Anomaly detection

Anomaly detection techniques in fraud detection, performance optimization, and data quality

Bjorn Blom, MSc
Jan Thiemen Postema, MSc
Rens IJsendijk, MSc
Judith Houtepen, MBA, AAG
Job Prince, MSc, FRM

Milliman

Introduction

Health insurance fraud is a large and growing problem throughout Europe. According to a 2011 report by SAS, a total of EUR 56 billion was spent on healthcare fraud in the region.[1] Even though hard numbers are not available, the Dutch court of audit estimated in 2022 that in the Netherlands a couple billion euros are spent on fraudulent healthcare claims each year and, in 2018, the Ministry of Internal Affairs in Bayern estimated those costs to be about EUR 14 billion a year in its state.[2,3] This excess cost places a burden on both consumers and the healthcare system as a whole, but there are methods technologists and others can use to detect fraud and optimize performance. Anomaly detection is a technique that tries to find the odd duck in the data set, making it incredibly useful in detecting unwanted cases.

Anomaly detection is one of the oldest research areas in the field of statistics, and the techniques developed have been applied in a wide variety of use cases. Virtually every real-life data set could contain anomalies, which makes these methods one of the fundamentals in the tool set of every data practitioner. The wealth of knowledge that has been gained over the years by data professionals can be applied to new business situations.

Our Milliman colleagues have implemented anomaly detection techniques in different fields, such as health insurance and retail, and in several areas such as fraud detection, performance optimization, and improving data quality. In this paper, we explore these areas by looking at some use cases and provide more insight into the different anomaly detection techniques.

Applications of outlier detection

There are numerous application areas for outlier detection. Within this paper we focus on the following three, which, in our opinion, could have a big impact in the insurance sector:

1. Fraud detection
2. Performance optimization
3. Data quality

This section gives an overview of each of these objectives and shortly describes some methods that could be used for finding outliers.

[1] SAS (May 16, 2011). Future Bright Fighting Fraud. Retrieved March 3, 2023, from https://issuu.com/sas-institue/docs/future-bright-fighting-fraud.

[2] Rekenkamer (April 14, 2022). Aanpak zorgfraude is vooral vergaderen. Retrieved March 3, 2023 from https://www.rekenkamer.nl/actueel/nieuws/2022/04/14/aanpak-zorgfraude-is-vooral-vergaderen.

[3] Bavarian State Ministry of the Interior (March 27, 2018). Betrug im Gesundheitswesen. Retrieved March 3, 2023, from https://www.stmi.bayern.de/med/pressemitteilungen/pressearchiv/2018/100b/index.php.

A. FRAUD DETECTION

Fraud detection is widely used as an application area for outlier detection. With respect to insurance companies, it mainly concerns healthcare and non-life insurance providers. The objective of outlier detection methods is to detect either fraudulent individual claims or to detect third-party claim handlers that submit fraudulent claims, such as car repair shops that submit excessive repair costs.

Fraud detection is not limited to the insurance sector. Credit card fraud, telecommunications, online auctions, and smart meter data fraud are common areas for which fraud detection also plays an important role.[4] The methods that are used in these different sectors are not exclusively applicable to a single sector. Over the years, many methods for fraud detection have been proposed. These methods can be roughly separated into three categories: parametric, statistical, and machine learning.

Parametric methods are methods that assume an a priori distribution of the data in such a way that data points outside a predetermined interval are considered outliers. Statistical methodologies do not necessarily assume an a priori distribution and can be fitted to data. They might use distance-based methods to determine outliers, such as Cook's distance. Advantages of these kinds of methods are their intuitiveness and explainable characteristics.

Machine learning methods can be very powerful tools in the detection of fraud. These methods are strong in mitigating bias in the data, which can result in a more optimal classification of fraudulent data points. However, using these methods also comes with a risk when a method is not correctly implemented or the results that are produced are not properly understood. The latter risk is often referred to as creating a black-box model. There are substantial risks to the end user in the form of exposure to litigation by unjustified classification of individuals as fraudulent. Therefore, when using such models, keeping a human-in-the-loop and properly explaining the results—for example, by using explainable artificial intelligence (AI) techniques—remain key.

