and seems confused

based sampling, in order to improve the performance on models that are heavily skewed, like in automobile insurance claims; finally, they estimated "the influence of the insured and claim characteristics on the probability of committing fraud" as well as investigating the problem of misclassification. Caudill, Ayuso, and Guillén (2005) further investigated the problem of misclassification.

Stijn Viaene, Richard A. Derrig, et al. (2002) conducted a thorough comparison of then state-of-the-art techniques for detection of fraud in automobile insurance claims. They tested logistic regression, C4.5 decision trees, k-nearest neighbor (kNN), Bayesian learning multilayer perceptron neural networks, least-squares support vector machine, naive Bayes and tree-augmented naive Bayes classification.

Brockett, Richard A. Derrig, et al. (2002) used principal component analysis of *RIDIT* scores, a method they call *PRIDIT* and which they evaluated in the context of automobile insurance fraud. Ai et al. (2013) continued the work on the PRIDIT method, by proposing a method to estimate the fraud rate in a data set of claims by using PRIDIT-based fraud rate estimation (PRIDIT-FRE).

S. Viaene, R.A. Derrig, and G. Dedene (2004) proposed the application of "the weight of evidence reformulation of AdaBoosted naive Bayes scoring". The claimed that this method effectively combined the advantages of boosting and the explanatory power of the weight of evidence scoring framework. Shortly after that, Viaene, Dedene, and Derrig (2005) they proposed Bayesian Learning Neural Networks which "explored the explicative capabilities of neural network classifiers with automatic relevance determination weight regularization".

A lot of the subsequent work on insurance fraud detection tried either to exploit the benefits of over- or under-sampling techniques or find other ways to deal with skewed data typical of automobile insurance claims. Pérez et al. (2005) used an oversampled automobile insurance data set in order to compare the performance of Consolidated Trees[10] versus the performance of the C4.5 tree algorithm. Bermúdez et al. (2008) presented a bayesian dichotomous logit model with asymmetric link which enabled it to deal with skewed data. Sundarkumar, Ravi, and Siddeshwar (2015) proposed undersampling the majority class based on one class support vector machine (OCSVM) models. Sundarkumar and Ravi (2015) went on to use the same technique (OCSVM) in conjunction with k-reverse nearest neighborhood models.

A number of similar approaches appeared in the following years: Hassan and Abraham (2016) proposed a different way of undersampling the majority class, Subudhi and Panigrahi (2017) proposed Genetic Algorithm-Based Fuzzy Cmeans Clustering (GAFCM), Subudhi and Panigrahi (2018) presented an adaptive oversampling method, Bouzgarne et al. (2020) used a Synthetic Minority Oversampling Technique (SMOTE) combined with a kNN algorithm, while Majhi et al. (2019) and Majhi (2021) tried hybrid techniques which at first applied fuzzy clustering in order to deal with the unbalanced data and then passed the modified data to Logit, Random Forest and XGBoost classifiers.

Some authors preferred a financial approach to the problem of insurance fraud and focused their research on finding classifiers that optimize cost savings instead of accuracy. Phua, Alahakoon, and Lee (2004) proposed a classifier that outperformed other widely used techniques in terms of cost saving. Stijn

[10] The authors describe Consolidate Trees as "classification trees induced from multiple subsamples but without loss of explaining capacity"

Viaene, Stijn Viaene, et al. (2007) investigated methods that minimized the cost of the investigation process instead of the error/misclassication rate. Bolance, Ayuso, and Guillen (2012) treated the problem from an operational risk point of view and calculated Value-at-Risk based loss estimations using non-parametric methods. Recently, Zelenkov (2019) proposed a method along the same lines but examined the case of *example-dependent cost-sensitive* (ECS) classification tasks[11] with the use of an AdaBoost classifier.

Another important branch of research on the subject involved treating automobile insurance fraud detection as an anomaly/outlier detection problem. Yan and Y. Li (2015) proposed an algorithm for determining whether an observation is an outlier by its distance to its nearest neighbor. Nian et al. (2016) introduced the Unsupervised Spectral Ranking for Anomaly (SRA) method. Shaeiri and Kazemitabar (2020) developed SRA further and provided algorithms which enable its use in real time on big data sets. Yan, Y. Li, et al. (2020) proposed an anomaly detection methodology that performs Kernel Ridge Regression with the assistance of a technique called an Artificial Bee Colony algorithm.

Anomaly detection methods have been particularly popular in the application of unsupervised machine learning for fraud detection. Besides the techniques mentioned in the previous paragraph, there have been a number of important applications of such methodologies in sectors other than automobile insurance. Stripling et al. (2018) used isolation forests -an effective and popular unsupervised anomaly detection method- for detecting worker's compensation fraud. Bauder, Da Rosa, and Khoshgoftaar (2018) compare the capabilities of different unsupervised learning methods in the context of health care insurance fraud. They apply Isolation Forests and Unsupervised Random Forests for the first time for detecting health care insurance fraud, while also using Local Outlier Factor (Breunig et al. 2000), autoencoders, and k-Nearest Neighbors. The Local Outlier Factor presents the best results in their data set. Jiang et al. (2021) also use a methodology based on Isolation Forests for detecting health care insurance fraud (specifically the problem of drug reselling)

Vosseler (2022) introduce a new outlier detection model, the Bayesian Histogram Anomaly Detector (BHAD). This model has desirable computational characteristics, as it scales linearly with the input data making it extremely fast compared to certain other methods when applied to big data sets. Their study also compares BHAD to other outlier detection algorithms,showing that it provides reliable results besides being computationally efficient.

Gomes, Jin, and Yang (2021) approach the problem of fraud detection across various industry sectors by focusing on identifying the most important variables using unsupervised deep learning methods, namely Auto Encoders (AE) and Variational Auto Encoders (VAE)

Returning to the automobile insurance sector, two interesting studies have been published very recently. Tumminello et al. (2023) approach fraud detection as a social phenomenon: they make use of bipartite networks to investigate the relationships between subjects and accidents or vehicles and accidents and then they develop filtering rules in order to uncover networks of criminal activity. They apply their methodology to a real database of Italian automobile insurance claims and validate the performance of their methodology when compared to

[11] i.e. classification tasks where the costs vary not only within classes but also between examples

out-of-sample fraudulent claims.

Duval, Boucher, and Pigeon (2023) explore the potential for new fraud detection methods in the new usage-based automobile insurance paradigm.The authors describe usage-based insurance (UBI) as "a fairly new insurance scheme mostly used in vehicle insurance, in which the insured's premium is estimated by making use of their driving data"[12]. They propose an anomaly detection method combined with a classification step in order to specify whether fraudulent activity took place. Their work presents the novel way of detecting anomalies based on both a "routine" and a "peculiarity" profile. The data sets available to the authors consist of telematics data from each insured vehicle. They separate the data into different trips and then they try to detect anomalous observations in two different ways: in the *local* scheme, each trip is compared to every other trip made by the same driver, which constitutes the trip's *routine* score. In the *global* scheme each trip is compared to every other trip made by all drivers, which accordingly corresponds to the trip's peculiarity score. They use three different ways of estimating anomaly scores: the Mahalanobis' distance, the Local Outlier Factor algorithm and Isolation Forests. In order to classify whether an observation is fraudulent or not they use Elastic-Net Regularized Logistic Regression on the anomaly scores.

4 Machine Learning Methods

4.1 Supervised versus Unsupervised Learning

Let us first describe in a simple manner the problem we are trying to tackle: we are given a data set $\mathcal{D}$, which contains various information about automobile insurance claims. The data set will be composed of a matrix X, each row corresponding to a claim and each column to an independent variable, a *feature* in the machine learning lingo, which may contain useful information related to the problem at hand. Finally, the dataset may contain a response variable Y but that is not always the case[13]. In machine learning these variables are usually called *labels*. When they exist, then our data set is constituted of N pairs of observations $(x_1^\intercal, y_1), (x_2^\intercal, y_2), \ldots, (x_N^\intercal, y_N)$.

In the context of automobile insurance fraud these response variables or *labels* may take a number of forms: for example a binary variable called "fraud" which indicates whether the claim was fraudulent. Alternatively, there may be a binary variable indicating whether the claim was deemed suspicious or fraudulent. Another variation of the same concept would be a variable "fraud" which takes three values, each corresponding to three classes of claims: non-fraudulent, suspicious and fraudulent.

Our ultimate goal is to utilize the data set in such a way, that we create statistical models and algorithms that enable us to detect fraudulent (or simply suspicious) claims when given new data points/sets. This is, at its core, a classification task. Our models seek to find the class to which a new data point

[12]The driving data could be recorded by a specialized on-board diagnostics device, but, nowadays, simply using a smart phone is preferred because of its cost efficiency.

[13]One could argue that the data set should at least contain a variable that indicates whether a claim was legally proven to be fraudulent, since that information is available to the insurer. That is not true however when dealing with live or recent data.

belongs, either the class of legitimate claims or the class of fraudulent ones. For this purpose one may typically use supervised machine learning, but the domain unsupervised learning is also promising.

The main difference between supervised and unsupervised models is usually reduced to whether they require and use labels (supervised) or they do not (unsupervised). A more rigorous description for each would be the following: in supervised learning we have access to the random variables X and Y. Supposing that they have some joint probability density $Pr(X, Y)$ then supervised learning can be treated as a density estimation problem focusing on the conditional density $Pr(Y|X)$. A model is trained by taking the predictions $\hat{y}_i$ it makes for each $x_i^\intercal$ given, and finding the one which minimizes some *loss function*[14] $L(y, \hat{y})$. We know from Bayes theorem that:

$$Pr(Y|X) = Pr(X, Y)/Pr(X)$$

In supervised learning, $Pr(X)$, i.e. the marginal density of only the X values is "typically of no concern" (Hastie, Tibshirani, and Friedman 2017). However in unsupervised learning X and the joint density of each of its constituent row vectors, $Pr(X)$, is all we have: hence it becomes our main concern. Hastie, Tibshirani, and Friedman (2017) make some general observations: "the dimension of X is sometimes much higher than in supervised learning, and the properties of interest are often more complicated than simple location estimates." This last part is particularly true in our application. The unsupervised machine learning methods we will use will focus on finding abnormalities in the data. What constitutes an abnormality will depend on each method. For example, in a clustering setting, anomalies in the data would be data points that lie a great distance from the center of the cluster to which they were assigned. In our application, legitimate insurance claims should form a relatively compact cluster, with fraudulent claims lying quite far away from that cluster. To achieve our goal, however, we cannot simply employ unsupervised machine learning techniques. We must make use of supplementary models to turn our measure of data abnormality into a decision on whether the data point (claim) under examination is fraudulent or not.

In various applications in data science, unsupervised machine learning is employed simply because the data is not labeled and consequently it is the only option. However, the data we use in this thesis as well as the data typically used in real world applications in the automobile insurance sector *does* have labels. One may then reasonably wonder why we would use unsupervised machine learning in the first place. Supervised machine learning is less subjective (Hastie, Tibshirani, and Friedman 2017) and is perfectly suited to modeling relations between data (Gomes, Jin, and Yang 2021); under ideal conditions supervised machine learning models lead to robust estimates (Debener, Heinke, and Kriebel 2023). The literature also shows that supervised methods have been researched much more extensively than their counterparts in the context of insurance fraud (J. Li et al. 2008; Benedek, Ciumas, and Nagy 2022; Debener, Heinke, and Kriebel 2023)

The answer to the question posed above is quite simple: due to the nature of our research domain and of the data sets that we encounter in this domain.

[14]For example, one of the most common loss functions is the Mean Squared Error (MSE) $L(y, \hat{y}) = (y - \hat{y})^2$

Here we will repeat some of the characteristics of the automobile insurance sector and the data sets concerning claim fraud, but now we will be able to see how they hinder the use of supervised machine learning methods. First of all, the labels available in our data set may only contain very few detected fraud cases[15] which could potentially hinder robust model estimation (Debener, Heinke, and Kriebel 2023) and will almost certainly result to models that can detect only a very small portion of total fraud. Even in the case where the labels also include suspected fraud and a supervised model can be trained on a lot of data labeled as fraudulent or potentially fraudulent, problems persist: the creation of these labels does not happen by itself as part of the business cycle (Gomes, Jin, and Yang 2021), instead being a manual cost- and time-intensive process (Gomes, Jin, and Yang 2021; Stijn Viaene, Stijn Viaene, et al. 2007). Unsupervised models present the obvious advantage of not requiring such a process while also providing another benefit. Since these labels are based on subjective judgments by a company's staff, they have implicit biases within them, while also not being perfectly accurate nor complete[16]. A well trained supervised model will by its very nature propagate these biases onto its predictions, only being able to detect cases similar to those encountered before, while also missing any kind of fraudulent activity that the companies staff could not identify(Debener, Heinke, and Kriebel 2023). In contrast, an unsupervised machine learning model in the absence of subjectively labeled data could avoid any bias or misguidance precipitated by them (Gomes, Jin, and Yang 2021), while also being able to detect fraudulent activity that differs from what has been detected before(Debener, Heinke, and Kriebel 2023).

In conclusion, there is no ideal machine learning method for our application. A lot of the benefits of unsupervised models were mentioned in the previous paragraph, but they also have their shortcomings. Generally supervised methods provide results that are much easier to interpret and/or evaluate, and when the labels fed to them are reasonably accurate they are extremely efficient at their predictions. Moreover, they do not require careful selection of the explanatory variables/features which are supplied to them since they can usually disregard any information that is not petrinent to the problem at hand. Unsupervised models on the other hand can discover patterns in the data that are not at all useful for our purposes (Debener, Heinke, and Kriebel 2023). Therefore the selection of variables to include in an unsupervised model should be done with care and requires domain knowledge (Stripling et al. 2018).

4.2 Overview of Anomaly Detection Techniques

In our literature overview we saw that the application of unsupervised machine learning methods in the domain of automobile insurance fraud is limited to a small number of techniques, such as Isolation Forests, Local Outlier Factor and Autoencoders. In the more general context of the anomaly detection literature, one is able to find a wealth of different techniques, whose capabilities in detecting automobile insurance fraud have not yet been evaluated.

Our empirical application will be a simulation study of the performance of various anomaly detection techniques, many of which have not yet appeared

[15] Typically those proven as such in court

[16] It is inevitable that these labels will contain both false positives and false negatives as discussed in a previous chapter.

Table 3: Table of Anomaly Detection Methods Utilized In This Study

Abbreviation	Name	Family
FastABOD	Fast Angle-Based Outlier Detection using approximation	Probabilistic
ECOD	ECDF-based Outlier Detection	Probabilistic
COPOD	Copula-Based Outlier Detection	Probabilistic
SOS	Stochastic Outlier Selection	Probabilistic
QMCD	Quasi-Monte Carlo Discrepancy Outlier Detection	Probabilistic
KDE	Kernel Density Functions based Outlier Detection	Probabilistic
Sampling	Rapid distance-based outlier detection via sampling	Probabilistic
GMM	Probabilistic Mixture Modeling based Outlier Detection	Probabilistic
PCA	Principal Component Analysis based Outlier Detection	Linear Model
MCD	Minimum Covariance Determinant based Outlier Detection	Linear Model
CD	Cook's distance based Outlier Detection	Linear Model
OCSVM	One-Class Support Vector Machines based Outlier Detection	Linear Model
LOF10	Local Outlier Factor (10 neighbours)	Proximity-Based
LOF20	Local Outlier Factor (20 neighbours)	Proximity-Based
LOF100	Local Outlier Factor (100 neighbours)	Proximity-Based
COF	Connectivity-Based Outlier Factor	Proximity-Based
CBLOF	Clustering-Based Local Outlier Factor	Proximity-Based
HBOS	Histogram-based Outlier Score	Proximity-Based
kNN	kNN based Outlier Detection (max distance)	Proximity-Based
kNN-avg	kNN based Outlier Detection (avg. distance)	Proximity-Based
kNN-median	kNN based Outlier Detection (median distance)	Proximity-Based
IForest	Isolation Forest	Ensembles
INNE	Isolation-based Anomaly Detection Using Nearest-Neighbor Ensembles	Ensembles
DIF	Deep Isolation Forest for Anomaly Detection	Ensembles / Neural Networks
FB	Feature Bagging	Ensembles
LODA	Lightweight On-line Detector of Anomalies	Ensembles
LUNAR	Unifying Local Outlier Detection Methods via Graph Neural Networks	Graph-Based/Neural Networks
Beta-VAE	Variational AutoEncoder	Neural Networks

in the automobile insurance fraud literature. In order to achieve this, we will be leveraging PyOD (Zhao, Nasrullah, and Z. Li 2019), a python library which includes implementations of more than 30 different anomaly detection methods.

