EXPLAINABLE UNSUPERVISED ANOMALY DETECTION FOR HEALTHCARE INSURANCE DATA

Hannes De Meulemeester[1], Frank De Smet[2, 3], Johan Van Dorst[3], Elise Derroitte[3], Bart De Moor[1]
[1] ESAT-STADIUS, KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{name.lastname}@esat.kuleuven.be
[2] Department of Public Health and Primary Care, Environment and Health, 3000 Leuven, Belgium
[3] Christian Health Insurance Fund, 1031 Brussels, Belgium
{name.lastname}@cm.be

ABSTRACT

Background: Waste and fraud are important problems for health insurers to deal with. With the advent of big data, these insurers are looking more and more towards data mining and machine learning methods to help in detecting waste and fraud. However, labeled data is costly and difficult to acquire as it requires expert investigators and known care providers with atypical behavior.

Methods: In this work we show how recent advances in machine learning can be used to set up a workflow that can aid investigators in discovering practitioners or groups of practitioners with unusual resource use in order to more efficiently combat waste and fraud. We combine three different techniques, which have not been used in the context of healthcare insurance anomaly detection: categorical embeddings to deal with high-cardinality categorical variables, state-of-the-art unsupervised anomaly detection techniques to detect anomalies and Shapley additive explanations (SHAP) to explain the model output.

Results: The method has been evaluated on providers with a known anomalous profile and with the help of experts of the largest health insurance fund in Belgium. The quantitative experiments show that categorical embeddings offer a significant improvement compared to standard methods and that the state-of-the-art unsupervised anomaly detection techniques generally show an improvement over traditional methods. In a practical setting, the proposed workflow with SHAP was able to detect a previously unknown, anomalous trend among general practitioners.

Conclusions: The proposed workflow is able to detect known care providers with atypical behaviour and helps expert investigators in making informed decisions concerning possible fraud or overconsumption in the health insurance field.

***K*eywords** Health Insurance · Anomaly Detection · Unsupervised Machine Learning

Background

Healthcare and, in turn, health insurance providers are a vital pillar of a modern welfare state. However, healthcare comes at an ever increasing cost. The European Union currently spends approximately 1.5 trillion euro yearly on healthcare which, on average, corresponds to 10.9% of GDP per country in 2020. [18]. In the United States, healthcare spending reached 4.3 trillion dollar [11] which is 18.3% of its GDP. The increased standard of care that is present in modern society means having access to the latest technological advances in medicine and care which comes at a premium [29].

It is no wonder that all stakeholders involved are looking to find areas of improvement such that healthcare spending can be decreased without compromising on quality of care. One of such areas is the detection of waste and fraud. In general, waste or fraud are almost never considered on its own. But part as a bigger whole containing waste, errors, abuse, fraud and corruption [17]. The main difference between these types of infringements is the intention

and degree of severity. Only in the case of abuse, fraud and corruption can it be said that there is an intent to do wrong by finding loopholes or stretching a rule (abuse) or by actively and intentionally breaking the law (fraud and corruption). In the case of errors, a rule is unintentionally broken. Similar to this is waste or overconsumption, which includes care that is unnecessary or needlessly expensive. When simply looking at the billing data, as available to healthcare insurers, it can be difficult to differentiate between these types of infringements and care providers with valid resource use. Therefore, it is common to either group the categories together and look for general anomalies. Alternatively, specific, thematic, investigations can be performed to investigate known types of fraud.

In this work, we focus on the perspective of health insurers. Health insurance funds in Belgium are co-managers of the health insurance system and have a vested interest in managing this system as optimal as possible. Additionally, these health insurers are the closest to the source as they are the first to receive detailed billing data, making them the first line of defense.

However, waste and fraud are one of the most difficult problems for health insurers to deal with. Therefore, they are looking for new ways to help them with this task. With the advent of big data, these insurers are looking more and more towards data mining and machine learning methods to help in detecting anomalies in resource use data. Investigating waste or fraud in healthcare practitioners in principle requires a business specialist to go over the resource use records of all the care providers and to subsequently interview the practitioners with a potential anomalous profile. Such manual investigations are costly, inefficient and require a lot of time without tools to guide the investigator by pre-selecting the care providers with a profile that could have a higher risk of inappropriate resource use. Such investigations are also required to acquire labeled data if one wants to apply supervised machine learning techniques. This need for expert investigators to label the training data makes supervised methods more difficult. Additionally, supervised problems are limited in that only previously known types of waste and fraud are considered. As a result, the labeled data can contain additional false negatives with respect to inappropriate resource use in general. Finally, it is usually desirable to be able to identify new types of anomalies that warrant further investigation.

This research has been performed at the Christian Health Insurance Fund (CM) [38]. The CM, which has over 4.5 million members, is the largest Belgian mutual health insurer and currently processes over 400 million expenditure records every year.

Contribution This work shows how a number of recent advances in machine learning can be used to improve anomaly detection on health insurance resource use data and to help healthcare experts create informed decisions about possible fraud or overconsumption in the field. We combine three different state-of-the-art techniques, which have not been used in the context of healthcare insurance anomaly detection previously, to create a novel workflow. Categorical embeddings [14] are employed to deal with the large number of categorical features and values. These embeddings allow for a more efficient representation of the data, which in turn speeds up machine learning methods and significantly increases their performance. A large variety of unsupervised anomaly detection techniques, each with their own unique advantages, are used during the experiments. We consider both classic methods, which have been a staple in the anomaly detection field, and novel methods, which have not been employed in the context of health insurance data. Finally, we show illustrate how Shapley additive explanations (SHAP)[28] can be used to help experts identify anomalies and novel patterns in the data.