In the Netherlands, such a scenario has occurred in the so-called fiscal benefit affair (the "Toeslagenaffaire"), which led to the resignation of the cabinet in 2021. In this affair, a model was used that was trained on protected data such as someone's Dutch citizenship status and the age of their "burgerservicenummer," an indication of the time a person has lived in the Netherlands.[5] The outcomes of this model were then used to decide whose applications should be manually investigated. The usage of such potentially discriminatory variables led to a fine by the Dutch Data Protection Authority.[6]

Outlier detection techniques are widely used in the financial sector to detect fraud, albeit with a human-in-the-loop. Statistical and parametric methods show that simpler methods can also provide aid in detecting fraud. Machine learning methods show great promise when used for fraud detection, but they must be well understood and they require experts for correct implementation.

B. PERFORMANCE OPTIMIZATION

An outlier in data or an outlier in the predictions of an algorithm can be an indicator of underperformance for the business. When the underperformance is identified, it can be optimized. The optimization itself is of course dependent on the business. For example, in a claim-handling process at an insurer, costs could be saved by prioritizing claims based on the "uniqueness" of a claim, as they are likely to be more time-consuming or could, based on the characteristics, directly be referred to a claim handler with the relevant expertise. Another example is in the retail sector, where outlier detection can be used to focus an account manager's attention on underperforming stores. For a more extensive example of this application, see also the retailer use case at the end of this report. By extension, outlier detection in the financial sector can be used to identify underperforming or over-performing products.

---

[4] Omar, Sinayobye & Kiwanuka, Fred & Swaib, Kaawaase (2018). A state-of-the-art review of machine learning techniques for fraud detection research. 11-19. 10.1145/3195528.3195534.

[5] Tweede Kamer der Staten-Generaal (December 17, 2020). Ongekend onrecht: Verslag – Parlementaire ondervragingscommissie Kinderopvangtoeslag. Retrieved March 3, 2023, from https://www.tweedekamer.nl/sites/default/files/atoms/files/20201217_eindverslag_parlementaire_ondervragingscommissie_kinderopvangtoeslag.pdf.

[6] Autoriteit Persoonsgegevens. Belastingdienst/Toeslagen: De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag. Retrieved March 3, 2023, from https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf.

In a broader sense, every aspect of the business that is directly related to the quantitative performance measure can potentially be optimized using outlier detection. Due to the nature of performance outliers, these data points should be examined to decide what actions to take. They can range from actions such as sending an account manager over, to more complex situations where outliers can have a significant impact on business performance, to decisions about underperformance or over-performance of certain business units.

C. DATA QUALITY

The quality of data plays an important role in the performance of models. Outliers in the data can greatly affect the performance of predictive models and in turn result in unreliable model outcomes. Finding these outliers can help to better understand and improve model performance. Because an outlier is not necessarily an erroneous data point, not all outliers need to be removed. It is possible that extreme values in data play an important role in understanding the data. These points thus need to be identified at an exploratory stage.

Depending on the purpose of the model, outliers might need to be removed. But this requires careful consideration as, for example, removing too many data points before the model is estimated can cause overfitting and thus produce undesirable results. Also, the need to remove outliers from a data set depends on the type of model that is used. Generally, models such as tree-based models and regression and parametric models do show a greater dependence on the presence of outliers.

Therefore, when building a model, whether it is a statistical or machine learning model, it is important to be aware of any outliers in the data and to have an expert decide whether they are erroneous and ought to be removed. This is true for every model, including, e.g., those used for financial reporting in insurance companies. It is not for naught that regulatory frameworks such as Solvency II require extensive data quality controls. For example, our colleagues explored the usage of outlier detection to improve the goodness-of-fit of a least squares Monte Carlo (LSMC) proxy model for the calculation of the best estimate liabilities (BEL) for a life insurance company by cleaning the data of unreasonable net asset values (NAVs) produced by the cash flow model,[7] which showed very promising results.

Techniques

The terms “outliers” and “anomalies” can be used largely interchangeably, and the literature generally does not make a clear distinction between the two. However, colloquially, practitioners often refer to outliers as points that ought to be removed because they are likely faults in the data. On the other hand, anomaly often refers to a broader set of data points that can include outliers but also other points that are not ”normal,” including, for example, fraud cases. This is also the definition that we will use in the remainder of this section.