Due to large number of different anomaly detection methods we will be utilizing, a description of each one is beyond the scope of this work. Instead we will be providing a more general description of the common elements that all anomaly detection techniques share, as well as description for “families” of such techniques, i.e. groupings of techniques that work in a similar way. We will also incorporate a table that enumerates the techniques we will be using in our simulation study including a short descriptive name, which we will be using in order to present the results of our simulations (see Table 3). The interested reader is encouraged to look at the PyOD library as well as at the related work ADBench (Anomaly Detection Benchmark) (Han et al. 2022; Zhao, Nasrullah, and Z. Li 2019) for more details than what we provide here on the different anomaly detection methodologies.

We can identify the following shared commonalities between all anomaly detection methods: given a design matrix X containing our p explanatory variables they all have a way to compute a mapping $f : \mathbb{R}^p \to \mathbb{R}$. This mapping f maps every row $X_{j,\cdot}$ of our X matrix (i.e. every observation in our data set) to a real number which in the anomaly/outlier detection literature is referred to as an outlier score. The last step of these algorithms is to compute a mapping (given the outlier scores for each observation) $g : \mathbb{R} \to \mathrm{I} = \{0, 1\}$ between the real numbers representing the outlier scores obtained in the previous step to a label taking values in $\{0, 1\}$[17]. The composition $g \circ f$ of these two mappings results in the following mapping: $g \circ f : \mathbb{R}^p \to \mathrm{I} = \{0, 1\}$, which essentially describes the whole procedure: For each observation in our data (each row j of X) we assign an outlier score to it and based on the outlier scores we, finally, assign a label indicating whether an observation was deemed to be an anomaly.

Let us delve deeper into how these mappings work by working backwards and initially explaining how the mapping between outlier scores and labels is achieved. If we know the exact ratio of anomalies to the number of total observations (typical and anomalies), which in the anomaly detection literature is commonly referred to as *comtamination* or *contamination ratio*, this mapping is quite straightforward. Given a contamination ratio, we simply choose a percentage of our observations with the most extreme anomaly scores and mark them as anomalies (set their label to 1); this percentage is of course equal to the contamination ratio. Although this method is extremely convenient due to its simplicity, in most cases we cannot a priori know the correct value for the contamination parameter. Domain expertise may be able to provide us with a rough estimate for the parameter, but in this way we will be over- or underestimating the correct ratio of anomalies in each data set we are given. It is important to note that in the anomaly detection literature a breadth of more sophisticated approaches for automatically determining the contamination factor exist. These approach usually rely on statistical measures that describe aspects of the distribution of outlier scores (see for example (Perini, Buerkner, and Klami 2023)). Such approaches are definitely worthy of more attention and

[17]There are alternative ways to encode whether an observation is an outlier, (for example scikit-learn uses the 1 label for typical observations and -1 for outliers). The encoding we use here, (which is also what PyOD uses) encodes typical observations with 0 and anomalies with 1

research, but they are outside of the scope of this particular study; as such we will be using the more naive approach where the contamination parameter is pre-specified. As will be explained later, in our simulation study we took steps to mitigate some of the unrealistic aspects of this naive approach.

Concluding, we will give a small glimpse into how the mapping from observations to outliers is achieved. In regards to this aspect, every method is different, but there exist some overarching general patterns of operation of different classes of anomaly detection techniques. A significant portion of the methods we will use are proximity-based; the calculation of distances between neighboring data points are the core concept behind their operation. Another portion are probabilistic: the use of various statistical concepts and techniques are the backbone of their operation. Ensemble-based outlier detection methods are also present, including the widely used Isolation Forest. Other techniques may rely on linear models or neural networks

5 Simulation Study

5.1 Reasoning behind our choice of a simulation study

At this point we encounter one of the problems we mentioned earlier concerning data availability: the data sets concerning automobile insurance claims are in most cases confidential. This can be attributed to two factors: first of all this data contains sensitive personal information for the insurance company's customers; moreover this information has inherent value to the company and its dissemination to competitors could have detrimental economic effects. As such, it is difficult for researches to get access to that information. Most of the research on the subject relies on proprietary information that is not made available to the public (Benedek, Ciumas, and Nagy 2022). The most insight we get into the data set studies use is at best some descriptive statistics in most cases (e.g. number of observations, type of features, etc). It is also not uncommon to encounter data sets where a anonymization preprocessing step (for example the use of a PCA transformation) has altered or removed the natural interpretation of each feature (see for example Palacio (2019), which is a study on property insurance fraud which uses such a dataset)

Due to the aforementioned challenges in acquiring a data set that is suitable for our application, we chose instead to focus on a simulation study. Our approach is guided in large part by a number of different techniques that are proposed in the unsupervised anomaly detection literature. (Han et al. 2022; Steinbuss and Böhm 2021). The central concept characterizing the approaches described in the aforementioned papers is the creation of "realistic" synthetic data[18], by utilizing a real data set as a "seed" for the creation of synthetic data sets that are similar to it, which are in turn "contaminated" with synthetic anomalies. In turn, we apply these techniques in the domain of automobile insurance fraud.

[18]this process is also referred to as "synthetic reconstruction"

Table 4: Distribution of categorical variable levels

Incident Severity	Incident Type	Police Report	Frequencies
Major Damage	Collision	NO	0.196
		YES	0.080
Minor Damage	Collision	NO	0.177
		YES	0.089
	Parked Car	NO	0.063
		YES	0.021
	Vehicle Theft	NO	0.065
		YES	0.029
Total Loss	Collision	NO	0.185
		YES	0.095

5.2 Simulation procedure for typical observations

Given a real data set one may generate a synthetic one that is similar to it using a number of different approaches. For example, during the last decade specific classes of artificial neural networks such as Generative Adversarial Networks have been proposed and used for this application. In our simulation study we will adopt simpler parametric techniques that may not be able to recreate the original data set with the fidelity of a neural network based approach, but will instead enable us to add different kinds of anomalies to the data set besides the typical observations. The different kinds of anomalies can be finely tuned based on the parametric description of our data set in order to allow us to evaluate the efficacy of a wide number of anomaly detection techniques.

Our original automobile insurance claims data set is composed of three categorical variables and 16 continuous variables. The categorical variables represent the type of incident (taking the values {Collision, Vehicle Theft and Parked Car}), the incident severity ({Minor Damage, Major Damage, Total Loss}) and whether a Police Report was available. We present the relative frequencies for each distinct combination of categorical variable levels in Table 4. The continuous variables are the following: [months_as_customer, age, policy_deductible, policy_annual_premium, umbrella_limit, capital_gains, capital_loss, incident_hour_of_the_day, number_of_vehicles_involved, bodily_injuries, witnesses, total_claim_amount, injury_claim, property_claim, vehicle_claim, auto_year]. We present the mean and standard deviation for the numerical columns of the whole dataset in Table 5. We also present a heatmap of the correlation matrix of the numerical columns in Figure 1

The first logical step is to obtain a parametric representation of our original data set. We identify the frequency of appearance of each combination of the different levels of our original data. We model the probability of each combination appearing by a multinomial distribution. The next step is modeling the distribution of the continuous variables. This is achieved via a mixture of multivariate gaussian distributions. For each combination of the categorical variables we isolate the observations of the continuous variables and fit a multivariate normal distribution to these observations. Consequently, we obtain as many multivariate normal distributions for the continuous variables as we have combinations of the levels of the categorical variables. The exact steps are presented

Table 5: Mean and Standard Deviation of Numerical Columns (Whole Dataset)

	Mean	Standard Deviation
age	38.948	9.140
auto_year	2005.103	6.016
bodily_injuries	0.992	0.820
capital_gains	25126.100	27872.188
capital_loss	-26793.700	28104.097
incident_hour_of_the_day	11.644	6.951
injury_claim	7433.420	4880.952
months_as_customer	203.954	115.113
number_of_vehicles_involved	1.839	1.019
policy_annual_premium	1256.406	244.167
policy_deductible	1136.000	611.865
property_claim	7399.570	4824.726
total_claim_amount	52761.940	26401.533
umbrella_limit	1101000.000	2297406.598
vehicle_claim	37928.950	18886.253
witnesses	1.487	1.111

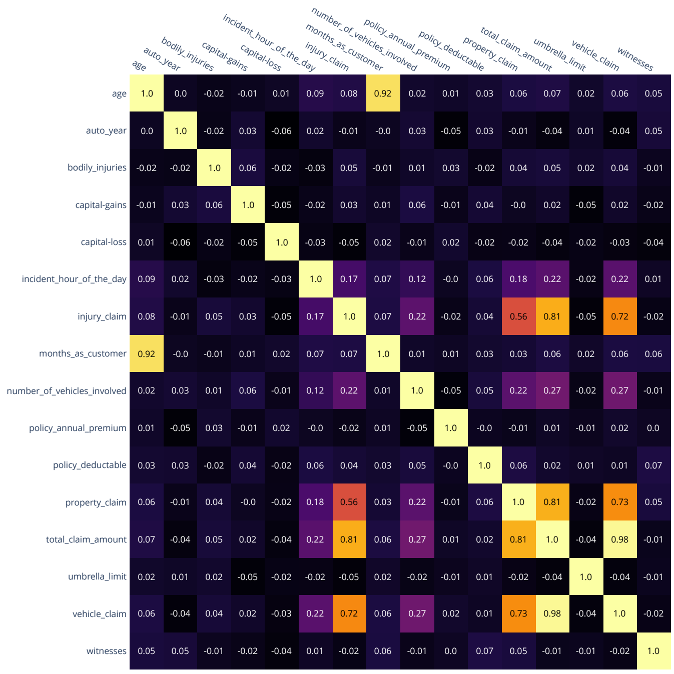

Figure 1: Correlation Matrix of the numerical features (Whole Dataset)

in algorithm 1.

Input : $\boldsymbol{X}$: $n \times p$ matrix of independent variables in original dataset
Output: $\boldsymbol{\pi}$: $\mathbb{R}^m$ vector of probabilities for multinomial distribution (with $n = 1$) describing the probabilities of each combination of categorical variables occuring
normal_dist_spec $:= \{(\boldsymbol{\mu_1}, \Sigma_1), \ldots, (\boldsymbol{\mu_m}, \Sigma_m)\}$ m pairs of mean vectors and covariance matrices describing the different multivariate normal distributions for the continuous variables

initialize π, normal_dist_spec;
(optional) remove outliers from X;
$m \leftarrow$ number of distinct combinations of the levels of the categorical variables in X;
for $i \leftarrow 1$ **to** m **do**
  select the i-th distinct combination of categorical variable levels in X;
  $\pi_i \leftarrow$ frequency of occurrence of current combination;
  isolate the observations in X corresponding to the current combination;
  fit a multivariate normal distribution to the continuous features isolated in the previous step;
  normal_dist_spec[i] $\leftarrow (\boldsymbol{\mu}, \Sigma)$ describing the distribution fitted in the previous step;
end

Algorithm 1: Distribution Fitting Procedure

The steps mentioned above allow us to generate synthetic samples of typical/non-fraudulent observations. We present the algorithm used in algorithm 2. By sampling from the multinomial distribution we obtained in the fitting step, we randomly pick a combination of categorical variable levels. We place the categorical variables in the first three columns of the X matrix of simulated samples, with each taking a value determined by the combination chosen. We finally sample from the multivariate normal distribution that is paired with the combination of categorical variable levels in order to generate the continuous features in X. Finally, we employ a post-processing steps to ensure that our variables take “sane” values. A normal distribution is a convenient but inaccurate description of many of our variables. As such, we have to post-process the random values generated: for ordinal values we discretize them; we also apply bounds where appropriate (for example claims should take only positive values)

5.3 Simulation procedure for anomalies

The literature on unsupervised anomaly detection proposes different parametric procedures for the generation of different kinds of anomalies[19]. In Steinbuss and Böhm (2021) the authors propose the generation of *Local*, *Global* and *Dependency* anomalies. Han et al. (2022) propose the use of *Clustered* Anomalies

[19]The terms “anomaly” and “outlier” are used interchangeably

Input : $\boldsymbol{\pi}$: $\mathbb{R}^m$ vector of probabilities for multinomial distribution (with $n = 1$) describing the probabilities of each combination of categorical variables occuring
normal_dist_spec $:= \{(\boldsymbol{\mu_1}, \Sigma_1), \ldots, (\boldsymbol{\mu_m}, \Sigma_m)\}$ m pairs of mean vectors and covariance matrices describing the different multivariate normal distributions for the continuous variables
p: number of features (categorical and continuous) in the original data set
p_categ: number of categorical features in the original data set
n: number of simulated samples to generate

Output: $\boldsymbol{X}$: $n \times p$ matrix of simulated samples
$\boldsymbol{y}$: $\mathbb{R}^n$ vector of simulated dependent variable

initialize X, y;
for $i \leftarrow 1$ **to** n **do**
  $k \leftarrow$ sample from a multinomial distribution with probabilities $\boldsymbol{\pi}$, $n = 1$ and outcomes in $\{1, \ldots, m\}$;
  $X_{i,1}, \ldots, X_{i,p\_categ} \leftarrow$ levels corresponding to the k-th combination of categorical variable levels;
  $(\boldsymbol{\mu}, \Sigma) \leftarrow$ normal_dist_spec[k];
  $X_{i,p\_categ+1}, \ldots, X_{i,p} \leftarrow$ random sample from a multivariate normal characterized by the unmodified $(\boldsymbol{\mu}, \Sigma)$ pair;
  $y_i \leftarrow 0$
end

Algorithm 2: Generation Of Typical Samples in addition to the other types of anomalies. A general comment regarding all anomaly generation procedures is that we tried to tune their parameters so that most of our anomaly detection algorithms achieved scores somewhere in the interval $(0.5, 1)$.

5.3.1 Global Anomalies

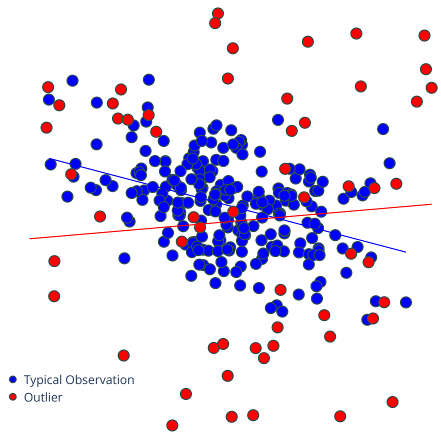

Figure 2: Bivariate Demonstration of Global Anomalies

Global anomalies are data points which differ from the rest of the data set. The proposed method for their generation found in the literature is to sample for each X_j continuous feature in the data set from a uniform distribution in the interval $[min(X_j), max(X_j)]$. A somewhat wider interval may also be used.