Related Work

Advanced statistical, data mining and machine learning techniques are a mainstay for fraud detection in many different areas [31, 3]. However, health care insurers are slower to implement these advanced types of techniques due to complex regulations, large decentralized data systems and high costs of commercial fraud detection packages[41] with questionable or at least unproven efficacy. Further, preparing the data and labeled data for machine learning is costly and manually intensive. Finally, investigating anomalies and subsequently, if necessary, deal with care providers with inappropriate resource use is a difficult and long process. Notwithstanding the slow integration, there has been some research into the field of unsupervised fraud detection in healthcare.

Chandola et al.[12] showed how text mining, social network analysis and temporal analysis can be applied to big healthcare data. For text mining, they performed topic modeling using Latent dirichlet allocation and determined that certain diagnosis topics were more likely to be related to fraudulent claims. Concerning social network analysis, graphs were created with the providers as nodes. They found that another fraudulent provider was present in 40% of the cases within 2-hops from a known fraudulent provider. Their temporal analysis showed that anomalies could be detected using temporal features. They also showed that features of a temporal nature are useful as input to supervised models such as logistic regression classifiers.

Thornton et al.[39] used k-means clustering, using custom metrics to perform anomaly analysis. The metrics were developed in cooperation with experts and incorporated features such as visit length, patient retention, referral rate and visit frequency. The domain knowledge used is common in their research and they conclude that this expertise is vital in order to develop healthcare fraud detection techniques[40].

Van Capelleveen et al. [42] performed experiments on dentist Medicaid data and showed that previously unknown patterns can be detected using simple unsupervised methods such as linear models and k-means clustering. The authors worked closely together with field experts and concluded that their methods could be beneficial for a decision support.

Bauder et al. [7] show that it is possible to classify health care professionals into their respective fields based on billing data . The practitioners for which the models have difficulty classifying them are flagged as outlying. The authors also developed a novel univariate probabilistic programming anomaly detection method which provided interpretability on the feature level [6]. The method was later extended with a MARS regression model [8, 19]. Bauder et al. additionally employed traditional unsupervised anomaly detection models where it was shown that Local Outlier Factor [10] performed the best [9].

From the existing research, it can be seen that machine learning and data mining techniques can be beneficial when used in the field of unsupervised healthcare insurance anomaly detection. However, limited research has been performed in cooperation with European health care insurers and fewer still has been performed with the latest anomaly detection and explainability techniques.

Methods

In Figure 1 a general overview of the workflow can be found. First, the data is preprocessed such that a suitable data set is created for the machine learning models. Next, a wide variety of unsupervised anomaly detection techniques is employed to detect and rank the actual anomalous profiles. For the anomalous samples, which have been detected, SHAP is employed such that for each practitioner the feature attribution, or importance, scores are calculated. Finally, the results are summarized and visualized to help the expert investigators in interpreting the results and guiding the investigation by helping the experts to efficiently select the care provider profiles that might warrant further investigation. When employing the workflow, it was found useful to have an expert in the loop at all times such that the process can be further optimized. For example, it might turn out that a certain variable, that was included in the preprocessing, is not useful to determine anomalies or has an unwanted impact on the anomaly detection methods.

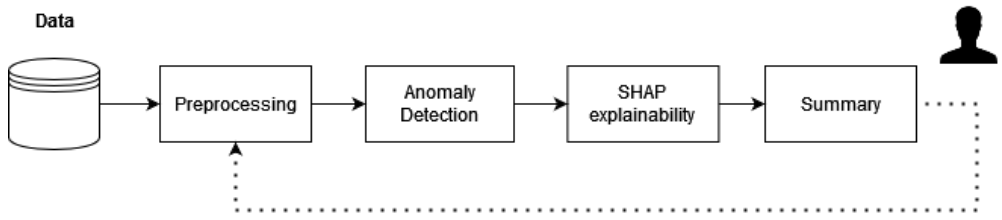


Figure 1: The general overview of the workflow to help expert investigators in detecting anomalous behavior and to pre-select care providers with a profile that could have a higher risk of inappropriate resource use. The health insurance data is first filtered and preprocessed to obtain a single, rich, feature vector for each practitioner. Next, anomaly detection is performed by a number of traditional and state-of-the-art anomaly detection techniques. The model output, for a sample to be investigated, is explained by using SHAP. Finally, the results are summarized in the form of visualisations and summary statistics to the expert investigator.

Preprocessing

Figure 2 shows a schematic overview of the preprocessing steps performed before employing the actual anomaly detection techniques. Resident experts were involved in the entire process, including the preprocessing, in order to identify useful features and interpret the data correctly. This allowed us to reduce the number of available features from hundreds to 66 different features that were identified to be potentially useful in characterizing anomalous behaviour.

Typically, tabular data contains both real and categorical data. Most machine learning methods however cannot by default be applied to categorical features. Currently, the standard is to use one-hot encoding, also known as dummy encoding [22], for categorical data. For a single categorical variable, such encoding is performed as follows. Consider a categorical variable C with n unique categories $\{c_1, c_2 \ldots, c_n\}$. Then a category c_i is transformed into a vector $x \in \mathbb{R}^n$ where each element is defined by equation 1.