In his 1980s book on the identification of outliers, Douglas Hawkins described an outlier as ”an observation which **deviates so much from the other observations** as to arouse suspicions that it was **generated by a different mechanism**.” It therefore follows that, if we want to detect an outlier, we need to have a benchmark with which we can compare our observations. Over the years, a plethora of methods have been proposed to generate such a benchmark. In this section, we describe a few of these techniques, as well as the main criteria by which they can be categorized. In Figure 1, a more extensive overview of methods is given. These methods in the table are ranked on several criteria. This list is, however, by no means complete; others might include speed, consistency, flexibility, and scalability. The most important to consider fully depends on the use case and the business environment in which it is implemented.

[7] Chaudhry, A., Dudceac M., & Leitschkis, M. (December 2021). Machine learning approaches to outlier detection. Milliman White Paper. Retrieved March 3, 2023, from https://nl.milliman.com/nl-NL/insight/machine-learning-approaches-to-outlier-detection.

FIGURE 1: OVERVIEW OF ANOMALY DETECTION METHODS AND CHARACTERISTICS

	METHOD TYPE		MACHINE LEARNING TYPE		
	STATISTICAL	MACHINE LEARNING	SUPERVISED	UNSUPERVISED	EXPLAINABILITY
Cook's distance	✓				High
Tukey's method	✓				High
Benfords law	✓				High
Isolation forest		✓		✓	Low
kNN		✓		✓	Medium
K-means clustering		✓		✓	Medium
Artificial neural network		✓	✓		Low
Naive Bayes		✓	✓		High
Decision tree		✓	✓		Medium
Support vector machines		✓	✓		Medium
One-class support vector machine		✓		✓	Medium
Forecasting techniques	✓	✓	✓		Medium

OUTLIER DETECTION

Because outlier detection, in the sense that it was defined in the introduction above, is an important use case of anomaly detection, it will be the focus of this section. More specifically, we look at unsupervised methods that find outliers by benchmarking them against the other points in the data set.

Traditionally, outlier detection has been seen as a statistical problem. In fact, it is one of the oldest research areas in the field of statistics.[8] More recently with the advent of bigger data sets, it has also gained more interest from the machine learning (ML) community. This section starts by looking at methods originating in the realm of statistics, before moving on to discuss ML models.

Statistical methods

A very simple, yet effective, method for finding outliers is the Tukey criterion. This rule states that a data point is an outlier when it is 1.5 times the interquartile distance under the first quartile or over the third quartile. Due to its simplicity, this method is very effective in spotting an outlier over one axis. However, when the data set becomes multidimensional, the Tukey criterion is less effective.

Another very intuitive method to find outliers is Benford's law, according to which the various leading digits in a data set do not occur uniformly. Instead, the leading digit is more likely to be small: the number 1 should occur in about 30% of cases, whereas the number 9 should only occur in about 5%. This law is often shown to apply to a large variety of cases and is often applied to spot anomalies in financial accounting, among other uses. This law was also successfully implemented by our colleagues at a Dutch health insurance company. In this case, it was especially suitable because of its simplicity and, thereby, had wide applicability. This also meant that no additional privacy-

[8] Hawkins, D. M. *Identification of Outliers*. Springer Dordrecht; 2013. See https://link.springer.com/book/10.1007/978-94-015-3994-4#.

sensitive data was required and there was no risk of any unintentional bias. This use case is described shortly in the next section and more extensively in a dedicated Milliman report.[9]

A common way to find outliers is to create a model of reality and then find points that do not adhere to this model. A notable statistical method that could be used for outlier detection and uses this strategy is Cook's distance. This technique is often used in regression analysis to find the most influential data points. However, when a point is very influential, it most likely also has different characteristics than the other data points in the data set. Therefore, a high Cook's distance also indicates a high level of "outlierness." One of the main advantages of using this metric for finding outliers is that, when using a regression model, there is little extra work required to implement it. However, this is simultaneously also one of the main drawbacks: it requires fitting a regression model to the data. When such a model is not a good fit, Cook's distance is not a good indicator either. Another drawback is that the distance is merely an indicator. It still requires a lot of expert judgment to set a threshold above which the data points are to be considered an outlier. This method is described in more detail in another paper by our colleagues.[10]

One of the benefits that all models mentioned in this section share is that they are inherently explainable. All three involve an intuitive theory that is easy to understand and explain, which makes them very suitable when explainability is a firm requirement. However, these methods lack the wide applicability and scalability that machine learning models might offer, especially when it comes to the number of dimensions in the data.