In our case this method results in anomalies that are extremely easy to differentiate from the typical observations. Consequently a lot of the outlier detection methods produce excellent results, resulting in limited or no information on the relative performance of the methods. Our workaround is to sample only a (random) subset of the continuous features from a uniform distribution, while the rest of the features are generated in the usual manner from our mixture of multivariate normal distributions. We present our version generating global outliers in algorithm 3. In our simulations the size of the columns subset used was not constant and its size took random values in $\{1, \ldots, \lfloor c/4 \rfloor\}$ where c is the number of columns with numerical variables

5.3.2 Clustered Anomalies

Clustered anomalies could be considered a sub-case of global anomalies (i.e. anomalies that differ from all of the rest of the data) but with an important difference: these anomalies are bunched up together. Thus they could pose a problem to methods that are proximity based.

Input : **normal_dist_spec** $:= \{(\boldsymbol{\mu_1}, \Sigma_1), \ldots, (\boldsymbol{\mu_m}, \Sigma_m)\}$ // m pairs of mean vectors and covariance matrices describing the different multivariate normal distributions for the continuous variables
p: number of features (categorical and continuous) in the original data set
p_categ: number of categorical features in the original data set
n: number of simulated samples to generate
$\boldsymbol{\alpha}$: bounds scaling factor
col_subset_size: size of the subset of continuous variables
Output: $\boldsymbol{X}$: $n \times p$ matrix of simulated samples
$\boldsymbol{y}$: $\mathbb{R}^n$ vector of simulated dependent variable

```
initialize X, y;
for i ← 1 to n do
    k ← sample from a multinomial distribution with equal
      probabilities 1/m, n = 1 and outcomes in {1, ... m};
    X_{i,1}, ..., X_{i,p_categ} ← levels corresponding to the k-th
      combination of categorical variable levels;
    (μ, Σ) ← normal_dist_spec[k];
    X_{i,p_categ+1}, ..., X_{i,p} ← random sample from a multivariate
      normal characterized by the unmodified (μ, Σ) pair (comment:
      some of the values are placeholders and will be replaced in the
      loop that follows);
    column_subset ← random subset of the columns of X with size
      col_subset_size;
    for j ← p_categ + 1 to p do
        if j in column_subset then
            bounds ← {min(X_original_{.,j}), max(X_original_{.,j})};
            scale the bounds pair so that the interval is α-times wider;
            X_{i,j} ← sample from uniform distribution in the interval
              specified by the bounds pair;
        end
    end
    y_i ← 1
end
```

Algorithm 3: Generation Of Global Anomalies

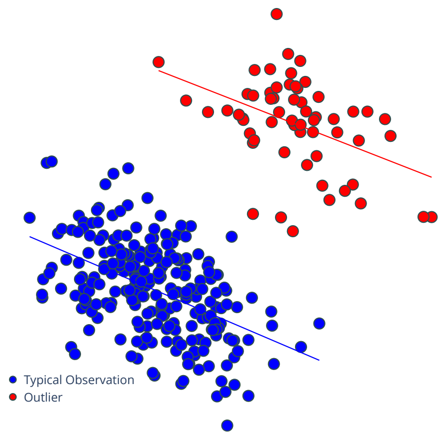

Figure 3: Bivariate Demonstration of Clustered Anomalies

Their generation is accomplished by adding to the vector of means that describes our multivariate normal distribution a constant factor α times each feature's standard deviation. The value we used for α in our simulations was 1. The detailed steps for clustered anomaly generation are described in algorithm 4

5.3.3 Local Anomalies

Local anomalies are data points which differ from their local neighborhood (Breunig et al. 2000). We create them by scaling the covariance matrices Σ that describe our multivariate normal distributions by a constant factor α and we then simulate data points from the scaled distribution. In our simulations we used $\alpha = 1.8$

5.3.4 Dependency Anomalies

Dependency Anomalies are data points which do not follow the dependence structure that characterizes normal data points. In the literature the proposed method for generating such anomalies is to model the dependency structure of typical samples using Vine Copulas; Kernel Density Estimation (KDE) is used in order to model the distributions of the variables (Steinbuss and Böhm 2021).

In this case we choose to stray away from the techniques we encounter in the literature. We already have ways of describing the dependency between our variables: namely, our continuous variables are characterized by a mixture of multivariate normal distributions, so the Σ_i can describe the correlation between the variables; moreover combination of categorical variable levels corresponds to a different multivariate normal distribution. We make use of these facts in order to generate a new type of outliers. First of all, we modify the Σ_i matrices describing our normal distribution by keeping only the elements on the diagonal. In this way we negate any correlation between variables. We also change the

Input : $\boldsymbol{\pi}$: $\mathbb{R}^m$ vector of probabilities for multinomial distribution (with $n = 1$) describing the probabilities of each combination of categorical variables occuring
normal_dist_spec $:= \{(\boldsymbol{\mu_1}, \Sigma_1), \ldots, (\boldsymbol{\mu_m}, \Sigma_m)\}$ m pairs of mean vectors and covariance matrices describing the different multivariate normal distributions for the continuous variables
p: number of features (categorical and continuous) in the original data set
p_categ: number of categorical features in the original data set
n: number of simulated samples to generate
$\boldsymbol{\alpha}$: location translation factor
Output: $\boldsymbol{X}$: $n \times p$ matrix of simulated samples
$\boldsymbol{y}$: $\mathbb{R}^n$ vector of simulated dependent variable

initialize X, y;
for $i \leftarrow 1$ **to** n **do**
  $k \leftarrow$ sample from a multinomial distribution with probabilities $\boldsymbol{\pi}$, $n = 1$ and outcomes in $\{1, \ldots, m\}$;
  $X_{i,1}, \ldots, X_{i,p\_categ} \leftarrow$ levels corresponding to the k-th combination of categorical variable levels;
  $(\boldsymbol{\mu}, \Sigma) \leftarrow$ normal_dist_spec[k];
  $n\_continuous\_variables \leftarrow$ number of elements in $\boldsymbol{\mu}$;
  for $j \leftarrow 1$ **to** $n\_continuous\_variables$ **do**
    $\mu_j \leftarrow \mu_j + \alpha \Sigma_{j,j}^{1/2}$;
  end
  $X_{i,p\_categ+1}, \ldots, X_{i,p} \leftarrow$ random sample from a multivariate normal characterized by the $(\boldsymbol{\mu}, \Sigma)$ modified pair;
  $y_i \leftarrow 1$
end

Algorithm 4: Generation Of Clustered Anomalies

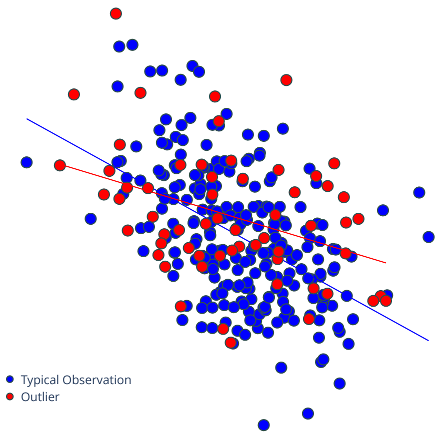

Figure 4: Bivariate Demonstration of Local Anomalies

Input : $\boldsymbol{\pi}$: $\mathbb{R}^m$ vector of probabilities for multinomial distribution (with $n = 1$) describing the probabilities of each combination of categorical variables occuring
normal_dist_spec $:= \{(\boldsymbol{\mu_1}, \Sigma_1), \ldots, (\boldsymbol{\mu_m}, \Sigma_m)\}$ m pairs of mean vectors and covariance matrices describing the different multivariate normal distributions for the continuous variables
p: number of features (categorical and continuous) in the original data set
p_categ: number of categorical features in the original data set
n: number of simulated samples to generate
$\boldsymbol{\alpha}$: covariance scaling factor

Output: $\boldsymbol{X}$: $n \times p$ matrix of simulated samples
$\boldsymbol{y}$: $\mathbb{R}^n$ vector of simulated dependent variable

initialize X, y;
for $i \leftarrow 1$ **to** n **do**
  $k \leftarrow$ sample from a multinomial distribution with probabilities $\boldsymbol{\pi}$, $n = 1$ and outcomes in $\{1, \ldots, m\}$;
  $X_{i,1}, \ldots, X_{i,p\_categ} \leftarrow$ levels corresponding to the k-th combination of categorical variable levels;
  $(\boldsymbol{\mu}, \Sigma) \leftarrow$ normal_dist_spec[k];
  $\Sigma \leftarrow \alpha\Sigma$;
  $X_{i,p\_categ+1}, \ldots, X_{i,p} \leftarrow$ random sample from a multivariate normal characterized by the $(\boldsymbol{\mu}, \Sigma)$ modified pair;
  $y_i \leftarrow 1$
end

Algorithm 5: Generation Of Local Anomalies

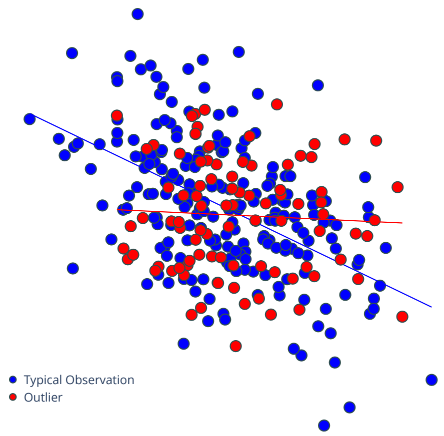

Figure 5: Bivariate Demonstration of Dependency Anomalies

probabilities of each combination of our categorical variables appearing to make them equally probable. We present our procedure in algorithm 6

5.3.5 Mixed Anomalies

The final type of anomalies generated are simply a mixture of the previous methods. This mix of anomaly generation procedures could simulate the existence of different types of anomalies in a dataset. It could also show which models tend to perform better under such circumstances.

5.4 Contamination present in the data

Another consideration for the purposes of our simulations is the prevalence of anomalies in data set, or *contamination* as anomaly detection literature commonly describes it, a subject we mentioned in an earlier part of this work. The anomaly detection methods we use take as a parameter the contamination in our data set. Since this ratio is something we control in our simulations, we could *naively* provide the true value for this parameter to the anomaly detection methods. This is, however, highly unrealistic. Even if we could rely on domain expertise to set the anomaly ratio to what is usually encountered in such datasets, knowing the *exact* ratio for each data set is, obviously, impossible

As such, we choose to utilize the estimates of (Insurance Information Institute 2023; Benedek, Ciumas, and Nagy 2022) for the prevalence of fraud in Central and Eastern Europe which places it somewhere in the range of $10\% - 20\%$. We set the contamination parameter required for the various models as input to 15%, and for each simulation we simulate a variable ratio of anomalies. It is obvious that in most cases, this results in models that over- or under-estimate the prevalence of anomalies in data. For generating the contamination of each simulated data set we sample from a uniform distribution in the interval $[0.05, 0.25]$.

Input : $\boldsymbol{X\_original}$: $n \times p$ matrix of independent variables in original dataset
normal_dist_spec := $\{(\boldsymbol{\mu_1}, \Sigma_1), \ldots, (\boldsymbol{\mu_m}, \Sigma_m)\}$ m pairs of mean vectors and covariance matrices describing the different multivariate normal distributions for the continuous variables
p: number of features (categorical and continuous) in the original data set
p_categ: number of categorical features in the original data set
n: number of simulated samples to generate
Output: $\boldsymbol{X}$: $n \times p$ matrix of simulated samples
$\boldsymbol{y}$: $\mathbb{R}^n$ vector of simulated dependent variable

initialize X, y;
for $i \leftarrow 1$ **to** n **do**
  $k \leftarrow$ sample from a multinomial distribution with equal probabilities $1/m$, $n = 1$ and outcomes in $\{1, \ldots m\}$;
  $X_{i,1}, \ldots, X_{i,p\_categ} \leftarrow$ levels corresponding to the k-th combination of categorical variable levels;
  $(\boldsymbol{\mu}, \Sigma) \leftarrow$ normal_dist_spec[k];
  set all non-diagonal elements of Σ to 0;
  $X_{i,p\_categ+1}, \ldots, X_{i,p} \leftarrow$ random sample from a multivariate normal characterized by the $(\boldsymbol{\mu}, \Sigma)$ modified pair;
  $y_i \leftarrow 1$
end

Algorithm 6: Generation Of Dependency Anomalies

Input : **p**: number of features (categorical and continuous) in the original data set
n: number of simulated samples to generate
Output: $\boldsymbol{X}$: $n \times p$ matrix of simulated samples
$\boldsymbol{y}$: $\mathbb{R}^n$ vector of simulated dependent variable

initialize X, y;
for $i \leftarrow 1$ **to** n **do**
  choose randomly and with equal probability an element in $\{clustered, local, global, dependency\}$;
  $X_{i,1}, \ldots, X_{i,p} \leftarrow$ random sample generated by the method chosen in the previous step;
  $y_i \leftarrow 1$
end

Algorithm 7: Generation Of Mixed Anomalies

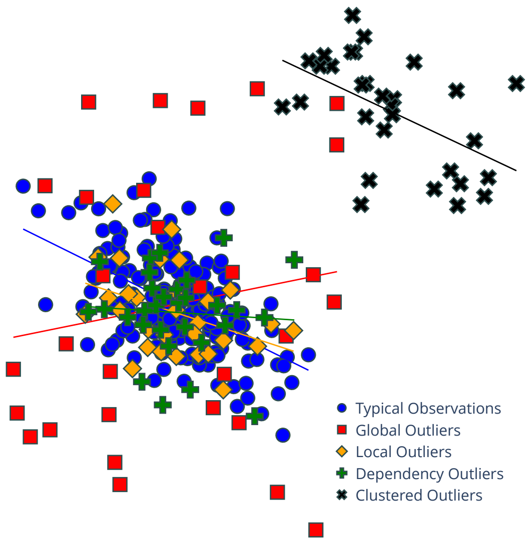

Figure 6: Bivariate Demonstration of Mixed Anomalies

An alternative approach could be to sample from a normal distribution with 0.15 mean and with a value for standard deviation which would ensure that almost all sampled values lie in the interval $[0.05, 0.25]$[20]. The possible advantage of using a normal distribution instead of a uniform would be that generated values would lie closer to the mean of the distribution, which may reflect reality better.

5.5 Presentation of results

In this section we will present the results of our simulation study. Our results are based on 10000 repetitions for each anomaly type. For every type of anomaly we will present two tables, one including the mean and 95% confidence intervals for the Area Under The Receiver Operating Characteristic Curve (ROC AUC) and another table with the mean and confidence intervals for Precision, Recall and F_1. In our evaluation of the results in the tables for Precision, Recall and F_1, we will for the most part comment on the F_1 score since it is the harmonic mean of Precision and Recall.

It is important to clarify here that due to the way of collecting our results, the ROC AUC results have been computed by providing more information to the programs compared to the results for Precision, Recall and F_1. The latter have been computed by comparing the true labels of our observations to the labels that were predicted by our anomaly detection algorithms. In contrast, in

[20]Of course, we would have to truncate the distribution, since, no matter how unlikely, we should not be able to generate values for the contamination that are not in the interval $(0, 0.5)$

order to compute the ROC AUC values, besides the true labels, the program did not have access to the predicted labels but instead to the probabilities of each observation being predicted as an anomaly, which allows the ROC AUC values to be more optimistic about the performance of each different method compared to when using only the predicted labels and not the probabilities that lead to label prediction.

5.5.1 Global Anomalies

When evaluating the results in the presence of global anomalies, as they are presented in Table 6, we are immediately surprised by the exceptional performance of MCD, which appears to have the ROC AUC of a perfect classifier when used on our simulated data. Other particularly well performing techniques are all the variants of kNN-Based Outlier Detection (kNN, kNN-avg, kNN-median), Unifying Local Outlier Detection Methods via Graph Neural Networks (LUNAR), Kernel Density Functions Based Outlier Detection (KDE) and Probabilistic Mixture Modeling based Outlier Detection (GMM). Among the worst performers are Histogram-Based Outlier Score (HBOS) and DIF

Viewing the F_1 scores we include in Table 7 the amazing performance of MCD is reaffirmed. The same can be said about the rest of the models that performed well in regards to their ROC AUC. As was also the case with local outliers, we once more see that despite its good ROC AUC score kNN-avg performs in an unsatisfactory way. In contrast, HBOS, which also performed badly in regards to its ROC AUC score, is not too far behind its better performing counterparts in its F_1 score.