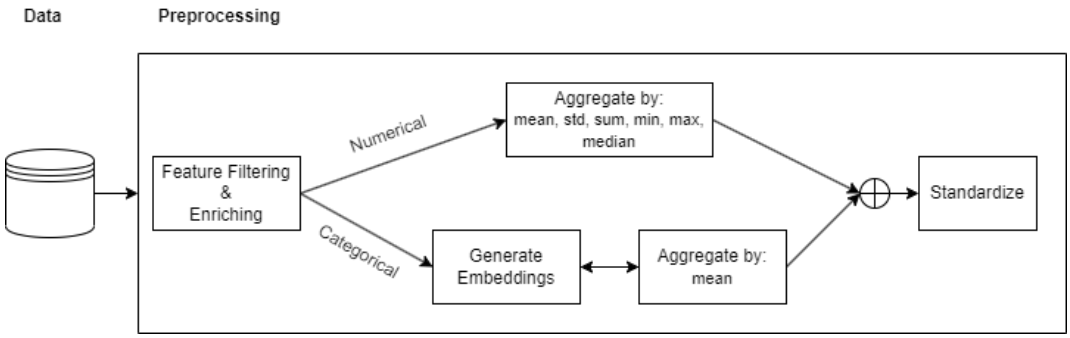


Figure 2: The preprocessing performed before performing anomaly detection. In close cooperation with experts, the features are filtered and enriched (e.g. using expert knowledge) such that a suitable set of the data and features are extracted that are useful in determining atypical behavior. Numerical features are aggregated per practitioner by their mean, standard deviation, sum, minimum, maximum and mean statistics. The categorical features are embedded using a categorical embedding method and aggregated using the mean. The data is then recombined and standardized by subtracting the mean and dividing by the standard deviation.

$$x_i = \begin{cases} 1 & \text{if } C = c_i \\ 0 & \text{if } C \neq c_i \end{cases} \tag{1}$$

However, for categorical features with a high cardinality this results in long feature vectors which greatly increases the dimensionality of the data set. Additionally, not all machine learning methods are able to efficiently handle high-dimensional inputs which can be detrimental to their performance. A better alternative is to transform categorical features by using categorical embeddings [14]. These embeddings have been shown to extract useful information, such as relations between the categorical variables, and result in a much smaller feature size. It was decided to use Latent Semantic Analysis (LSA) [15, 16] as the embedding method as this was shown to be the most consistent [14].

LSA originates from the information retrieval field where it is used to create word and document vector representations. For categorical variables it was shown that a similar process can be used to create vector representations of the categorical values. First, a co-occurrence matrix X is constructed with as rows all possible categories (from all categorical variables) and as columns all actions (of all practitioners). Then, a binary matrix is constructed such that each entry of the matrix indicates whether a specific categorical value is present in a practitioner action or not. At this point, each categorical value is represented in X by a large binary vector which could be used for machine learning tasks. However, this vector representation has as dimensionality p, the number of practitioner actions (rows) in the data set, which is often too large for any practical use. Therefore, in LSA, a rank-k approximation is made of X by means of the Singular Value Decomposition (SVD) as in figure 3. This approximation extracts the most important relations from the co-occurrence matrix while removing redundancy and noise. The categorical embeddings for all categories are obtained by considering only the category part of the decomposition: $XP = CS \in \mathbb{R}^{c \times k}$ where each row corresponds with the embedding of a category with dimension k. Finally, the resulting embeddings are l2-normalized. This normalization is typical in information retrieval as it allows, for example, to easily compare two vectors by using the cosine distance by a dot product [4].

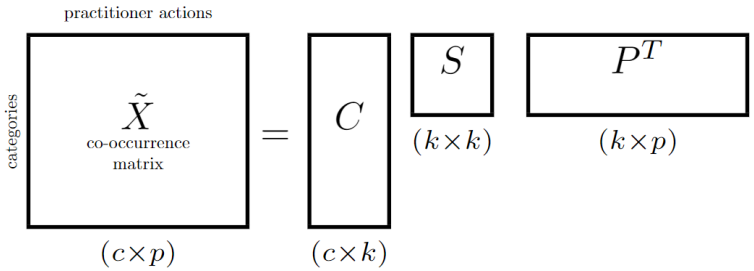


Figure 3: The Singular Value Decomposition step of LSA applied to categorical variables. The co-occurrence matrix, relates the categories with the data entries (the practitioner actions), gets decomposed into a rank-k approximation of the category space C, the practitioner actions space P and weights S which relates the two.

The embedding size k is a tuning parameter as the optimal value depends on the problem. Generally, a trade-off is made between the size and expressivity of the vectors. Larger embeddings can capture more information, up to a point, but make the learning process slower due to the larger dimensionality. In natural language processing, large embedding sizes (e.g. > 100) are usually preferred [43] while for categorical embeddings, smaller sizes already achieve a significant performance increase [14].

For the embedding size, it was decided to use 3 dimensions as this showed to provide a good performance improvement during our experiments while keeping the overall dimensionality manageable. Increasing the dimensionality offered little to no performance improvement while only increasing the computation complexity. The process described above creates an embedding, in our case a 3-dimensional vector, for each categorical value possible. Each value in the original data set is then replaced (embedded) by the corresponding embedding. After embedding the categorical features, the data set contains only real-valued features.

Healthcare insurance data is time series data where each sample is an action for a particular practitioner or institution. Because of the different possible actions, varying lengths and the scale of the data, the time series are typically aggregated into a single vector per practitioner. The aggregation follows previous work [21, 9] such that all numerical features are aggregated by their sum, mean, median, standard deviation, minimum and maximum statistics. For the embedded data, the features are aggregated using the mean. Finally, the resulting feature columns are standardized before training by, for each feature, subtracting the mean and dividing by the variance.