Purpose-built machine learning models

When we are looking at ML-based methods for outlier detection, we can again make a split, this time between methods that are purpose-built for outlier detection and "general" methods that can be adapted for this purpose. Isolation forests and local outlier factors (LOF) are examples of techniques that fall in the purpose-built group. Usually, these methods compare each data point against the other data points in the set. An LOF, for example, tries to find the k-nearest neighbors (kNN) for each data point. Then, when for some points the distance to its nearest neighbors is substantially larger than among most other points in the data set, those points are considered outliers. A similar strategy can be used when doing k-means clustering. The relative simplicity of this approach makes it easy to understand the model; however, how explainable the results are depends largely on the dimensionality of the input data. When the data is two-dimensional, it is easy to visualize it and understand what is happening. When more than three dimensions are involved, it gets substantially harder. Techniques such as an isolation forest are based on the same principle of finding data points that do not fit in with the rest of the data. An isolation forest is, however, substantially more difficult to explain, even though there are great post hoc techniques available, such as SHapley Additive exPlanations (SHAP) values.

General machine learning models

Then there is the other category of ML methods, which are not purpose-built for outlier detection, but can be adapted to serve this purpose. When we return to the definition, we read that an outlier is "generated by a different mechanism." When we use one of these methods for outlier detection, we aim to approximate the mechanism that generated the data. For example, one could fit a decision tree on a data set and then use that trained decision tree to give a prediction for all data points in the original data set. When the model is trained properly and there is no overfitting, we expect the prediction for the outlier to differ substantially from the actual target, as the model should not be able to generalize to this point. Other methods in this category, such as neural networks, follow the same principle.

We could say that all of these methods are easy to explain. After all, you compare the output of the model with the observation, which is an easy-to-understand and -explain process. However, when the user also wants to know how the output was generated, it becomes more difficult. In some cases—for example, when using a decision tree—it is possible to extract information about feature importance from the model. However, when using more black-box models, one might need to revert to post hoc techniques for explainable AI, such as SHAP values.

Time series models

[9] Carolissen, K. & Kacal, M. (June 2020). Anomaly analysis and detection in health insurance. Milliman White Paper. Retrieved March 3, 2023, from https://us.milliman.com/-/media/milliman/pdfs/articles/anonomaly-analysis.ashx.

[10] Chaudhry, A., Dudceac, M., & Leitschkis, M. (December 2021), op cit.

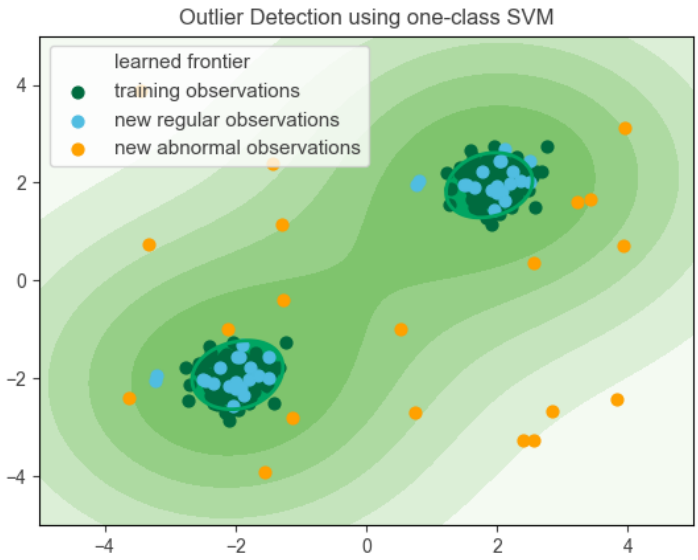
Outlier detection methods for time series data can be considered as a category in their own right, and feature both statistical and ML methods. There are many techniques in this realm, most of which again focus on building a model of reality and then testing the observations against that model. One popular way to do so is by building a time series forecasting method, e.g., an autoregressive integrated moving average (ARIMA) model or a long short-term memory (LSTM) neural network, and then comparing the (new) observations against the predictions of the model. When a data point is distant from the prediction, it is again likely to be an outlier.