5.5.2 Clustered Anomalies

Looking at Table 8 we immediately notice that Deep Isolation Forest (DIF) performs much worse than chance. Stochastic Outlier Selection (SOS), Connectivity Based Outlier Factor (CBLOF), Local Outlier Factor with 10 or 20 neighbours (LOF10, LOF20), and Angle Based Outlier Detection (FastABOD) perform worse than most of the methods. Among the highest performers are Copula-Based Outlier Detection (COPOD) and Minimum Covariance Based Outlier Detection (MCD). It is important to note that, while COPOD performs slightly worse than MCD, when looking at their confidence intervals we notice that COPOD is much more consistent.

The second table of results, Table 9 we see in action the exceptional performance of MCD, which has an F_1 score way higher than the rest of the models. COPOD, which had a comparable ROC AUC value to MCD, is the second best performer but its F_1 score is considerably lower than that of MCD. Another things that pops out in these results is that DIF is completely ineffectual. The same goes for kNN-avg, which in the ROC AUC results had a satisfactory performance but in practice is pretty much unusable due to its abysmally low Recall

5.5.3 Local Anomalies

Table 6: Results for ROC AUC: Global Anomalies

	ROC_AUC		
	.025	**Mean**	**.975**
MCD	0.998	1.000	1.000
GMM	0.931	0.965	0.994
kNN-avg	0.930	0.951	0.972
LUNAR	0.923	0.950	0.975
kNN	0.929	0.949	0.971
kNN-median	0.925	0.948	0.971
KDE	0.924	0.947	0.970
CD	0.903	0.939	0.974
LOF100	0.900	0.929	0.960
LOF20	0.888	0.929	0.965
FB	0.871	0.926	0.967
LOF10	0.854	0.913	0.962
INNE	0.861	0.907	0.952
COF	0.869	0.904	0.937
CBLOF	0.852	0.900	0.947
IForest	0.816	0.867	0.917
OCSVM	0.796	0.859	0.922
FastABOD	0.816	0.859	0.908
COPOD	0.811	0.857	0.906
Sampling	0.765	0.853	0.920
QMCD	0.813	0.846	0.894
Beta-VAE	0.759	0.811	0.870
PCA	0.760	0.810	0.868
ECOD	0.744	0.799	0.862
LODA	0.694	0.797	0.886
DIF	0.718	0.775	0.837
HBOS	0.712	0.766	0.828
SOS	0.640	0.702	0.801

Table 7: Results for Precision, Recall, f1: Global Anomalies

	Precision			Recall			f1		
	.025	**Mean**	**.975**	**.025**	**Mean**	**.975**	**.025**	**Mean**	**.975**
MCD	0.367	0.833	1.000	0.612	0.883	1.000	0.537	0.823	0.990
GMM	0.353	0.732	0.920	0.547	0.786	0.984	0.522	0.727	0.825
LUNAR	0.333	0.717	0.947	0.567	0.758	0.932	0.483	0.707	0.809
KDE	0.327	0.671	0.873	0.514	0.717	0.918	0.476	0.664	0.759
kNN	0.371	0.709	0.899	0.449	0.673	0.899	0.518	0.661	0.758
CD	0.333	0.662	0.860	0.506	0.711	0.932	0.486	0.657	0.745
kNN-median	0.403	0.730	0.911	0.390	0.630	0.881	0.521	0.646	0.753
LOF100	0.320	0.635	0.818	0.462	0.671	0.887	0.461	0.625	0.719
LOF20	0.348	0.625	0.774	0.358	0.619	0.900	0.462	0.593	0.714
INNE	0.307	0.583	0.740	0.423	0.634	0.877	0.451	0.582	0.675
FB	0.348	0.612	0.778	0.323	0.601	0.900	0.420	0.577	0.724
COF	0.280	0.581	0.787	0.444	0.619	0.803	0.404	0.574	0.671
CBLOF	0.287	0.567	0.753	0.422	0.608	0.839	0.410	0.563	0.671
LOF10	0.371	0.600	0.737	0.279	0.533	0.873	0.384	0.534	0.670
Sampling	0.247	0.519	0.753	0.350	0.554	0.777	0.344	0.514	0.639
IForest	0.253	0.519	0.727	0.391	0.554	0.760	0.366	0.513	0.614
OCSVM	0.273	0.492	0.627	0.346	0.540	0.782	0.392	0.493	0.579
COPOD	0.255	0.501	0.675	0.366	0.526	0.728	0.360	0.493	0.576
FastABOD	0.217	0.441	0.614	0.450	0.610	0.820	0.338	0.491	0.575
QMCD	0.220	0.451	0.627	0.339	0.485	0.679	0.316	0.448	0.532
LODA	0.187	0.393	0.613	0.227	0.426	0.673	0.233	0.391	0.537
PCA	0.207	0.390	0.520	0.283	0.424	0.623	0.297	0.389	0.463
Beta-VAE	0.208	0.387	0.514	0.278	0.422	0.624	0.298	0.387	0.460
ECOD	0.211	0.377	0.497	0.256	0.402	0.603	0.293	0.373	0.443
HBOS	0.173	0.365	0.527	0.279	0.387	0.542	0.248	0.360	0.442
SOS	0.180	0.311	0.433	0.226	0.340	0.550	0.244	0.310	0.375
kNN-avg	0.000	0.847	1.000	0.000	0.091	0.339	0.000	0.150	0.490
DIF	0.000	0.273	1.000	0.000	0.006	0.034	0.000	0.011	0.064

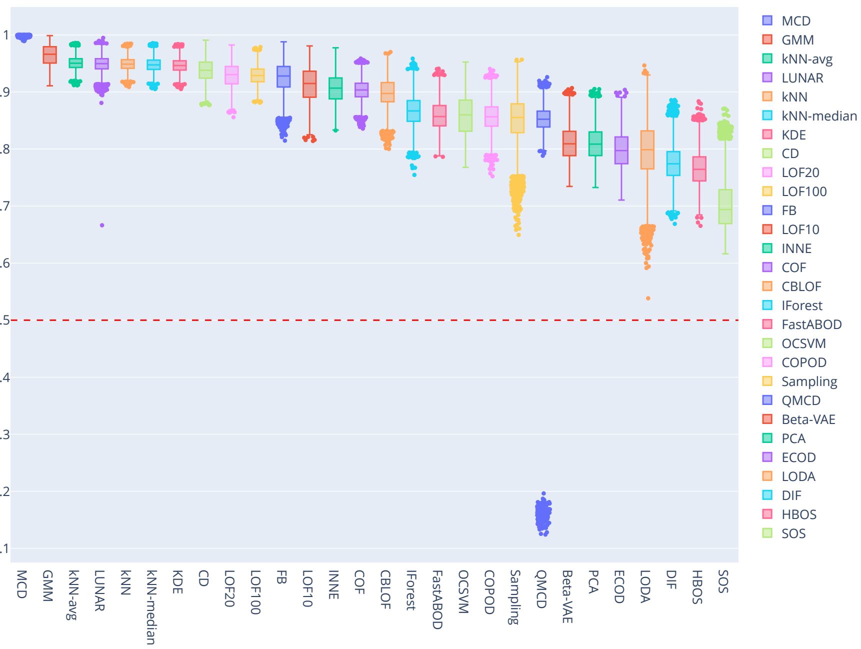

Figure 7: Boxplots for ROC AUC: Global Anomalies

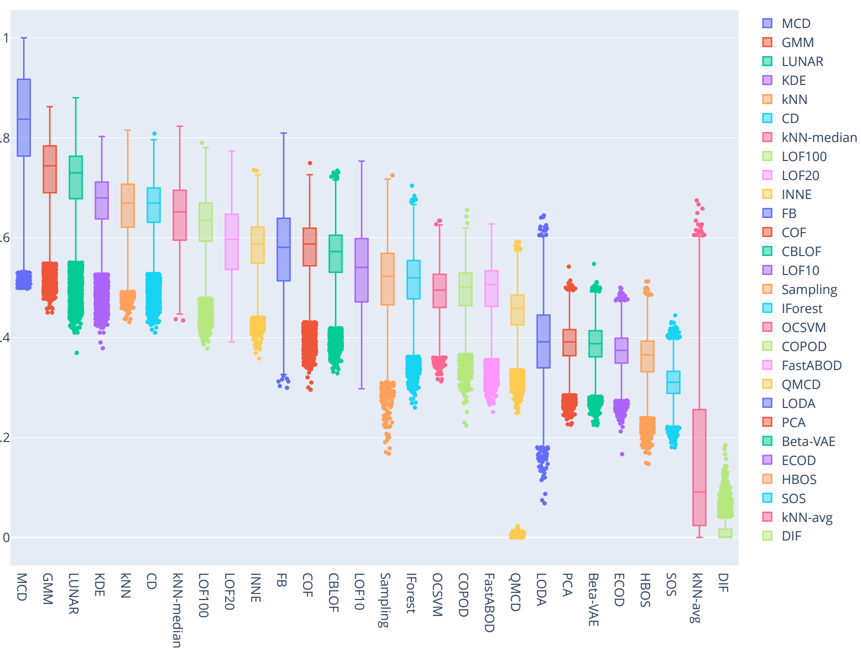

Figure 8: Boxplots for F_1 : Global Anomalies

Table 8: Results for ROC AUC: Clustered Anomalies

	ROC_AUC		
	.025	**Mean**	**.975**
MCD	0.620	0.966	1.000
COPOD	0.830	0.905	0.951
PCA	0.750	0.816	0.883
Beta-VAE	0.748	0.815	0.883
QMCD	0.739	0.809	0.877
ECOD	0.741	0.809	0.879
CD	0.661	0.789	0.932
LOF100	0.679	0.787	0.885
OCSVM	0.697	0.786	0.882
GMM	0.651	0.785	0.944
LODA	0.651	0.781	0.892
IForest	0.685	0.777	0.866
INNE	0.664	0.766	0.867
KDE	0.666	0.761	0.856
HBOS	0.680	0.750	0.824
kNN	0.636	0.740	0.849
kNN-avg	0.624	0.725	0.838
LUNAR	0.623	0.720	0.824
kNN-median	0.618	0.717	0.831
CBLOF	0.600	0.707	0.911
Sampling	0.520	0.703	0.866
FB	0.557	0.671	0.824
LOF20	0.533	0.651	0.814
FastABOD	0.562	0.636	0.728
LOF10	0.500	0.605	0.762
COF	0.494	0.595	0.758
SOS	0.508	0.548	0.613
DIF	0.300	0.355	0.418

Table 9: Results for Precision, Recall, f1: Clustered Anomalies

	Precision			Recall			f1		
	.025	**Mean**	**.975**	**.025**	**Mean**	**.975**	**.025**	**Mean**	**.975**
MCD	0.307	0.762	1.000	0.204	0.834	1.000	0.244	0.765	0.980
COPOD	0.263	0.594	0.831	0.382	0.618	0.844	0.376	0.577	0.711
PCA	0.227	0.427	0.573	0.311	0.466	0.672	0.324	0.426	0.509
Beta-VAE	0.225	0.424	0.572	0.307	0.464	0.673	0.322	0.424	0.507
QMCD	0.213	0.425	0.580	0.310	0.462	0.667	0.312	0.423	0.505
CD	0.287	0.412	0.500	0.260	0.477	0.831	0.315	0.421	0.526
ECOD	0.222	0.424	0.563	0.285	0.455	0.671	0.320	0.418	0.494
OCSVM	0.233	0.406	0.527	0.287	0.451	0.688	0.328	0.408	0.488
IForest	0.207	0.402	0.567	0.281	0.438	0.652	0.297	0.401	0.498
GMM	0.307	0.386	0.460	0.236	0.458	0.860	0.286	0.398	0.514
LODA	0.187	0.398	0.633	0.230	0.432	0.690	0.229	0.396	0.557
LOF100	0.231	0.353	0.448	0.211	0.399	0.685	0.255	0.357	0.463
HBOS	0.173	0.357	0.507	0.266	0.385	0.556	0.250	0.354	0.433
KDE	0.200	0.353	0.480	0.246	0.391	0.616	0.274	0.354	0.437
LUNAR	0.187	0.349	0.480	0.242	0.384	0.595	0.266	0.349	0.432
Sampling	0.140	0.345	0.620	0.169	0.375	0.653	0.169	0.343	0.549
CBLOF	0.164	0.337	0.480	0.227	0.382	0.765	0.204	0.340	0.526
INNE	0.213	0.333	0.440	0.218	0.378	0.645	0.258	0.338	0.424
kNN	0.210	0.360	0.488	0.193	0.329	0.562	0.244	0.327	0.418
kNN-median	0.202	0.352	0.485	0.159	0.277	0.479	0.212	0.295	0.392
FastABOD	0.109	0.231	0.347	0.238	0.334	0.486	0.168	0.262	0.335
FB	0.155	0.245	0.341	0.137	0.248	0.481	0.164	0.234	0.322
LOF20	0.137	0.223	0.321	0.123	0.229	0.456	0.143	0.215	0.301
COF	0.107	0.200	0.307	0.123	0.224	0.425	0.130	0.201	0.278
LOF10	0.107	0.201	0.310	0.097	0.179	0.345	0.114	0.180	0.256
SOS	0.073	0.175	0.287	0.127	0.183	0.268	0.098	0.171	0.233
kNN-avg	0.000	0.536	1.000	0.000	0.027	0.090	0.000	0.050	0.157
DIF	0.000	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000