Models

The experiments are performed using a wide variety of anomaly detection methods to investigate which types of algorithms are most suitable. This section gives a short overview of these methods. **Nearest neighbor** methods are commonly used methods due to their simplicity and ease to explain. These methods take into account the k nearest neighbors of a data point to calculate their outlyingness score. In the simplest case, for the K Nearest Neighbors algorithm, the distance between a sample and its k nearest neighbors can be calculated and averaged. This average distance can then be used to rank the samples such that anomalies get a higher outlyingness score because they're furthest from their neighbors. A disadvantage of these methods is that they are generally less effective for high dimensional problems as the effect of any single feature becomes less important when calculating the distance between two samples. K Nearest Neighbors (KNN) [33] and Local Outlier Factor (LOF) [10] have been successfully used on Medicare data [9] where these methods were the best performers. Additionally, we consider Angle-base Outlier Detection (ABOD) [23] and Subspace Outlier Detection (SOD) [24] which have been shown to work better on high dimensional data. The final nearest neighbor method that is employed is Learnable Unified Neighborhood-based Anomaly Ranking (LUNAR) [20], a state-of-the-art local anomaly detection technique. This model combines the nearest neighbor approach with graph neural networks.

Isolation based methods take the entire data set into account and attempt to partition the data into regions such that small outlying groups are easily isolated from the rest of the data. These methods generally employ decision trees which makes them explainable and fast because they're easily parallizable. Isolation Forest (IForest or IF) [27] use an ensemble of binary decision trees constructed from a random sample of the data. The anomalies are isolated by randomly splitting the data at each level based on a random feature. An outlyingness score is calculated based on the length of the path down the decision trees of a sample. Unsupervised Random Forest (URF) [2] uses the standard random forest in an unsupervised setting by introducing synthetic data. The proximity matrix provided by the random forest algorithm is then used to calculate an anomaly score. A more recent method in this category is called Isolation-based Nearest-Neighbor Ensembles (INNE) [5]. It was shown that this method outperforms iForest when it comes to detecting local anomalies and high dimensional data with many irrelevant features.

One-class methods attempt to learn the single class that is represented by the training data. Typically, such methods attempt to find suitable trade-off between a hypersphere that encompasses as much of the training data as possible while at the same time keeping the hypersphere small. The most well known of these methods is the One-Class Support Vector Machine (OCSVM) [36]. Deep Support Vector Data Description (DeepSVDD) [35] can be seen as an extension to the OCSVM where the encompassing hypersphere is learned together with non-linear feature representations by using a deep neural network .

Distribution estimation methods attempt to learn the data distribution, or an approximation thereof, which can then be used to identify anomalies. A Variational Autoencoder (VAE) is a probabilistic generative model in the form of an autoencoder. The reconstruction probability of a trained VAE can be used as an anomaly score. Lightweight On-line Detector of Anomalies (LODA) [32] is a method where the data set is projected on a number of random vectors of which an ensemble of one-dimensional histograms is then created. The method takes inspiration from ensemble and boosting methods and shows that a collection of weak models can achieve state-of-the-art results while being one of

the fastest anomaly detection methods. Copula-Based Outlier Detection (COPOD) [25] and Empirical-Cumulative-distribution-based Outlier Detection (ECOD) [26] are two recent parameter-free methods. The methods use different way to inspect, for each dimension, the tails of their univariate empirical cumulative distribution function and aggregate this information into an anomaly score.

Explainability

Machine learning approaches have been slow to be adapted for use in the healthcare insurance field. One important reason for this is that it can be difficult to interpret and explain model output. Understanding why a machine learning models makes a prediction is a requirement if it is to be used by experts in order to make a decision concerning the presence of fraud or overconsumption.

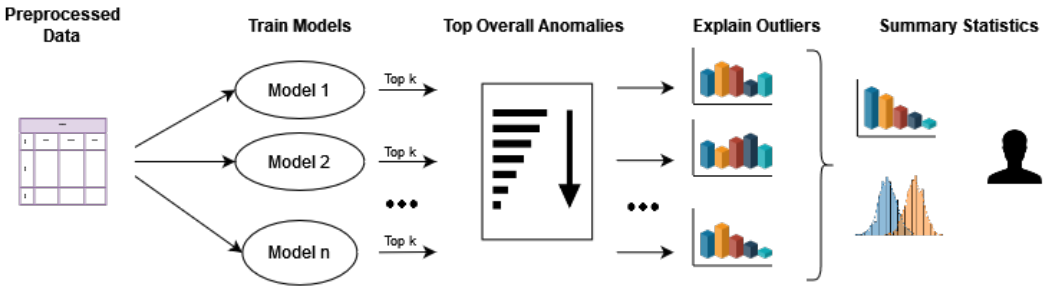


Figure 4: The process for explainable anomaly detection. Preprocessed data is provided to n anomaly detection models. The top k anomalies for each models are explained using SHAP. Next, the anomalies are ranked based on how many times they are detected by the anomaly detection methods. finally, The explanations in combination with standard statistics are summarized and provided to an expert to help them in their decision making.

Recently, a model-agnostic technique, called SHAP [28], was proposed that allows any model output to be explained. The method can assign an importance value to any feature for a single prediction or a group of predictions by using Shapley values [37], originating from cooperative game theory. Calculating SHAP values is computationally costly except when using linear or Tree-based models [28]. However, it is not necessary to calculate these values for the entire data set as only the top outlying samples are of interest. Figure 4 shows how SHAP is incorporated in the workflow. For the trained anomaly detection models, the top k outlying care providers are selected. Each of these samples may be considered separately and investigated individually. However, to further limit the search window, we found it informative to sort the anomalies based on number of times it is present in the top k for all models. As a result, an expert is provided a list of profiles with unusual resource use where the most anomalous, over all models, are ranked higher.