ANOMALY DETECTION

So far, we have looked primarily at detecting outliers. However, there are also other types of anomalies, anomalies that we might ”expect.” In those cases, we are likely to have a labeled data set of anomalies, which need to be detected when new observations come in. This problem can also be described as a two-class classification problem on highly imbalanced data (after all, by definition, an anomaly does not occur as often as a regular observation). Almost all classification techniques can be used to tackle this problem. However, some are better equipped to handle imbalanced data out of the box than others.

A one-class support vector machine (SVM), for example, is quite suitable for this kind of problem. Instead of attempting to properly distinguish between the two classes, it tries to find a boundary that describes the majority class. All observations outside of this boundary are then assigned to the minority group (the anomalies).

FIGURE 2: AN EXAMPLE OF A ONE-CLASS SVM



With some work, other classification methods can also be adapted to deal with this kind of data. For example, by over-sampling the minority, or under-sampling the majority classes, the problem can be mitigated. This is the most universal strategy for dealing with imbalanced data and can be used with virtually every classification technique, including, for example, naive Bayes, decision trees (and any of their variations such as gradient boosted trees), neural networks, and multi-class SVMs. Another approach would be to give more weight to observations in the minority class. This technique can be used in some models, such as neural networks. It makes the model pay more attention to examples with higher weights by penalizing wrong predictions for those data points more in the loss function. In other words, a wrong prediction on an example with weight 2 would be twice as bad as a wrong prediction on an example with weight 1.

Because the types of models that can be used to perform anomaly detection are virtually endless, it is nearly impossible to make a generally applicable remark about their explainability. However, one added difficulty when dealing with this problem is the severe class imbalance, which might make it substantially harder to give a meaningful explanation. In the ”explainable AI” use case that’s showcased below, one possible solution to this problem is presented.

Use cases

Several applications and techniques have been explained in the previous sections. Three use cases where anomaly detection was used are described in this section. The first use case uses anomaly detection to improve the performance of the scheduled check visits in stores, which is in the application area of performance optimization. In the last two use cases, anomaly detection is used for fraud detection.

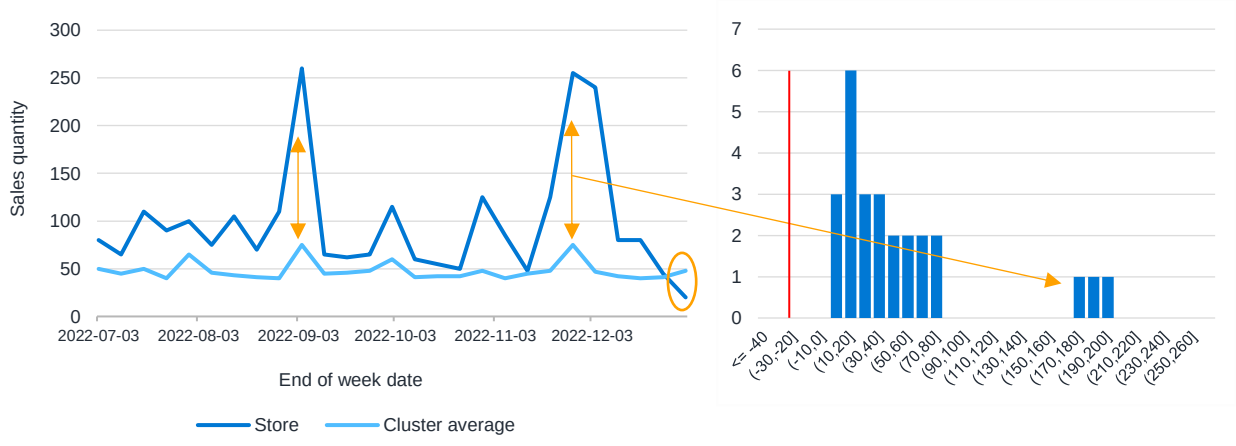
RETAILER USE CASE

Sales representatives of a large retailer regularly visit stores where products are sold. During those visits, they perform checks on multiple products, including inventory, price labels, and positioning. The current scheduling of sales representatives visiting stores is considered suboptimal. By using a more data-driven approach, the scheduling of store visits can be improved. The goal of this project was to make sales representatives more efficient in their visits to stores by detecting more anomalies, such that there is a higher likelihood of finding irregularities in the stores (e.g., unavailable inventory for a product). These irregularities can be remedied during these visits, with the ultimate aim to increase sales.