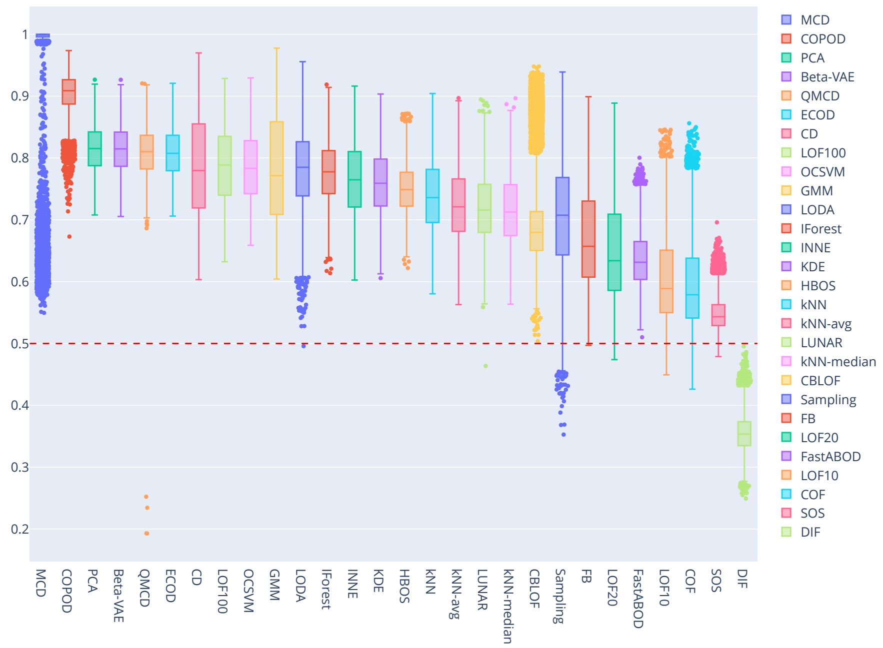

Figure 9: Boxplots for ROC AUC: Clustered Anomalies

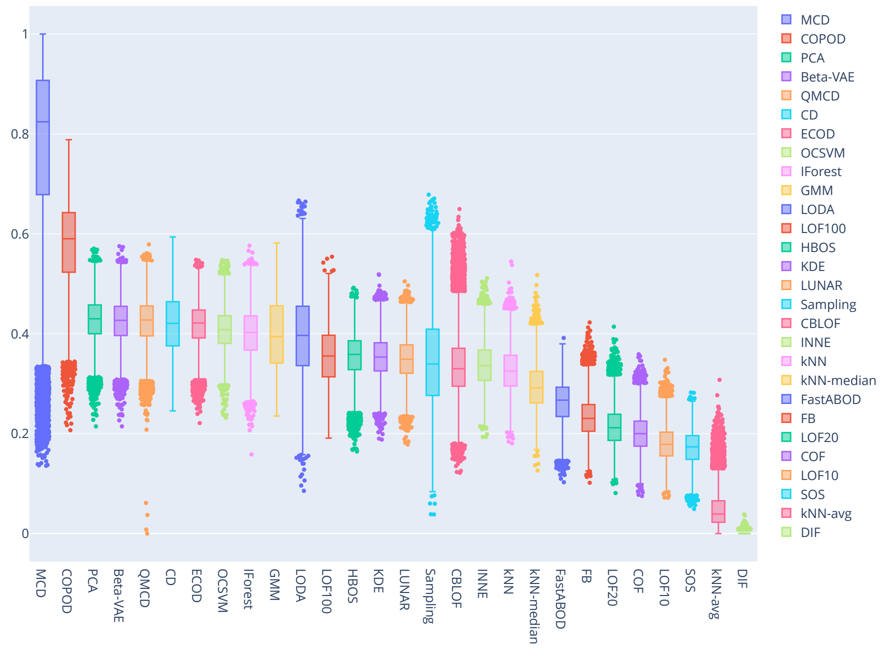

Figure 10: Boxplots for F_1 : Clustered Anomalies

Table 10: Results for ROC AUC: Local Anomalies

	ROC_AUC		
	.025	**Mean**	**.975**
GMM	0.805	0.845	0.883
LOF20	0.789	0.832	0.872
INNE	0.789	0.831	0.871
FB	0.788	0.830	0.871
CD	0.786	0.829	0.869
OCSVM	0.786	0.827	0.867
LOF10	0.783	0.827	0.867
LOF100	0.784	0.826	0.868
KDE	0.777	0.821	0.863
CBLOF	0.776	0.820	0.861
kNN	0.769	0.814	0.857
kNN-avg	0.763	0.810	0.855
kNN-median	0.758	0.805	0.851
MCD	0.736	0.802	0.865
ECOD	0.753	0.796	0.837
PCA	0.752	0.794	0.837
Beta-VAE	0.751	0.794	0.837
COF	0.747	0.791	0.835
IForest	0.744	0.789	0.833
QMCD	0.746	0.788	0.829
SOS	0.735	0.781	0.828
LUNAR	0.729	0.781	0.832
LODA	0.716	0.775	0.825
COPOD	0.729	0.774	0.817
FastABOD	0.713	0.760	0.809
Sampling	0.699	0.755	0.809
DIF	0.687	0.736	0.785
HBOS	0.678	0.725	0.772

Based on Table 10, we see that in the presence of local outliers all the methods perform quite well. There are some variations between the different methods but we cannot identify any methods that perform significantly better (or worse) than the others.

Based on the results for Precision, Recall and F_1 as they are found in Table 11 we see that despite the good ROC AUC value for the kNN-avg model, its F_1 score is noticeably lower than all the other methods. It is the only model with a performance considerably different from all the other models.

5.5.4 Dependency Anomalies

Among the first things one may notice in Table 12 is that about half the models used are below the 0.500 value for ROC AUC that a random classifier has. This can be attributed to the characteristics of our original “seed” dataset in conjunction with the way we generated dependency anomalies: the anomalies

Table 11: Results for Precision, Recall, f1: Local Anomalies

	Precision			**Recall**			**f1**		
	.025	**Mean**	**.975**	**.025**	**Mean**	**.975**	**.025**	**Mean**	**.975**
GMM	0.227	0.534	0.780	0.453	0.554	0.695	0.330	0.521	0.623
LOF20	0.232	0.536	0.781	0.407	0.508	0.648	0.325	0.499	0.595
LOF100	0.218	0.515	0.760	0.428	0.525	0.655	0.313	0.498	0.601
INNE	0.213	0.510	0.753	0.432	0.528	0.661	0.312	0.497	0.600
CD	0.220	0.510	0.753	0.431	0.528	0.663	0.317	0.496	0.600
FB	0.232	0.535	0.780	0.402	0.501	0.639	0.325	0.495	0.591
KDE	0.213	0.506	0.753	0.431	0.524	0.652	0.308	0.493	0.598
OCSVM	0.213	0.505	0.747	0.429	0.523	0.660	0.310	0.492	0.595
kNN	0.235	0.540	0.785	0.386	0.481	0.610	0.323	0.488	0.581
CBLOF	0.207	0.501	0.740	0.423	0.517	0.650	0.304	0.487	0.590
LOF10	0.243	0.550	0.792	0.368	0.470	0.609	0.329	0.485	0.574
kNN-median	0.248	0.557	0.802	0.349	0.442	0.569	0.325	0.473	0.561
LUNAR	0.193	0.472	0.707	0.399	0.487	0.613	0.283	0.459	0.565
SOS	0.187	0.454	0.680	0.382	0.469	0.597	0.273	0.442	0.543
COF	0.187	0.448	0.673	0.379	0.462	0.588	0.269	0.436	0.538
ECOD	0.190	0.447	0.669	0.356	0.454	0.585	0.272	0.432	0.532
PCA	0.187	0.438	0.660	0.367	0.454	0.580	0.265	0.427	0.530
IForest	0.187	0.438	0.667	0.365	0.453	0.581	0.265	0.427	0.531
Beta-VAE	0.185	0.438	0.662	0.367	0.453	0.580	0.264	0.426	0.528
Sampling	0.180	0.432	0.667	0.356	0.445	0.567	0.256	0.420	0.530
FastABOD	0.149	0.374	0.578	0.415	0.504	0.627	0.232	0.411	0.526
COPOD	0.175	0.422	0.637	0.335	0.430	0.556	0.253	0.408	0.508
LODA	0.173	0.418	0.653	0.303	0.431	0.566	0.241	0.406	0.524
QMCD	0.173	0.417	0.627	0.346	0.432	0.560	0.252	0.406	0.503
MCD	0.113	0.397	0.713	0.159	0.415	0.623	0.136	0.388	0.585
HBOS	0.133	0.345	0.540	0.280	0.354	0.462	0.194	0.335	0.432
kNN-avg	0.600	0.888	1.000	0.054	0.110	0.197	0.102	0.192	0.319
DIF	0.000	0.592	1.000	0.000	0.012	0.049	0.000	0.024	0.091

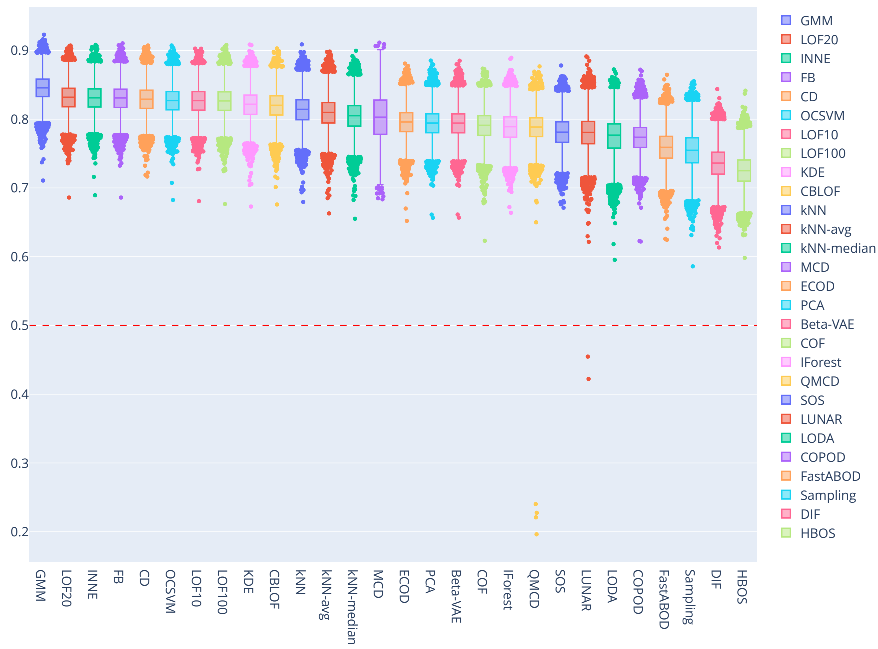

Figure 11: Boxplots for ROC AUC: Local Anomalies

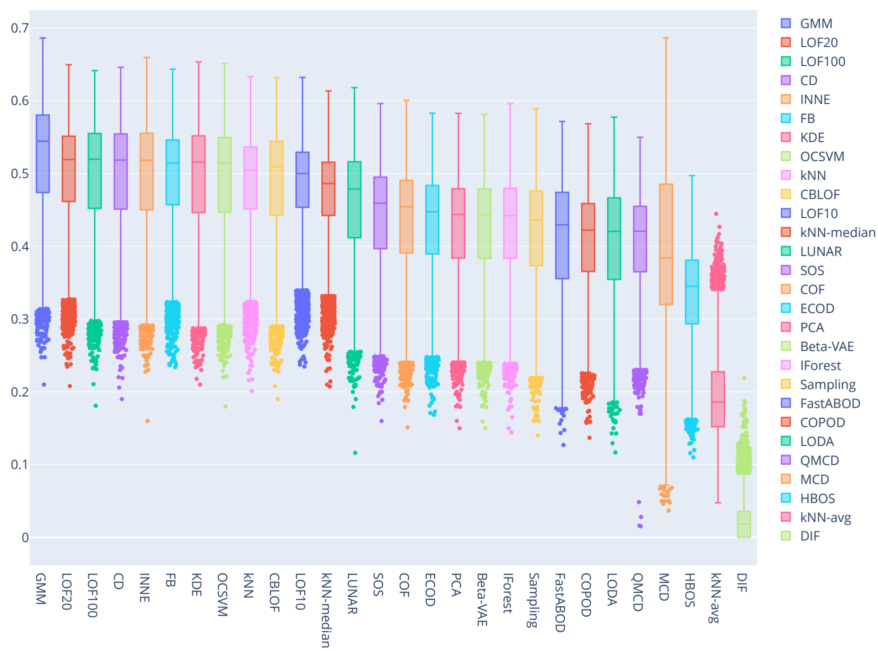

Figure 12: Boxplots for F_1 : Local Anomalies

Table 12: Results for ROC AUC: Dependency Anomalies

	ROC_AUC		
	.025	**Mean**	**.975**
DIF	0.532	0.590	0.653
HBOS	0.535	0.589	0.649
Beta-VAE	0.525	0.579	0.639
PCA	0.524	0.578	0.638
IForest	0.519	0.576	0.642
ECOD	0.508	0.560	0.618
COPOD	0.508	0.560	0.620
LODA	0.482	0.555	0.631
QMCD	0.436	0.545	0.617
OCSVM	0.469	0.521	0.578
MCD	0.417	0.520	0.636
INNE	0.451	0.502	0.561
SOS	0.444	0.496	0.551
LOF10	0.436	0.493	0.552
COF	0.436	0.492	0.551
FB	0.436	0.491	0.547
LOF20	0.432	0.489	0.547
CD	0.432	0.485	0.546
GMM	0.430	0.482	0.544
LOF100	0.417	0.471	0.526
Sampling	0.393	0.470	0.571
CBLOF	0.411	0.468	0.529
FastABOD	0.385	0.441	0.499
KDE	0.379	0.433	0.494
LUNAR	0.376	0.432	0.497
kNN	0.371	0.427	0.487
kNN-median	0.371	0.425	0.484
kNN-avg	0.370	0.424	0.484

Table 13: Results for Precision, Recall, f1: Dependency Anomalies

	Precision			Recall			f1		
	.025	**Mean**	**.975**	**.025**	**Mean**	**.975**	**.025**	**Mean**	**.975**
HBOS	0.080	0.219	0.360	0.158	0.226	0.316	0.116	0.213	0.292
Beta-VAE	0.079	0.207	0.342	0.149	0.213	0.304	0.108	0.201	0.279
PCA	0.080	0.206	0.340	0.148	0.212	0.300	0.107	0.200	0.276
IForest	0.073	0.198	0.333	0.137	0.204	0.301	0.102	0.192	0.274
QMCD	0.067	0.191	0.327	0.111	0.197	0.286	0.090	0.186	0.267
ECOD	0.063	0.188	0.316	0.105	0.185	0.267	0.082	0.181	0.268
LODA	0.060	0.185	0.353	0.110	0.189	0.298	0.081	0.180	0.292
MCD	0.067	0.181	0.347	0.088	0.195	0.337	0.080	0.179	0.296
COPOD	0.062	0.185	0.317	0.105	0.183	0.265	0.082	0.178	0.265
OCSVM	0.047	0.152	0.273	0.096	0.153	0.220	0.067	0.146	0.219
COF	0.047	0.149	0.273	0.091	0.150	0.216	0.060	0.143	0.218
FastABOD	0.038	0.126	0.228	0.107	0.173	0.243	0.057	0.140	0.217
SOS	0.040	0.141	0.260	0.085	0.141	0.200	0.056	0.135	0.207
LOF20	0.043	0.148	0.271	0.077	0.133	0.194	0.057	0.134	0.205
Sampling	0.040	0.139	0.267	0.074	0.140	0.231	0.053	0.134	0.221
FB	0.041	0.148	0.274	0.074	0.131	0.191	0.054	0.134	0.205
CD	0.040	0.134	0.240	0.080	0.135	0.197	0.055	0.129	0.196
LOF10	0.042	0.151	0.279	0.065	0.121	0.181	0.052	0.128	0.198
CBLOF	0.033	0.128	0.240	0.075	0.135	0.196	0.049	0.126	0.197
GMM	0.040	0.130	0.240	0.077	0.131	0.194	0.053	0.125	0.192
LOF100	0.034	0.132	0.247	0.071	0.129	0.187	0.047	0.125	0.196
INNE	0.033	0.127	0.233	0.071	0.128	0.184	0.049	0.122	0.191
LUNAR	0.040	0.126	0.233	0.073	0.127	0.188	0.050	0.121	0.186
KDE	0.033	0.121	0.227	0.068	0.121	0.179	0.046	0.116	0.182
kNN	0.032	0.120	0.230	0.048	0.098	0.151	0.038	0.103	0.165
kNN-median	0.029	0.119	0.231	0.038	0.083	0.133	0.034	0.094	0.152
kNN-avg	0.000	0.108	0.667	0.000	0.002	0.014	0.000	0.004	0.027
DIF	0.000	0.116	1.000	0.000	0.001	0.013	0.000	0.003	0.025