A large advantage of SHAP is that Shapley values are fully additive, this means that the respective SHAP values of multiple features can be added to get their corresponding total importance. This is especially useful when working with embeddings, or feature statistics, where a single feature is transformed into a new representation with a larger dimensionality. In such cases, it is often not easy to calculate an importance score for the original feature using traditional methods whereas using SHAP we simply need to sum the scores for all dimensions. This process is illustrated in figure 5. The left graphs shows the feature contribution of the categorical embeddings and numerical feature statistics. For both categorical and numerical features, the dimensionality of the input features is larger than that of the original features. The feature contributions of all new dimensions are summed to find the total importance of a feature with regards to predicting the anomaly score of a practitioner.

At this point, the top anomalous samples over all models are listed together with their SHAP feature contributions. The investigator is able to see why a model assigns a high anomaly score to a practitioner. This is useful as different techniques might flag a sample as outlying for different reasons. However, a single explanation for all models provides a better overview, especially when working with many different models. Therefore, all SHAP scores are re-scaled and the mean over all models is calculated to provide a single plot that shows the feature contributions over all models. Finally, it was found to be useful to provide additional statistics from the original time series data. Specifically, for an outlying sample, the SHAP values of top the aggregated features are shown together with the univariate feature distributions of the time series data for the top outlying features. For the categorical features, the distribution of the practitioner is showed together with that of the population. For the numerical features, the values are compared with the average population values and standard univarate statistics such as the Z-score are shown.

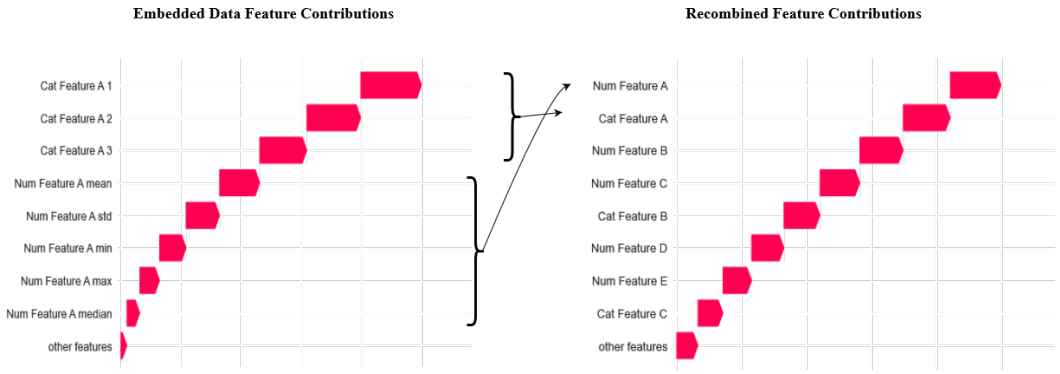


Figure 5: The SHAP feature contributions for the feature embeddings are summed such that the contribution of the original feature is achieved. For a numerical feature, all values are aggregated per practitioner using the mean, standard deviated, minimum, maximum and median. For a categorical variable, the values are embedded and aggregated by the mean. Both methods expand a feature into multiple dimensions which means there are as many feature contributions (importance scores) per original feature. SHAP has the advantage that these scores can be summed to achieve the overall importance score for the original feature.

One important feature, that is provided to the expert in the form of summary statistics, is the nomenclature for the action performed by the institution or practitioner. For example, in the United States, these are the HCPCS codes [13]. In Belgium, these codes are simply referred to as nomenclature codes [34].

The final result is a clear overview where a field expert can immediately identify why a sample is flagged as outlying and whether further investigation for the sample is needed. An example of an importance plot with univariate statistics for a categorical and numerical feature are shown in figure 6. The top anomalous practitioners are investigated by first inspecting the overall SHAP importance plot. The expert will then investigate the summary statistics of the most important features in combination with the nomenclature codes in order to identify the nature of the anomaly. The proposed workflow filters and sorts the anomalous practitioners, reducing the search window, and provides an explanation for the detection.

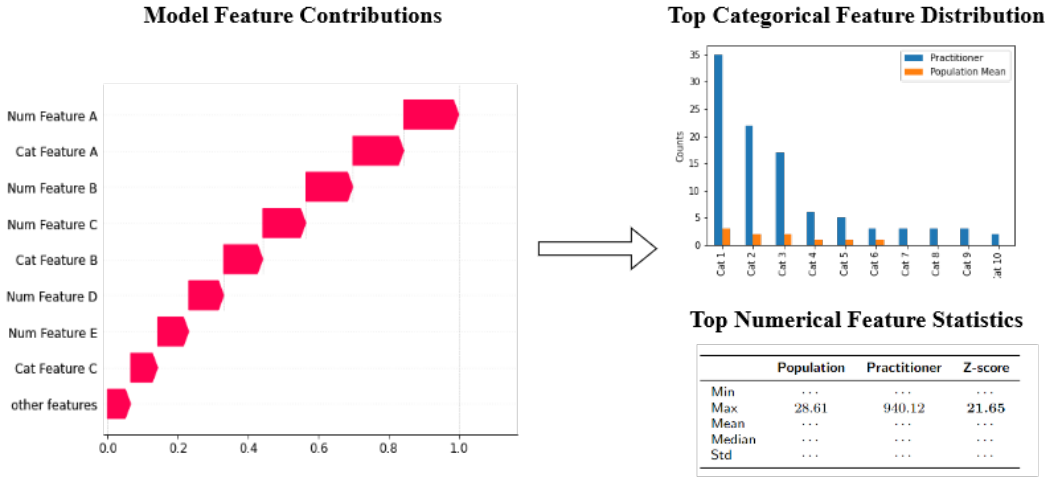


Figure 6: The SHAP model feature contributions, for a single sample, indicate which feature contribute most to the model output i.e. the anomaly score. The features contributing most to the sample being considered anomalous are then further investigated. For a categorical feature, the distribution of the categorical values are shown and compared to the average of all practitioners. For a numerical feature, the aggregate features (i.e. the minimum, maximum, mean, median and standard deviation) are shown, compared with the mean of the all practitioners and the Z-score is provided. This provides a clear overview for the expert investigator that shows why a practitioner's behavior is flagged as anomalous.