We have helped developing a model to find those products with a two-step method. It starts by clustering together stores that behaved similarly in the past. This clustering is performed for every single product separately. Then we compare the behavior of each store over the last period, relative to the other stores in the cluster, to its behavior in the latest week. Based on the deltas between the average of the cluster and the store’s sales, a probability density function can be estimated. Based on this probability density function, the likelihood of the last week’s occurrence can be calculated. If this behavior is not as expected, we consider it an outlier.

In Figure 3, an illustrative example is given about step 2 of the two-step method. In the graph on the left, the sales of this store are compared with the average of the cluster. The graph on the right is showing a frequency histogram of the differences between the sales quantity of the store compared to the average of the cluster. Based on this delta frequency histogram, the outlier detection method is performed: How unlikely the last occurrence was to happen compared to the deltas in the last period, with the last week’s occurrence delta, is shown with a red line.

FIGURE 3: ILLUSTRATIVE EXAMPLE OF ONE STORE-PRODUCT COMBINATION THAT IS UNDERPERFORMING IN THE LAST WEEK, SALES QUANTITY OVER TIME STORE VERSUS CLUSTER AVERAGE AND DELTA FREQUENCY HISTOGRAM



This whole process is performed on store-product combinations, i.e., the sales of product A in store X are compared to the sales of product A in the other stores in the market.

The first results of this model were promising. An uplift in sales has been detected in stores where visits were performed, based on the outcome of this model.

HEALTH FRAUD DETECTION BY INSURER

Healthcare fraud is considered a material risk in the Netherlands. According to the Dutch Court of Audit, several billions of euros are lost each year due to healthcare fraud in the Netherlands.[11] Tackling healthcare fraud became one of the priorities of the last coalition agreement. In the Netherlands, healthcare providers bill their services directly to the customer's health insurer. Currently it is very easy to start a new healthcare provider company, even without a medical degree or patients. According to a Dutch health insurer, this results in a lot of fraudulent companies and claims.[12] Dutch insurers would like to be able to detect fraudulent claims more efficiently. This became even more relevant after the Dutch Minister of Health confirmed that the number of fraudulent investigations doubled from 2020 to 2021.[13]

In this research, the goal was to limit the number of healthcare companies that need to be checked by domain experts on fraudulent claims based on the outcome of different models. In this particular case, it was estimated that 2% of the claim data contains irregularities and thus possible fraudulent claims. Milliman used two models to detect possible fraudulent claims and out of 5,000 health companies created a ranked list of 50 companies that were most likely sending fraudulent claims.

The two models used to rank these companies are Benford's law and isolation forest. After extensive investigation of the ranked companies by the domain experts, nine out of the 50 companies from the isolation forest model and two out of the 50 from the Benford's law model were considered fraudulent.

More information on this research can be found in our Milliman paper.[14]

EXPLAINABLE AI

When basing a decision on the outcome of an artificial intelligence model, it is important to understand how that outcome came to be. This explains why, in recent years, explainable AI has seen a substantial rise in popularity. One area where this is especially useful is in fraud detection. In this use case, there is usually a human-in-the-loop, who checks all of the cases the model has marked as fraudulent. To help this person decide whether the case is indeed fraudulent, it can be incredibly useful to have an explanation on how the model came to its decision. However, because the model has marked all the cases the person sees as fraudulent, all explanations are likely to be similar and might not be that useful after all.