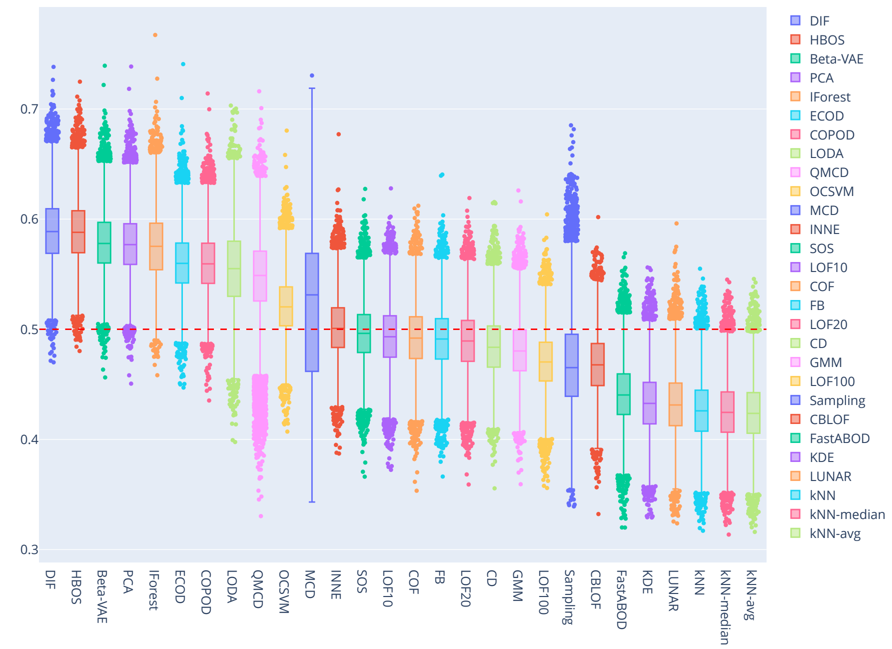

Figure 13: Boxplots for ROC AUC: Dependency Anomalies

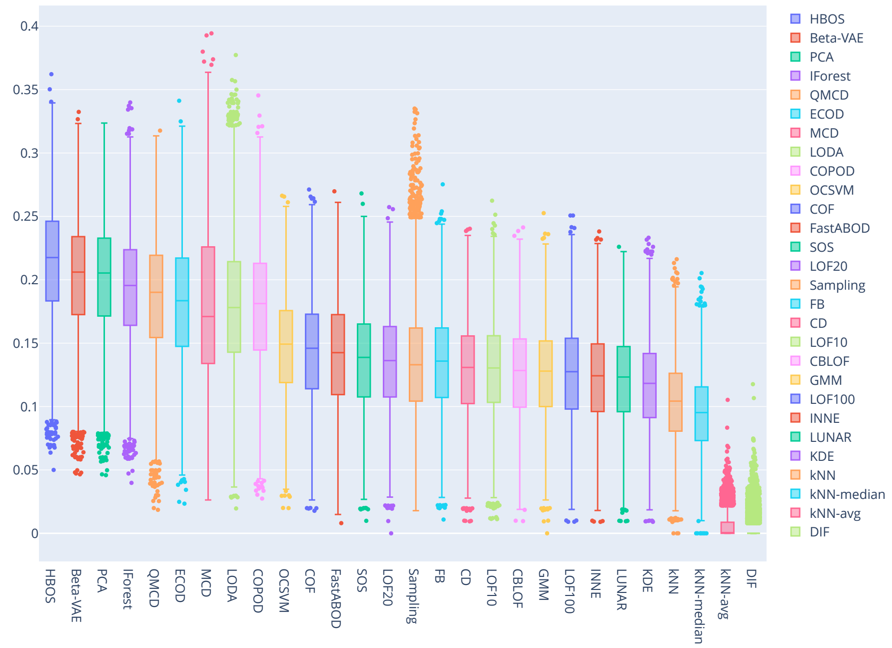

Figure 14: Boxplots for F_1 : Dependency Anomalies

that we generated are not significantly different from the typical samples so the outlier detection techniques struggle. It is worthwhile to note that when using a different "seed", the method we propose for generating outliers can be tuned in order to generate outliers that differ more from typical observations. We sadly could not achieve that in this case.

Moving on the evaluation of the results in Table 12, we notice that DIF, which performed terribly in the other cases, is the best performer here. HBOS, which also struggled in the presence of global and local anomalies performs practically as well as DIF. In regards to the worst performers, we observe that all variants of kNN struggle the most, with KDE and LUNAR also performing badly

Observing Table 13, we notice that the methods achieving the highest F_1 scores are Principal Component Analysis based Outlier Detection (PCA), Beta-Variational AutoEncoder (Beta-VAE), and HBOS. kNN-avg and DIF have extremely low F_1 scores and kNN-median is not much better.

5.5.5 Mixed Anomalies

Figure 15: Boxplots for ROC AUC: Mixed Anomalies

The results in Table 14 remind us somewhat of the results for local anomalies with their common characteristic being that the performance of most methods is quite similar to the rest. Unlike the case of local outliers, here we can identify a few bad performers, namely DIF, SOS, and FastABOD.

Moving on the Precision, Recall and F_1 results in Table 15, the latter show that FastABOD and SOS, while not being anywhere near the top, are not so

Table 14: Results for ROC AUC: Mixed Anomalies

	ROC_AUC		
	.025	**Mean**	**.975**
COPOD	0.754	0.800	0.840
OCSVM	0.759	0.798	0.839
IForest	0.752	0.796	0.840
INNE	0.757	0.795	0.834
ECOD	0.753	0.794	0.837
Beta-VAE	0.753	0.794	0.835
PCA	0.752	0.793	0.835
QMCD	0.751	0.791	0.834
LOF100	0.748	0.787	0.825
FB	0.742	0.783	0.829
GMM	0.744	0.782	0.827
LOF20	0.739	0.781	0.827
CBLOF	0.721	0.778	0.819
KDE	0.727	0.770	0.812
LODA	0.709	0.768	0.820
LOF10	0.717	0.766	0.820
kNN	0.720	0.763	0.807
MCD	0.716	0.763	0.808
CD	0.719	0.760	0.811
kNN-avg	0.713	0.758	0.803
kNN-median	0.709	0.754	0.801
LUNAR	0.696	0.746	0.795
COF	0.696	0.745	0.799
HBOS	0.700	0.745	0.793
Sampling	0.657	0.727	0.788
FastABOD	0.650	0.697	0.751
SOS	0.637	0.684	0.746
DIF	0.577	0.629	0.679

Table 15: Results for Precision, Recall, f1: Mixed Anomalies

	Precision			Recall			f1		
	.025	**Mean**	**.975**	**.025**	**Mean**	**.975**	**.025**	**Mean**	**.975**
CBLOF	0.207	0.511	0.747	0.406	0.510	0.625	0.303	0.492	0.595
LOF100	0.209	0.503	0.743	0.425	0.521	0.641	0.305	0.490	0.591
OCSVM	0.207	0.490	0.720	0.418	0.518	0.643	0.303	0.482	0.580
INNE	0.207	0.483	0.713	0.410	0.511	0.633	0.301	0.476	0.574
COPOD	0.205	0.488	0.736	0.388	0.502	0.625	0.294	0.475	0.590
kNN	0.228	0.518	0.752	0.371	0.470	0.597	0.316	0.473	0.562
KDE	0.200	0.480	0.713	0.412	0.507	0.625	0.295	0.472	0.574
GMM	0.207	0.476	0.700	0.408	0.505	0.633	0.301	0.469	0.563
FB	0.226	0.494	0.711	0.359	0.474	0.618	0.318	0.463	0.551
LUNAR	0.193	0.469	0.707	0.406	0.492	0.603	0.283	0.460	0.565
LOF20	0.223	0.486	0.696	0.361	0.473	0.625	0.315	0.458	0.542
IForest	0.200	0.465	0.700	0.393	0.491	0.625	0.286	0.457	0.561
kNN-median	0.243	0.532	0.763	0.332	0.430	0.562	0.318	0.456	0.539
ECOD	0.201	0.465	0.683	0.363	0.479	0.611	0.294	0.452	0.546
QMCD	0.193	0.450	0.667	0.381	0.476	0.604	0.277	0.443	0.538
PCA	0.193	0.448	0.667	0.379	0.474	0.603	0.277	0.441	0.539
Beta-VAE	0.191	0.446	0.667	0.377	0.473	0.604	0.279	0.439	0.537
CD	0.193	0.444	0.653	0.377	0.472	0.604	0.283	0.438	0.530
LOF10	0.231	0.483	0.691	0.314	0.423	0.583	0.312	0.431	0.510
LODA	0.173	0.419	0.653	0.312	0.443	0.583	0.252	0.412	0.535
Sampling	0.167	0.416	0.660	0.327	0.438	0.565	0.242	0.409	0.531
COF	0.180	0.409	0.607	0.338	0.436	0.575	0.257	0.404	0.497
HBOS	0.153	0.372	0.573	0.310	0.391	0.512	0.219	0.365	0.462
FastABOD	0.131	0.317	0.490	0.359	0.448	0.573	0.207	0.356	0.458
MCD	0.107	0.355	0.593	0.188	0.378	0.575	0.143	0.350	0.514
SOS	0.147	0.336	0.507	0.275	0.357	0.483	0.215	0.331	0.412
kNN-avg	0.667	0.916	1.000	0.029	0.082	0.173	0.056	0.149	0.290
DIF	0.000	0.389	1.000	0.000	0.006	0.029	0.000	0.012	0.055

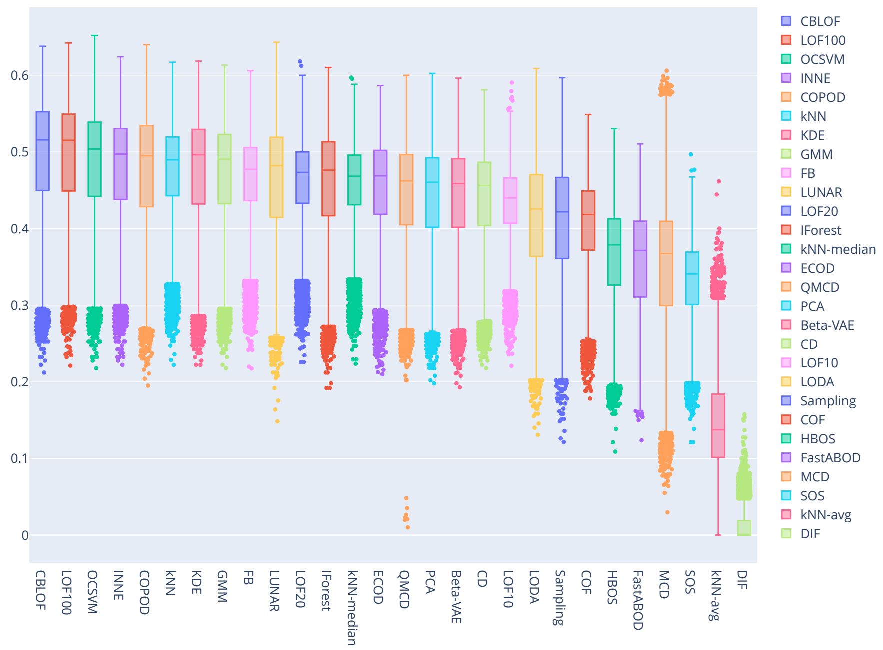

Figure 16: Boxplots for F_1 : Mixed Anomalies

bad as the ROC AUC score would have us believe. DIF appears unusable.

5.5.6 General Observations

Table 16: ROC AUC of Anomaly Detection Methods in the presence of different type of anomalies

	clustered	local	global	dependency	mixed
FastABOD	0.636	0.759	0.859	0.441	0.697
ECOD	0.808	0.796	0.799	0.561	0.794
COPOD	0.904	0.773	0.857	0.560	0.800
SOS	0.548	0.781	0.702	0.496	0.684
QMCD	0.809	0.788	0.846	0.545	0.791
KDE	0.760	0.821	0.947	0.433	0.770
Sampling	0.704	0.755	0.852	0.470	0.727
GMM	0.784	0.845	0.965	0.482	0.783
PCA	0.815	0.794	0.810	0.578	0.794
MCD	0.965	0.802	1.000	0.521	0.762
CD	0.787	0.829	0.938	0.485	0.761
OCSVM	0.786	0.827	0.859	0.521	0.799
LOF10	0.604	0.827	0.913	0.493	0.767
LOF20	0.650	0.831	0.929	0.490	0.781
LOF100	0.787	0.826	0.929	0.471	0.787
COF	0.595	0.791	0.903	0.492	0.746
CBLOF	0.706	0.820	0.899	0.468	0.778
HBOS	0.750	0.725	0.766	0.589	0.745
kNN	0.739	0.814	0.949	0.427	0.764
kNN-avg	0.725	0.809	0.951	0.424	0.758
kNN-median	0.717	0.805	0.948	0.425	0.754
IForest	0.777	0.789	0.867	0.576	0.796
INNE	0.765	0.831	0.906	0.502	0.795
DIF	0.355	0.736	0.775	0.590	0.629
FB	0.670	0.830	0.925	0.491	0.784
LODA	0.781	0.775	0.797	0.555	0.768
LUNAR	0.719	0.780	0.949	0.432	0.746
Beta-VAE	0.815	0.794	0.810	0.579	0.794

Concluding the evaluation of our results in the presence of different kinds of anomalies, we include Table 16 and Table 17, two tables that show the ROC AUC and F_1 scores for the methods we utilized across the different kinds of anomaly simulations.

In the previous sections we saw that among all the anomaly detection models we evaluated there is no single one that consistently performs better than all the others. There are however two methods that consistently perform worse in regards to their F_1 score: kNN-avg and DIF. We can conclude the following: in the presence of unknown types of anomalies (as would be the case when dealing with real data) it would be best to choose those methods which have the most consistent performance across all different anomaly regimes.

In order to present the relative performance of the anomaly detection methods we utilized across all our simulations, we present Table 18 and Table 19, two tables that show the ranks for the ROC AUC and F_1 scores respectively.

Table 17: F_1 Score of Anomaly Detection Methods in the presence of different type of anomalies

	clustered	local	global	dependency	mixed
FastABOD	0.262	0.410	0.491	0.141	0.357
ECOD	0.418	0.431	0.373	0.181	0.454
COPOD	0.576	0.408	0.492	0.178	0.476
SOS	0.171	0.441	0.310	0.136	0.332
QMCD	0.423	0.405	0.447	0.186	0.444
KDE	0.354	0.493	0.663	0.116	0.474
Sampling	0.345	0.420	0.513	0.134	0.410
GMM	0.398	0.520	0.726	0.126	0.470
PCA	0.427	0.427	0.388	0.201	0.443
MCD	0.765	0.388	0.822	0.180	0.351
CD	0.420	0.496	0.656	0.129	0.439
OCSVM	0.408	0.492	0.492	0.147	0.484
LOF10	0.180	0.485	0.534	0.129	0.432
LOF20	0.215	0.499	0.592	0.134	0.460
LOF100	0.357	0.497	0.624	0.125	0.492
COF	0.201	0.435	0.574	0.143	0.405
CBLOF	0.340	0.487	0.563	0.126	0.494
HBOS	0.354	0.334	0.359	0.213	0.365
kNN	0.327	0.487	0.661	0.103	0.474
kNN-avg	0.050	0.193	0.151	0.004	0.147
kNN-median	0.294	0.473	0.645	0.094	0.458
IForest	0.401	0.426	0.512	0.193	0.459
INNE	0.338	0.496	0.581	0.122	0.477
DIF	0.000	0.024	0.011	0.003	0.012
FB	0.234	0.495	0.576	0.134	0.465
LODA	0.396	0.406	0.391	0.179	0.414
LUNAR	0.349	0.458	0.705	0.121	0.461
Beta-VAE	0.424	0.426	0.386	0.202	0.441

Table 18: ROC AUC Descending Ranks of Anomaly Detection Methods in the presence of different type of anomalies

	clustered	**local**	**global**	**dependency**	**mixed**
FastABOD	24	25	17	23	26
ECOD	6	15	24	6	5
COPOD	2	24	19	7	1
SOS	27	21	28	13	27
QMCD	5	20	21	9	8
KDE	14	9	7	24	14
Sampling	21	26	20	21	25
GMM	10	1	2	19	11
PCA	3	16	23	4	7
MCD	1	14	1	11	18
CD	7	5	8	18	19
OCSVM	9	6	18	10	2
LOF10	25	7	12	14	16
LOF20	23	2	9	17	12
LOF100	8	8	10	20	9
COF	26	18	14	15	23
CBLOF	20	10	15	22	13
HBOS	15	28	27	2	24
kNN	16	11	5	26	17
kNN-avg	17	12	3	28	20
kNN-median	19	13	6	27	21
IForest	12	19	16	5	3
INNE	13	3	13	12	4
DIF	28	27	26	1	28
FB	22	4	11	16	10
LODA	11	23	25	8	15
LUNAR	18	22	4	25	22
Beta-VAE	4	17	22	3	6