Results and Discussion

Belgian Health Insurance Data

The CM data analysed in this paper has only been used for research purposes, as a proof of concept, without any operational goals. All research has been performed on pseudonymized data at the CM under supervision of the chief medical officer. No re-idenfitication of the data points has been performed at any time, regardless of whether the samples have been identified as anomalous or not.

In this section we consider the data set of all general practitioners (GPs) in Belgium that have more than 50 entries per year in the CM database. This results in 11851 different GPs. A set of 38 general practitioners was provided who were known for their anomalous profiles due to excessive use of certain billing codes. This setting is a decidedly imbalanced task where only 0.3% of the data set are identified as known anomalies. It is important to note that there could be, and likely are, more unknown anomalies present in the data. The original data contains 12899707 individual medical acts, in the form of resource use data, of all general practitioners. Each practitioner action has 66 features of which 54 contain categorical features and 12 contain numerical features. The cardinality for the categorical features varies between 2 and 159155. After the preprocessing, described in the previous section, the data sets have the final number of features and samples as shown in Table 1.

Table 1: The number of samples and features of the data sets used during the experiments. The original data is aggregated and transformed into one-hot encoded and embedding data sets.

Data	**Samples**	**Features**
Original	12899707	66
One-hot	11851	191707
Embeddings	11851	234

Experiment Results

The performance of the anomaly detection methods is quantified using the area under the ROC curve (AUROC). All models output a real valued anomaly score which is used to sort the samples with more anomalous samples being ranked higher. The ROC curve is calculated using these anomaly scores and labels indicating whether the practitioner has a known anomalous profile.

The AUROC has two advantages which makes it suitable to the discussed setting. It's immune to class imbalance, which is certainly the case, and it assigns a higher score when the outlying class is ranked higher than the other class. This second point makes it especially suitable as one goal is to investigate which anomaly detection methods can best help experts with identifying novel and outlying samples, which would be the methods that are able to rank anomalies higher.

The experiments were performed using Python where the models have been implemented using the Scikit-learn[30], PyOD [44] and Tensorflow [1] packages.

In Table 2 the results of the experiments on the transformed CM data sets can be found.

A clear difference can be observed between the one-hot encoded and the embedding data. For almost all methods, the embedding method significantly outperforms the current standard, one-hot encoding. Healthcare insurance data sets are rich in categorical features with often thousands of different values. One-hot encoding makes the number of features explode in size while not providing any additional useful information which is detrimental for the performance and time complexity of most machine learning methods. The only exception to this observation is the LUNAR method. The learnability aspect of the graph neural network approach is able to overcome the limitation of standard neighbor methods where their performance tends to deteriorate for very high dimensional data.

Overall, the best performing methods are VAE and LODA with INNE, OCSVM and COPOD following closely behind with the proposed embedding method. It can be seen that the more recent and state-of-the-art anomaly detection methods generally outperform the methods such as KNN and isolation forest which are traditionally used for health insurance data. The exception to this is the one-class SVM which is one of the best performing methods on the embedding data.

The experiments show that unsupervised anomaly detection methods can help to detect or prioritize the same anomalies that have been detected by experts. However, to do so effectively, special care must be taken during the preprocessing step, particularly with regards to the categorical features.

Table 2: Area under the ROC curve of for the one-hot and embedding encoded CM general practitioner data for the following algorithms: Local Outlier Factor (LOF), K Nearest Neighbors (KNN), Isolation Forest (IF), Unsupervised Random Forest (URF), Angle-base Outlier Detection (ABOD), Empirical-Cumulative-distribution-based Outlier Detection (ECOD), Copula-Based Outlier Detection (COPOD), Variational Autoencoder (VAE), Histogram-based Outlier Score (HBOS), Isolation-based Nearest-Neighbor Ensembles (INNE), One-Class Support Vector Machine (OCSVM), Deep Support Vector Data Description (DeepSVDD), Subspace Outlier Detection (SOD), Learnable Unified Neighborhood-based Anomaly Ranking (LUNAR) and Lightweight On-line Detector of Anomalies (LODA).
For the stochastic algorithms, the performance shown is the average with standard deviation over 10 different runs. The embedding method offers a clear performance increase for all methods. VAE and LODA achieve the best performance with INNE, OCSVM and COPOD following closely behind.

Model	**Onehot**	**Embeddings**
LOF	0.60	0.69
KNN	0.60	0.78
IF	0.53 ($\pm$0.04)	0.76 ($\pm$0.03)
URF	0.51 ($\pm$0.01)	0.62 ($\pm$0.01)
ABOD	0.60	0.76
ECOD	0.60	0.75
COPOD	0.60	0.81
VAE	0.51 ($\pm$0.00)	**0.82 ($\pm$0.00)**
HBOS	0.51	0.79
INNE	0.52 ($\pm$0.01)	0.80 ($\pm$0.01)
OCSVM	0.60	0.81
DeepSVDD	0.58 ($\pm$0.02)	0.61 ($\pm$0.03)
SOD	0.61	0.77
LUNAR	0.77 ($\pm$0.01)	0.78 ($\pm$0.01)
LODA	0.56 ($\pm$0.01)	**0.82 ($\pm$0.01)**

Explainability

In this section, we show how unsupervised machine learning methods in combination with SHAP can be used to aid decision making and help experts in discovering new types of anomalies. This will be done by considering a real-life example of a novel trend that was discovered while performing the experiments.