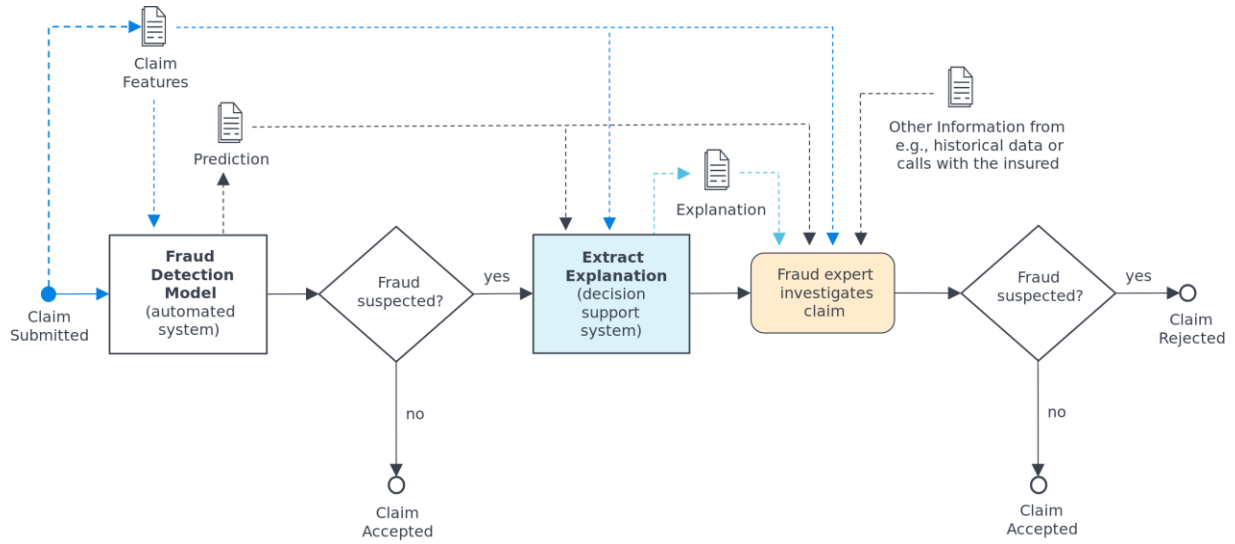
[11] Rekenkamer (April 14, 2022). Aanpak zorgfraude is vooral vergaderen, p. 26. Retrieved March 3, 2023, from https://www.rekenkamer.nl/actueel/nieuws/2022/04/14/aanpak-zorgfraude-is-vooral-vergaderen.

[12] NOS (January 2, 2020). Controle zorgfraude faalt, OM en verzekeraars luiden noodklok. Accessed at https://nos.nl/nieuwsuur/artikel/2317099-controle-zorgfraude-faalt-om-en-verzekeraars-luiden-noodklok

[13] AD. (2022, 29 November). Aantal fraudedossiers in de zorg is verdubbeld: 'Sommigen maken tonnen winst, krankzinnig'. Retrieved March 3, 2023, from https://www.ad.nl/politiek/aantal-fraudedossiers-in-de-zorg-is-verdubbeld-sommigen-maken-tonnen-winst-krankzinnig~aca9d264/.

[14] Carolissen, K. & Kacal, M. (June 2020), op cit.

FIGURE 4: EXAMPLE OF A FRAUD DETECTION PROCESS WITH A HUMAN-IN-THE-LOOP



One alternative is to train a meta-learning model on just the cases that were flagged as fraud (i.e., it is trained to distinguish between true results and false positives). When explaining the outcomes of this meta-learning model, the explanations are not ”contaminated” by the non-fraudulent cases. Our colleagues explored this method at a Dutch health insurance company and wrote an article about it.[15]

Conclusion

With the increasing amount of data available in insurance, retail, and other fields, it becomes desirable to apply anomaly detection techniques. Methods to detect anomalies can be used to detect fraudulent claims in insurance, particularly in product types that typically have a large frequency of payments, such as healthcare. Anomaly detection also can be used to optimize performance, as seen in the retailer use case, or to improve the data quality. With all these application areas, the ability to apply different methods and to understand their limitations is increasingly important. Being able to explain the results of the outcomes from the method is also critical, because it eases the deployment and adoption of anomaly detection methods in the business.

The number of different techniques has increased in recent years and will continue to increase. This field of data science is still in evolution and will be used more and more in the future as we see an ever-increasing amount of data.

[15] Zitouni, I., Postema, J.T., Sznajder D., & van Es, R. (December 22, 2022). Explainable AI in fraud detection. Milliman Insight. Retrieved March 3, 2023, from https://www.milliman.com/en/insight/Explainable-AI-in-fraud-detection.

Milliman

Milliman is among the world's largest providers of actuarial, risk management, and technology solutions. Our consulting and advanced analytics capabilities encompass healthcare, property & casualty insurance, life insurance and financial services, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT

Bjorn Blom
bjorn.blom@milliman.com

Jan Thiemen Postema
janthiemen.postema@milliman.com

Rens IJsendijk
rens.ijsendijk@milliman.com

Judith Houtepen
judith.houtepen@milliman.com

Job Prince
job.prince@milliman.com