Table 19: F_1 Descending Ranks of Anomaly Detection Methods in the presence of different type of anomalies

	clustered	local	global	dependency	mixed
FastABOD	21	21	19	12	24
ECOD	7	16	24	6	14
COPOD	2	22	18	9	5
SOS	26	14	26	13	26
QMCD	5	24	20	5	15
KDE	14	7	4	24	7
Sampling	16	20	15	15	21
GMM	10	1	2	20	8
PCA	3	17	22	3	16
MCD	1	25	1	7	25
CD	6	4	6	17	18
OCSVM	8	8	17	10	3
LOF10	25	11	14	18	19
LOF20	23	2	9	14	11
LOF100	12	3	8	21	2
COF	24	15	12	11	22
CBLOF	17	10	13	19	1
HBOS	13	26	25	1	23
kNN	19	9	5	25	6
kNN-avg	27	27	27	27	27
kNN-median	20	12	7	26	13
IForest	9	19	16	4	12
INNE	18	5	10	22	4
DIF	28	28	28	28	28
FB	22	6	11	16	9
LODA	11	23	21	8	20
LUNAR	15	13	3	23	10
Beta-VAE	4	18	23	2	17

Table 20: Mean Ranks for ROC AUC and F_1 Score of Anomaly Detection Methods

	ROC AUC
GMM	8.6
OCSVM	9.0
MCD	9.0
INNE	9.0
Beta-VAE	10.4
PCA	10.6
COPOD	10.6
IForest	11.0
LOF100	11.0
ECOD	11.2
CD	11.4
QMCD	12.6
FB	12.6
LOF20	12.6
KDE	13.6
LOF10	14.8
kNN	15.0
kNN-avg	16.0
CBLOF	16.0
LODA	16.4
kNN-median	17.2
LUNAR	18.2
HBOS	19.2
COF	19.2
DIF	22.0
Sampling	22.6
FastABOD	23.0
SOS	23.2

	F_1
GMM	8.2
OCSVM	9.2
LOF100	9.2
CD	10.2
KDE	11.2
COPOD	11.2
MCD	11.8
LOF20	11.8
INNE	11.8
CBLOF	12.0
IForest	12.0
PCA	12.2
kNN	12.8
LUNAR	12.8
Beta-VAE	12.8
FB	12.8
ECOD	13.4
QMCD	13.8
kNN-median	15.6
LODA	16.6
COF	16.8
LOF10	17.4
Sampling	17.4
HBOS	17.6
FastABOD	19.4
SOS	21.0
kNN-avg	27.0
DIF	28.0

We also include Table 20 in order to evaluate the mean rank of each method across all different anomaly types.

The first observation when inspecting the mean ranks of the different anomaly detection methods is that the results for ROC AUC and F_1 score are somewhat different. This is an issue which we commented on in an earlier section of our work. The results for ROC AUC represent the potential of each anomaly detector when using an optimal threshold. The results for F_1 are obtained by using the thresholding technique we described in an earlier chapter[21].

Despite the differences between the ROC AUC and F_1 scores, there are also a lot of similarities. We see that the GMM method is in the first place. We have to note however that this may not be representative of actual performance in real data sets, since the simulation method we used generates data based on Gaussian mixtures, which may skew the results. Additionally, we observe that One Class Support Vector Machines are in the second place. Other good performers are Minimum Covariance Determinant based Outlier Detection (MCD), Local Outlier Factor with 100 neighbours (LOF100), Cook's Distance based Outlier Detection (CD), Copula Based Outlier Detection (COPOD), Isolation Forest (IForest) and Isolation-based Anomaly Detection Using Nearest-Neighbor Ensembles (INNE). Among the worst performers are Histogram-based Outlier Score (HBOS), Stochastic Outlier Selection (SOS), Fast Angle Based Outlier Detection (FastABOD), Deep Isolation Forest (DIF) and Rapid distance-based outlier detection via sampling (Sampling).

6 Conclusion

Our work concerned the use of unsupervised machine learning techniques in order to detect fraud in automobile insurance, or to be more precise automobile insurance claims. As we saw, the problem we are trying to tackle falls under the more general umbrella of anomaly detection techniques.

Due to the confidential nature of automobile insurance claims data sets, we had very limited access to data sets that we could use in our work, so we chose to move forward with a simulation study. Our simulations were based on a real dataset and we used various parametric techniques in order to generate different kinds of outliers. A wide array of different anomaly detection methods were evaluated across different kinds of simulations. We observed that the performance of the various methods depended quite a lot on the types of outlier present in each simulated data set. We could not identify any method that performed better than the others across all different simulations. However, we identified two methods that consistently performed worse, indicating that it would be best to avoid their use.

Concluding our work we are left with a number of open questions which could provide the inspiration for further research on the subject. An interesting question would be whether we can find ways to detect the type of outliers that exist in our data and use the model/models which perform better in the presence of such outliers. Another appealing avenue for research would be to conduct a simulation study similar to the one we present here, but using various

[21] To reiterate, we select as outliers a constant number of observations with the highest outliers scores. The number of observations is chosen based on the contamination factor we use

statistical measures to estimate the prevalence of outliers in the data, instead of pre-specifying a contamination factor to the anomaly detection algorithms as we did.

References

Ai, Jing et al. (Mar. 2013). "A Robust Unsupervised Method for Fraud Rate Estimation". In: *Journal of Risk and Insurance* 80.1, pp. 121–143. ISSN: 00224367. DOI: 10.1111/j.1539-6975.2012.01467.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6975.2012.01467.x (visited on 10/23/2023) (cit. on p. 13).

Artís, Manuel, Mercedes Ayuso, and Montserrat Guillén (Mar. 1999). "Modelling Different Types of Automobile Insurance Fraud Behaviour in the Spanish Market". In: *Insurance: Mathematics and Economics* 24.1-2, pp. 67–81. ISSN: 01676687. DOI: 10.1016/S0167-6687(98)00038-9. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167668798000389 (visited on 10/20/2023) (cit. on pp. 9, 11).

— (Sept. 2002). "Detection of Automobile Insurance Fraud With Discrete Choice Models and Misclassified Claims". In: *Journal of Risk and Insurance* 69.3, pp. 325–340. ISSN: 0022-4367, 1539-6975. DOI: 10.1111/1539-6975.00022. URL: https://onlinelibrary.wiley.com/doi/10.1111/1539-6975.00022 (visited on 10/20/2023) (cit. on p. 11).

Bauder, Richard, Raquel Da Rosa, and Taghi Khoshgoftaar (July 2018). "Identifying Medicare Provider Fraud with Unsupervised Machine Learning". In: *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. 2018 IEEE International Conference on Information Reuse and Integration for Data Science (IRI). Salt Lake City, UT: IEEE, pp. 285–292. ISBN: 978-1-5386-2659-7. DOI: 10.1109/IRI.2018.00051. URL: https://ieeexplore.ieee.org/document/8424722/ (visited on 11/07/2023) (cit. on p. 14).

Belhadji, El-Bachir and Georges Dionne (1998). "Development of an Expert System for the Automatic Detection of Automobile Insurance Fraud". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.134768. URL: http://www.ssrn.com/abstract=134768 (visited on 10/23/2023) (cit. on pp. 7, 11).

Belhadji, El Bachir, George Dionne, and Faouzi Tarkhani (Oct. 2000). "A Model for the Detection of Insurance Fraud". In: *The Geneva Papers on Risk and Insurance - Issues and Practice* 25.4, pp. 517–538. ISSN: 1018-5895, 1468-0440. DOI: 10.1111/1468-0440.00080. URL: http://link.springer.com/10.1111/1468-0440.00080 (visited on 11/07/2023) (cit. on p. 10).

Benedek, Botond, Cristina Ciumas, and Bálint Zsolt Nagy (Aug. 2, 2022). "Automobile Insurance Fraud Detection in the Age of Big Data – a Systematic and Comprehensive Literature Review". In: *Journal of Financial Regulation and Compliance* 30.4, pp. 503–523. ISSN: 1358-1988, 1358-1988. DOI: 10.1108/JFRC-11-2021-0102. URL: https://www.emerald.com/insight/content/doi/10.1108/JFRC-11-2021-0102/full/html (visited on 10/17/2023) (cit. on pp. 5, 7, 8, 10, 16, 20, 30).

Benedek, Botond and Bálint Zsolt Nagy (2023). "Traditional versus AI-Based Fraud Detection: Cost Efficiency in the Field of Automobile Insurance". In: *Financial and Economic Review* 22.2, pp. 77–98. ISSN: 24159271, 2415928X. DOI: 10.33893/FER.22.2.77. URL: https://en-hitelintezetiszemle.mnb.hu/letoltes/fer-22-2-st3-benedek-nagy.pdf (visited on 10/17/2023) (cit. on p. 9).

Bermúdez, Ll. et al. (Apr. 2008). “A Bayesian Dichotomous Model with Asymmetric Link for Fraud in Insurance”. In: *Insurance: Mathematics and Economics* 42.2, pp. 779–786. ISSN: 01676687. DOI: 10.1016/j.insmatheco.2007.08.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167668707000947 (visited on 06/02/2023) (cit. on p. 13).

Bolance, Catalina, Mercedes Ayuso, and Montserrat Guillen (Mar. 2012). “A Nonparametric Approach to Analyzing Operational Risk with an Application to Insurance Fraud”. In: *The Journal of Operational Risk* 7.1, pp. 57–75. ISSN: 17446740. DOI: 10.21314/JOP.2012.103. URL: http://www.risk.net/journal-of-operational-risk/technical-paper/2164348/nonparametric-approach-analyzing-operational-risk-application-insurance-fraud (visited on 10/23/2023) (cit. on p. 14).

Bouzgarne, Itri et al. (2020). “Empirical Oversampling Threshold Strategy for Machine Learning Performance Optimisation in Insurance Fraud Detection”. In: *International Journal of Advanced Computer Science and Applications* 11.10. ISSN: 21565570, 2158107X. DOI: 10.14569/IJACSA.2020.0111054. URL: http://thesai.org/Publications/ViewPaper?Volume=11&Issue=10&Code=IJACSA&SerialNo=54 (visited on 10/23/2023) (cit. on p. 13).

Breunig, Markus M. et al. (May 16, 2000). “LOF: Identifying Density-Based Local Outliers”. In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. SIGMOD/PODS00: ACM International Conference on Management of Data and Symposium on Principles of Database Systems. Dallas Texas USA: ACM, pp. 93–104. ISBN: 978-1-58113-217-5. DOI: 10.1145/342009.335388. URL: https://dl.acm.org/doi/10.1145/342009.335388 (visited on 11/13/2023) (cit. on pp. 14, 27).

Brockett, Patrick L., Richard A. Derrig, et al. (Sept. 2002). “Fraud Classification Using Principal Component Analysis of RIDITs”. In: *Journal of Risk and Insurance* 69.3, pp. 341–371. ISSN: 0022-4367, 1539-6975. DOI: 10.1111/1539-6975.00027. URL: https://onlinelibrary.wiley.com/doi/10.1111/1539-6975.00027 (visited on 10/23/2023) (cit. on p. 13).

Brockett, Patrick L., Xiaohua Xia, and Richard A. Derrig (June 1998). “Using Kohonen’s Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud”. In: *The Journal of Risk and Insurance* 65.2, p. 245. ISSN: 00224367. DOI: 10.2307/253535. JSTOR: 253535. URL: https://www.jstor.org/stable/253535?origin=crossref (visited on 10/23/2023) (cit. on pp. 9, 11).

Caudill, Steven B., Mercedes Ayuso, and Montserrat Guillén (Dec. 2005). “Fraud Detection Using a Multinomial Logit Model With Missing Information”. In: *Journal of Risk and Insurance* 72.4, pp. 539–550. ISSN: 0022-4367, 1539-6975. DOI: 10.1111/j.1539-6975.2005.00137.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6975.2005.00137.x (visited on 10/21/2023) (cit. on p. 13).

Comité Européen des Assurances (May 1996). “The European Insurance Anti-fraud Guide”. In: *CEA Info Special Issue* 4 (cit. on p. 7).

Debener, Jörn, Volker Heinke, and Johannes Kriebel (Sept. 2023). “Detecting Insurance Fraud Using Supervised and Unsupervised Machine Learning”. In: *Journal of Risk and Insurance* 90.3, pp. 743–768. ISSN: 0022-4367, 1539-6975. DOI: 10.1111/jori.12427. URL: https://onlinelibrary.wiley.com/doi/10.1111/jori.12427 (visited on 10/17/2023) (cit. on pp. 9, 10, 16, 17).

Derrig, Richard A. (Sept. 2002). "Insurance Fraud". In: *Journal of Risk and Insurance* 69.3, pp. 271–287. ISSN: 0022-4367, 1539-6975. DOI: 10.1111/1539-6975.00026. URL: https://onlinelibrary.wiley.com/doi/10.1111/1539-6975.00026 (visited on 10/23/2023) (cit. on p. 6).

Derrig, Richard A. and Krzysztof M. Ostaszewski (Sept. 1995). "Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification". In: *The Journal of Risk and Insurance* 62.3, p. 447. ISSN: 00224367. DOI: 10.2307/253819. JSTOR: 253819. URL: https://www.jstor.org/stable/253819?origin=crossref (visited on 10/23/2023) (cit. on pp. 10, 11).

Dionne, Georges and Claire Laberge-Nadeau, eds. (1999). *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation*. Red. by J. David Cummins. Vol. 20. Huebner International Series on Risk, Insurance, and Economic Security. Boston, MA: Springer US. ISBN: 978-1-4613-6817-5 978-1-4615-4058-8. DOI: 10.1007/978-1-4615-4058-8. URL: http://link.springer.com/10.1007/978-1-4615-4058-8 (visited on 08/09/2023) (cit. on p. 7).

Duffield, Grace M. and Peter N. Grabosky (2001). *The Psychology of Fraud*. Canberra: Australian Institute of Criminology. ISBN: 978-0-642-24224-2 (cit. on p. 5).

Duval, Francis, Jean-Philippe Boucher, and Mathieu Pigeon (June 2023). "Enhancing Claim Classification with Feature Extraction from Anomaly-detection-derived Routine and Peculiarity Profiles". In: *Journal of Risk and Insurance* 90.2, pp. 421–458. ISSN: 0022-4367, 1539-6975. DOI: 10.1111/jori.12418. URL: https://onlinelibrary.wiley.com/doi/10.1111/jori.12418 (visited on 11/07/2023) (cit. on p. 15).

Federal Bureau of Investigation (2023). *Insurance Fraud*. URL: https://www.fbi.gov/stats-services/publications/insurance-fraud (visited on 11/01/2023) (cit. on p. 6).

Gomes, Chamal, Zhuo Jin, and Hailiang Yang (Sept. 2021). "Insurance Fraud Detection with Unsupervised Deep Learning". In: *Journal of Risk and Insurance* 88.3, pp. 591–624. ISSN: 0022-4367, 1539-6975. DOI: 10.1111/jori.12359. URL: https://onlinelibrary.wiley.com/doi/10.1111/jori.12359 (visited on 10/23/2023) (cit. on pp. 14, 16, 17).