When investigating the top anomalous profiles of general practitioners, a particular practitioner had high SHAP feature contribution scores for the features indicating the discipline, location, subcategory and the total billing amount of the medical intervention. When investigating the provided univariate feature statistics, the billing amount was significantly higher than is standard for general practitioners. The feature distribution for the discipline showed that most actions were part of the orthopedics discipline. The type of the action showed a large amount of actions related to surgeries. The location indicated that these took place at general hospitals. Finally, the distribution of nomenclature codes revealed actions such as arthroplasty of the knee or hip using a prosthesis, meniscectomy etc. As a summary, this investigation into the practitioner showed that they mainly aided in surgeries involving orthopedics in a hospital.

During further investigation, multiple practitioners were identified by the algorithm with similar behavior, indicating a trend which was previously unknown to experts at the CM. While this behavior by itself does not indicate fraud or waste, it is clearly outlying for a standard general practitioner and, as such, provides novel avenues of investigation for experts. Important for the experts was that the results of the used anomaly detection techniques were interpretable: the investigator was able to immediately identify the anomalous behavior using the combination of SHAP and simple univariate statistics and visualisations.

Conclusion

In this work, we discussed a workflow for explainable anomaly detection in the healthcare insurance field. The workflow implements and applies a number of recent advances in machine learning: categorical embeddings, state-of-the-art anomaly detection methods and SHAP explainability, which are novel to unsupervised anomaly detection in healthcare insurance. The experiments show that unsupervised anomaly detection techniques can successfully detect and prioritize practitioners with known anomalous behaviour which were provided by experts. It was shown that the current standard

of one-hot encoding categorical features is detrimental for most models due to the high cardinality of the categorical features, which are common in healthcare insurance data. Instead, we propose to use categorical embeddings, which provide a significant increase in performance.
Finally, we showed how the machine learning techniques can be used in combination with SHAP and simple summary statistics to explain the predicted anomalies and aid expert investigators. We illustrated how the approach was successfully used in the field by providing an example of a previously unknown anomalous trend for general practitioners

Funding

KU Leuven: Research Fund (projectsi BOF/23/064, C16/15/059, C3/19/053, C24/18/022, C3/20/117, C3I-21-00316), Industrial Research Fund (Fellowships 13-0260, IOFm/16/004) and several Leuven Research and Development bilateral industrial projects; Flemish Government Agencies: FWO: EOS Project no G0F6718N (SeLMA), SBO project S005319N, Infrastructure project I013218N, TBM Project T001919N; PhD Grant (SB/1SA1319N), EWI: the Flanders AI Research Program, VLAIO: CSBO (HBC.2021.0076) Baekeland PhD (HBC.20192204) European Commission: European Research Council under the European Union's Horizon 2020 research and innovation programme (ERC Adv. Grant grant agreement No 885682); Other funding: Foundation 'Kom op tegen Kanker', CM (Christelijke Mutualiteit)

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`.

[2] Nelson Lee Afanador, Agnieszka Smolinska, Thanh N. Tran, and Lionel Blanchet. Unsupervised random forest: a tutorial with case studies. *Journal of Chemometrics*, 30(5):232–241, 2016. doi: https://doi.org/10.1002/cem.2790.

[3] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md. Rafiqul Islam. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278 – 288, 2016. ISSN 0167-739X. doi: https://doi.org/10.1016/j.future.2015.01.001. URL `http://www.sciencedirect.com/science/article/pii/S0167739X15000023`.

[4] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, Toronto, 1999. ISBN 0-201-39829-X.

[5] Tharindu R. Bandaragoda, Kai Ming Ting, David Albrecht, Fei Tony Liu, Ye Zhu, and Jonathan R. Wells. Isolation-based anomaly detection using nearest-neighbor ensembles. *Computational Intelligence*, 34(4):968–998, 2018. doi: https://doi.org/10.1111/coin.12156.

[6] R. A. Bauder and T. M. Khoshgoftaar. A probabilistic programming approach for outlier detection in healthcare claims. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 347–354, Dec 2016. doi: 10.1109/ICMLA.2016.0063.

[7] R. A. Bauder, T. M. Khoshgoftaar, A. Richter, and M. Herland. Predicting medical provider specialties to detect anomalous insurance claims. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 784–790, Nov 2016. doi: 10.1109/ICTAI.2016.0123.

[8] Richard A. Bauder and Taghi M. Khoshgoftaar. Multivariate outlier detection in medicare claims payments applying probabilistic programming methods. *Health Services and Outcomes Research Methodology*, 17(3): 256–289, Dec 2017. ISSN 1572-9400. doi: 10.1007/s10742-017-0172-1. URL `https://doi.org/10.1007/s10742-017-0172-1`.

[9] Richard A. Bauder, Raquel C. Da Rosa, and Taghi M. Khoshgoftaar. Identifying medicare provider fraud with unsupervised machine learning. *Proceedings - 2018 IEEE 19th International Conference on Information Reuse and Integration for Data Science, IRI 2018*, pages 285–292, 2018. doi: 10.1109/IRI.2018.00051.

[10] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, may 2000. ISSN 0163-5808. doi: 10.1145/335191.335388.

[11] Centers for Medicare & Medicaid Services. European health care expenditure. URL `https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nationalhealthaccountshistorical`.

[12] Varun Chandola, Sreenivas R. Sukumar, and Jack C. Schryver. Knowledge discovery from massive healthcare claims data. In *KDD*, 2013.

[13] CMS. Hcpcs - centers for medicare and medicaid services. URL https://www.cms.gov/medicare/coding/medhcpcsgeninfo.

[14] H. De Meulemeester and B. De Moor. Unsupervised embeddings for categorical variables. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020. doi: 10.1109/IJCNN48605.2020.9207703.

[15] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[16] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[17] EHFCN. The EHFCN waste typology matrix. URL http://www.ehfcn.org/what-is-fraud/ehfcn-waste-typology-matrix/.

[18] Eurostat. European health care expenditure. URL https://ec.europa.eu/eurostat/databrowser/view/hlth_sha11_hf/default/table.

[19] Jerome H. Friedman. Multivariate adaptive regression splines. *Ann. Statist.*, 19(1):1–67, 03 1991. doi: 10.1214/aos/1176347963. URL https://doi.org/10.1214/aos/1176347963.

[20] Adam Goodge, Bryan Hooi, See-Kiong Ng, and Wee Siong Ng. Lunar: Unifying local outlier detection methods via graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6737–6745, Jun. 2022. doi: 10.1609/aaai.v36i6.20629. URL https://ojs.aaai.org/index.php/AAAI/article/view/20629.

[21] Matthew Herland, Taghi M. Khoshgoftaar, and Richard A. Bauder. Big data fraud detection using multiple medicare data sources. *Journal of Big Data*, Sep 2018. ISSN 2196-1115. doi: 10.1186/s40537-018-0138-3. URL https://doi.org/10.1186/s40537-018-0138-3.

[22] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer, New York, 2013.

[23] Hans-Peter Kriegel, Matthias Schubert, and Arthur Zimek. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 444–452, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401946. URL https://doi.org/10.1145/1401890.1401946.

[24] Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Outlier detection in axis-parallel subspaces of high dimensional data. In Thanaruk Theeramunkong, Boonserm Kijsirikul, Nick Cercone, and Tu-Bao Ho, editors, *Advances in Knowledge Discovery and Data Mining*, pages 831–838, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[25] Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. Copod: Copula-based outlier detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1118–1123, 2020. doi: 10.1109/ICDM50108.2020.00135.

[26] Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George Chen. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2022. doi: 10.1109/TKDE.2022.3159580.

[27] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008. doi: 10.1109/ICDM.2008.17.

[28] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[29] Alberto Marino and Luca Lorenzoni. The impact of technological advancements on health spending. (113), 2019. doi: https://doi.org/https://doi.org/10.1787/fa3bab05-en. URL https://www.oecd-ilibrary.org/content/paper/fa3bab05-en.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[31] Johan L. Perols, Robert M. Bowen, Carsten Zimmermann, and Basamba Samba. Finding needles in a haystack: Using data analytics to improve fraud prediction. *Accounting Review*, 92(2):221 – 245, 2017. ISSN 00014826.

[32] Tomáš Pevný. Loda: Lightweight on-line detector of anomalies. *Machine Learning*, 102(2):275–304, Feb 2016. ISSN 1573-0565. doi: 10.1007/s10994-015-5521-0. URL https://doi.org/10.1007/s10994-015-5521-0.

[33] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. *SIGMOD Rec.*, 29(2):427–438, may 2000. ISSN 0163-5808. doi: 10.1145/335191.335437. URL https://doi.org/10.1145/335191.335437.

[34] RIZIV. Nomenclatuur van de geneeskundige verstrekkingen - riziv. URL https://www.riziv.fgov.be/nl/nomenclatuur/Paginas/default.aspx.

[35] Lukas Ruff, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Robert A. Vandermeulen, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4390–4399, Cambridge, MA, 2018. PMLR. URL http://proceedings.mlr.press/v80/ruff18a.html.

[36] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, July 2001. doi: 10.1162/089976601750264965. URL https://doi.org/10.1162%2F089976601750264965.

[37] Lloyd S. Shapley. *A Value for N-Person Games*. RAND Corporation, Santa Monica, CA, 1952. doi: 10.7249/P0295.

[38] The National Alliance of Christian Mutualities. International website | cm, 2023. URL https://www.cm.be/international.

[39] D Thornton, Guido van Capelleveen, Mannes Poel, Jos Hillegersberg, and Roland Mueller. Outlier-based health insurance fraud detection for u.s. medicaid data. *ICEIS 2014 - Proceedings of the 16th International Conference on Enterprise Information Systems*, 2:684–694, 01 2014.

[40] Dallas Thornton, Roland M. Mueller, Paulus Schoutsen, and Jos van Hillegersberg. Predicting healthcare fraud in medicaid: A multidimensional data model and analysis techniques for fraud detection. *Procedia Technology*, 9:1252 – 1264, 2013. ISSN 2212-0173. doi: https://doi.org/10.1016/j.protcy.2013.12.140. URL http://www.sciencedirect.com/science/article/pii/S2212017313002946. CENTERIS 2013 - Conference on ENTERprise Information Systems / ProjMAN 2013 - International Conference on Project MANagement/ HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies.

[41] Peter Travaille, Roland Mueller, Dallas Thornton, and Jos van Hillegersberg. Electronic fraud detection in the u.s. medicaid healthcare program: Lessons learned from other industries. In *Proceedings of the 17th Americas Conference on Information Systems (AMCIS)*, pages 1–11, United States, 8 2011. Association for Information Systems.

[42] Guido van Capelleveen, Mannes Poel, Roland M. Mueller, Dallas Thornton, and Jos van Hillegersberg. Outlier detection in healthcare fraud: A case study in the medicaid dental domain. *International Journal of Accounting Information Systems*, 21:18–31, 2016. ISSN 1467-0895. doi: https://doi.org/10.1016/j.accinf.2016.04.001. URL https://www.sciencedirect.com/science/article/pii/S1467089515300324.

[43] Zi Yin and Yuanyuan Shen. On the dimensionality of word embedding. *Advances in neural information processing systems*, 31, 2018.

[44] Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019. URL http://jmlr.org/papers/v20/19-011.html.