Han, Songqiao et al. (Sept. 16, 2022). *ADBench: Anomaly Detection Benchmark*. arXiv: 2206.09426 [cs]. URL: http://arxiv.org/abs/2206.09426 (visited on 12/28/2023). preprint (cit. on pp. 19, 20, 23).

Hassan, Amira Kamil Ibrahim and Ajith Abraham (2016). "Modeling Insurance Fraud Detection Using Imbalanced Data Classification". In: *Advances in Nature and Biologically Inspired Computing*. Ed. by Nelishia Pillay et al. Vol. 419. Cham: Springer International Publishing, pp. 117–127. ISBN: 978-3-319-27399-0 978-3-319-27400-3. DOI: 10.1007/978-3-319-27400-3_11. URL: http://link.springer.com/10.1007/978-3-319-27400-3_11 (visited on 10/23/2023) (cit. on p. 13).

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second edition, corrected at 12th printing 2017. Springer Series in Statistics. New York, NY: Springer. 745 pp. ISBN: 978-0-387-84857-0. DOI: 10.1007/b94608 (cit. on p. 16).

Insurance Europe (2019). *Insurance Fraud - Not a Victimless Crime*. URL: https://www.insuranceeurope.eu/publications/703/insurance-

fraud-not-a-victimless-crime/Insurance%20fraud%20-%20not%20a%20victimless%20crime.pdf (visited on 11/01/2023) (cit. on p. 7).
Insurance Information Institute (2023). *Background on: Insurance Fraud*. URL: https://www.iii.org/article/background-on-insurance-fraud (visited on 11/01/2023) (cit. on pp. 7, 30).
Jiang, Xiaoshan et al. (Aug. 17, 2021). "Medical Insurance Medication Anomaly Detection Based on Isolated Forest Proximity Matrix". In: *2021 16th International Conference on Computer Science & Education (ICCSE)*. 2021 16th International Conference on Computer Science & Education (ICCSE). Lancaster, United Kingdom: IEEE, pp. 512–517. ISBN: 978-1-66541-468-5. DOI: 10.1109/ICCSE51940.2021.9569723. URL: https://ieeexplore.ieee.org/document/9569723/ (visited on 10/20/2023) (cit. on p. 14).
Ken Dornstein (Jan. 1, 1996). "Accidentally, On Purpose: The Making of a Personal Injury Underworld in America". In: (cit. on p. 5).
Li, Jing et al. (Sept. 2008). "A Survey on Statistical Methods for Health Care Fraud Detection". In: *Health Care Management Science* 11.3, pp. 275–287. ISSN: 1386-9620, 1572-9389. DOI: 10.1007/s10729-007-9045-4. URL: http://link.springer.com/10.1007/s10729-007-9045-4 (visited on 11/09/2023) (cit. on p. 16).
Majhi, Santosh Kumar (Mar. 2021). "Fuzzy Clustering Algorithm Based on Modified Whale Optimization Algorithm for Automobile Insurance Fraud Detection". In: *Evolutionary Intelligence* 14.1, pp. 35–46. ISSN: 1864-5909, 1864-5917. DOI: 10.1007/s12065-019-00260-3. URL: http://link.springer.com/10.1007/s12065-019-00260-3 (visited on 10/23/2023) (cit. on p. 13).
Majhi, Santosh Kumar et al. (Mar. 26, 2019). "Fuzzy Clustering Using Salp Swarm Algorithm for Automobile Insurance Fraud Detection". In: *Journal of Intelligent & Fuzzy Systems* 36.3. Ed. by Sabu M. Thampi and El-Sayed M. El-Alfy, pp. 2333–2344. ISSN: 10641246, 18758967. DOI: 10.3233/JIFS-169944. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JIFS-169944 (visited on 11/07/2023) (cit. on p. 13).
Nagrecha, Saurabh, Reid A. Johnson, and Nitesh V. Chawla (Mar. 2018). "FraudBuster: Reducing Fraud in an Auto Insurance Market". In: *Big Data* 6.1, pp. 3–12. ISSN: 2167-6461, 2167-647X. DOI: 10.1089/big.2017.0083. URL: http://www.liebertpub.com/doi/10.1089/big.2017.0083 (visited on 10/23/2023) (cit. on p. 6).
Nian, Ke et al. (Mar. 2016). "Auto Insurance Fraud Detection Using Unsupervised Spectral Ranking for Anomaly". In: *The Journal of Finance and Data Science* 2.1, pp. 58–75. ISSN: 24059188. DOI: 10.1016/j.jfds.2016.03.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S2405918816300058 (visited on 06/02/2023) (cit. on p. 14).
Palacio, Sebastián M. (July 17, 2019). "Abnormal Pattern Prediction: Detecting Fraudulent Insurance Property Claims with Semi-Supervised Machine-Learning". In: *Data Science Journal* 18.1, p. 35. ISSN: 1683-1470. DOI: 10.5334/dsj-2019-035. URL: https://datascience.codata.org/article/10.5334/dsj-2019-035/ (visited on 10/29/2023) (cit. on p. 20).
Pérez, Jesús M. et al. (2005). "Consolidated Tree Classifier Learning in a Car Insurance Fraud Detection Domain with Class Imbalance". In: *Pattern Recognition and Data Mining*. Ed. by Sameer Singh et al. Red. by David Hutchison

et al. Vol. 3686. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 381–389. ISBN: 978-3-540-28757-5 978-3-540-28758-2. DOI: 10.1007/11551188_41. URL: http://link.springer.com/10.1007/11551188_41 (visited on 11/10/2023) (cit. on p. 13).

Perini, Lorenzo, Paul Buerkner, and Arto Klami (Oct. 17, 2023). *Estimating the Contamination Factor's Distribution in Unsupervised Anomaly Detection*. arXiv: 2210.10487 [cs, stat]. URL: http://arxiv.org/abs/2210.10487 (visited on 04/23/2024). preprint (cit. on p. 19).

Phua, Clifton, Damminda Alahakoon, and Vincent Lee (June 2004). "Minority Report in Fraud Detection: Classification of Skewed Data". In: *ACM SIGKDD Explorations Newsletter* 6.1, pp. 50–59. ISSN: 1931-0145, 1931-0153. DOI: 10.1145/1007730.1007738. URL: https://dl.acm.org/doi/10.1145/1007730.1007738 (visited on 11/07/2023) (cit. on p. 13).

Picard, Pierre (Dec. 1996). "Auditing Claims in the Insurance Market with Fraud: The Credibility Issue". In: *Journal of Public Economics* 63.1, pp. 27–56. ISSN: 00472727. DOI: 10.1016/0047-2727(95)01569-8. URL: https://linkinghub.elsevier.com/retrieve/pii/0047272795015698 (visited on 10/23/2023) (cit. on p. 8).

Shaeiri, Z and S.J. Kazemitabar (July 2020). "Fast Unsupervised Automobile Insurance Fraud Detection Based on Spectral Ranking of Anomalies". In: *International Journal of Engineering* 33.7. ISSN: 17281431, 17359244. DOI: 10.5829/ije.2020.33.07a.10. URL: http://www.ije.ir/article_108465.html (visited on 11/07/2023) (cit. on p. 14).

Steinbuss, Georg and Klemens Böhm (Aug. 31, 2021). "Benchmarking Unsupervised Outlier Detection with Realistic Synthetic Data". In: *ACM Transactions on Knowledge Discovery from Data* 15.4, pp. 1–20. ISSN: 1556-4681, 1556-472X. DOI: 10.1145/3441453. arXiv: 2004.06947 [cs, stat]. URL: http://arxiv.org/abs/2004.06947 (visited on 12/29/2023) (cit. on pp. 20, 23, 27).

Sternberg, M. and R.G. Reynolds (Nov./1997). "Using Cultural Algorithms to Support Re-Engineering of Rule-Based Expert Systems in Dynamic Performance Environments: A Case Study in Fraud Detection". In: *IEEE Transactions on Evolutionary Computation* 1.4, pp. 225–243. ISSN: 1089778X. DOI: 10.1109/4235.687883. URL: http://ieeexplore.ieee.org/document/687883/ (visited on 10/23/2023) (cit. on p. 11).

Stripling, Eugen et al. (July 2018). "Isolation-Based Conditional Anomaly Detection on Mixed-Attribute Data to Uncover Workers' Compensation Fraud". In: *Decision Support Systems* 111, pp. 13–26. ISSN: 01679236. DOI: 10.1016/j.dss.2018.04.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S016792361830068X (visited on 10/20/2023) (cit. on pp. 14, 17).

Šubelj, Lovro, Štefan Furlan, and Marko Bajec (Jan. 2011). "An Expert System for Detecting Automobile Insurance Fraud Using Social Network Analysis". In: *Expert Systems with Applications* 38.1, pp. 1039–1052. ISSN: 09574174. DOI: 10.1016/j.eswa.2010.07.143. arXiv: 1104.3904 [physics, stat]. URL: http://arxiv.org/abs/1104.3904 (visited on 10/23/2023) (cit. on p. 11).

Subudhi, Sharmila and Suvasini Panigrahi (2017). "Use of Optimized Fuzzy C-Means Clustering and Supervised Classifiers for Automobile Insurance Fraud Detection". In: *Journal of King Saud University - Computer and Informa-*

tion Sciences 32.5, pp. 568–575. ISSN: 13191578. DOI: 10.1016/j.jksuci.2017.09.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S1319157817301672 (visited on 11/07/2023) (cit. on p. 13).

Subudhi, Sharmila and Suvasini Panigrahi (Sept. 2018). "Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud". In: *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*. 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA). Changsha: IEEE, pp. 528–531. ISBN: 978-1-5386-8431-3. DOI: 10.1109/ICDSBA.2018.00104. URL: https://ieeexplore.ieee.org/document/8588973/ (visited on 10/23/2023) (cit. on p. 13).

Sundarkumar, G. Ganesh and Vadlamani Ravi (Jan. 2015). "A Novel Hybrid Undersampling Method for Mining Unbalanced Datasets in Banking and Insurance". In: *Engineering Applications of Artificial Intelligence* 37, pp. 368–377. ISSN: 09521976. DOI: 10.1016/j.engappai.2014.09.019. URL: https://linkinghub.elsevier.com/retrieve/pii/S0952197614002395 (visited on 06/02/2023) (cit. on p. 13).

Sundarkumar, G. Ganesh, Vadlamani Ravi, and V. Siddeshwar (Dec. 2015). "One-Class Support Vector Machine Based Undersampling: Application to Churn Prediction and Insurance Fraud Detection". In: *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). Madurai, India: IEEE, pp. 1–7. ISBN: 978-1-4799-7848-9 978-1-4799-7849-6. DOI: 10.1109/ICCIC.2015.7435726. URL: http://ieeexplore.ieee.org/document/7435726/ (visited on 10/20/2023) (cit. on p. 13).

Tennyson, Sharon and Pau Salsas-Forn (Sept. 2002). "Claims Auditing in Automobile Insurance: Fraud Detection and Deterrence Objectives". In: *Journal of Risk & Insurance* 69.3, pp. 289–308. ISSN: 0022-4367, 1539-6975. DOI: 10.1111/1539-6975.00024. URL: https://onlinelibrary.wiley.com/doi/10.1111/1539-6975.00024 (visited on 07/19/2023) (cit. on p. 8).

Tumminello, Michele et al. (June 2023). "Insurance Fraud Detection: A Statistically Validated Network Approach". In: *Journal of Risk and Insurance* 90.2, pp. 381–419. ISSN: 0022-4367, 1539-6975. DOI: 10.1111/jori.12415. URL: https://onlinelibrary.wiley.com/doi/10.1111/jori.12415 (visited on 10/20/2023) (cit. on p. 14).

Viaene, S, G Dedene, and R Derrig (Oct. 2005). "Auto Claim Fraud Detection Using Bayesian Learning Neural Networks". In: *Expert Systems with Applications* 29.3, pp. 653–666. ISSN: 09574174. DOI: 10.1016/j.eswa.2005.04.030. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417405000825 (visited on 10/23/2023) (cit. on p. 13).

Viaene, S., R.A. Derrig, and G. Dedene (May 2004). "A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis". In: *IEEE Transactions on Knowledge and Data Engineering* 16.5, pp. 612–620. ISSN: 1041-4347. DOI: 10.1109/TKDE.2004.1277822. URL: http://ieeexplore.ieee.org/document/1277822/ (visited on 11/07/2023) (cit. on p. 13).

Viaene, Stijn and Guido Dedene (Apr. 2004). "Insurance Fraud: Issues and Challenges". In: *The Geneva Papers on Risk and Insurance - Issues and Practice* 29.2, pp. 313–333. ISSN: 1018-5895, 1468-0440. DOI: 10.1111/j.1468-0440.2004.00290.x. URL: http://link.springer.com/10.1111/j.1468-0440.2004.00290.x (visited on 08/09/2023) (cit. on pp. 4–9).

Viaene, Stijn, Richard A. Derrig, et al. (Sept. 2002). "A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection". In: *Journal of Risk and Insurance* 69.3, pp. 373–421. ISSN: 0022-4367, 1539-6975. DOI: 10.1111/1539-6975.00023. URL: https://onlinelibrary.wiley.com/doi/10.1111/1539-6975.00023 (visited on 11/07/2023) (cit. on p. 13).

Viaene, Stijn, Stijn Viaene, et al. (Jan. 1, 2007). "Strategies for Detecting Fraudulent Claims in the Automobile Insurance Industry". In: *European Journal of Operational Research* 176.1, pp. 565–583. DOI: 10.1016/j.ejor.2005.08.005 (cit. on pp. 7, 13, 17).

Vosseler, Alexander (June 21, 2022). "Unsupervised Insurance Fraud Prediction Based on Anomaly Detector Ensembles". In: *Risks* 10.7, p. 132. ISSN: 2227-9091. DOI: 10.3390/risks10070132. URL: https://www.mdpi.com/2227-9091/10/7/132 (visited on 10/20/2023) (cit. on p. 14).

Weisberg, Herbert I and Richard A Derrig (1998). "Quantitative Methods For Detecting Fraudulent Automobile Bodily Injury Claims". In: (cit. on pp. 7, 10).

— (1991). "Fraud and Automobile Insurance: A Report on Bodily Injury Liability Claims in Massachusetts". In: *Journal of Insurance Regulation* 9.4, pp. 497–541 (cit. on p. 10).

Yan, Chun and Yaqi Li (Sept. 2015). "The Identification Algorithm and Model Construction of Automobile Insurance Fraud Based on Data Mining". In: *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*. 2015 Fifth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC). Qinhuangdao, China: IEEE, pp. 1922–1928. ISBN: 978-1-4673-7723-2. DOI: 10.1109/IMCCC.2015.408. URL: http://ieeexplore.ieee.org/document/7406190/ (visited on 10/23/2023) (cit. on p. 14).

Yan, Chun, Yaqi Li, et al. (June 2020). "An Artificial Bee Colony-Based Kernel Ridge Regression for Automobile Insurance Fraud Identification". In: *Neurocomputing* 393, pp. 115–125. ISSN: 09252312. DOI: 10.1016/j.neucom.2017.12.072. URL: https://linkinghub.elsevier.com/retrieve/pii/S0925231219310550 (visited on 11/07/2023) (cit. on p. 14).

Zelenkov, Yuri (Nov. 2019). "Example-Dependent Cost-Sensitive Adaptive Boosting". In: *Expert Systems with Applications* 135, pp. 71–82. ISSN: 09574174. DOI: 10.1016/j.eswa.2019.06.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417419304099 (visited on 10/23/2023) (cit. on p. 14).

Zhao, Yue, Zain Nasrullah, and Zheng Li (2019). "PyOD: A Python Toolbox for Scalable Outlier Detection". In: *Journal of Machine Learning Research* 20.96, pp. 1–7. URL: http://jmlr.org/papers/v20/19-011.html (cit. on p. 19).

Emacs 29.3 (Org mode 9.7)