NOVA IMS Information Management School

DSAA

Mestrado em Data Science and Advanced Analytics

Master Program in Data Science and Advanced Analytics

Fraud- and anomaly detection in healthcare – an unsupervised machine learning approach

Lennart Dangers

Internship report as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

FRAUD- AND ANOMALY DETECTION IN HEALTHCARE – AN UNSUPERVISED MACHINE LEARNING APPROACH

by

Lennart Dangers

Internship report presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics

Advisor: Nuno Miguel da Conceição António

September 2021

ABSTRACT

Fraud and abuse in healthcare are critical and cause significant damage. However, the auditing of healthcare encounters is cumbersome, and the detection of fraud and abuse is challenging and binds capacity. Data-driven fraud and anomaly detection models can help to overcome these issues. This work proposes several unsupervised learning methods to understand patterns and detect abnormal healthcare encounters which might be fraudulent or abusive. The ensemble of models is split into sub-processes and applied on a healthcare data set belonging to Future Healthcare group, a Portuguese group acting in health insurance. One major part of the ensemble is the implementation of the Isolation Forest algorithm, which achieves good results in precision and recall and detect new potential fraudulent abnormal behaviour. Due to unlabelled data and the application of unsupervised learning methods, the proposed model detects new fraudulent patterns instead of learning from existing patterns. Besides the model to predict whether new incoming medical encounters are fraudulent or abusive, this work illustrates a visual method to detect suspicious networks among medical providers. In addition, this work contains an approach to predict whether a customer will cancel the insurance based on anomalous behaviour. This internship report aims to contribute to science and be public, even though some parts could not be explained in detail due to confidentiality.

KEYWORDS

Fraud detection; Anomaly detection; Healthcare data; Unsupervised learning; Clustering; Machine learning; Isolation forest; Network analysis; Cancellation Prediction

ACKNOWLEDGEMENTS

This work is a practical project related to an academic internship as a partial requirement for obtaining a Master's degree in Data Science and Advanced Analytics. I would like to thank the Future Healthcare Group for giving me the opportunity, especially Ana Pina, for all the support during and after the project.

I would also like to thank my thesis advisor Nuno António for his great help and effort before and during the project phase.

Finally, I would like to thank my family and friends, without whom I would not be at this point.

Thank you!

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CRISP-DM	Cross-industry process for data mining
DBSCAN	Density-based spatial clustering of applications with noise
EDA	Exploratory data analysis
ETL	Extract, transform, and load
FN	False negatives
FP	False positives
IDE	Integrated development environment
iForest	Isolation Forest
GAN	Generative adversarial network
GMM	Gaussian mixture models
PR-AUC	Precision-Recall Area under the curve
RFM	Recency, frequency, and monetary
SHAP	Shapley additive explanations
TN	True negatives
TP	True positives
XGBoost	Extreme Gradient Boosting

1. INTRODUCTION

Fraud and abuse in healthcare cause high financial harm, and medical encounters are rapidly increasing. Around 10 per cent of healthcare expenditures can be related to fraud and abuse (Zhang et al., 2020). While healthcare insurance companies already implemented auditing departments and specific rule-based techniques, detecting fraud and abuse is challenging and cumbersome (Carvalho et al., 2017). Due to the high and rising number of daily healthcare encounters, hand checking every encounter binds too many human resources and cannot be the desired goal. Nevertheless, human domain knowledge is a valuable and rare resource for insurance companies and should be used wisely. Besides the number of encounters, the difficulties in detecting fraud are varied. One reason is the small number of fraudulent or abusive encounters among all healthcare encounters. Due to this tiny ratio, it is challenging to detect those encounters because they might disappear into the crowd. Fraudsters adapt their strategy over time, and rule-based detection needs to be updated steadily. However, detecting novelties while implementing new fraudulent patterns rules based on business knowledge is not always possible (Zhang et al., 2020).

The detection of fraud and abuse, also called “anomalies”, can be considered either an unsupervised or a supervised classification task. Since the given raw data does not contain any labels, this work focuses on unsupervised learning techniques to detect anomalies containing fraudulent and abusive encounters. Labelling data is cumbersome, and unlabelled data increases the difficulty of measuring the model performance (Domingues et al., 2018). However, unsupervised learning methods applied to anomaly detection help detect new fraudulent patterns.

This work is a practical project related to an academic internship at Future Healthcare Group[1]. Future Healthcare provides corporate customers, such as insurance companies, with services related to health insurance and a healthcare provider network. Services are delivered through a digital platform in which the healthcare providers bill the corresponding healthcare encounters. The anonymised data from the encounters are the primary data set for this work. In this context, fraud and abusive behaviour can be perpetrated by the healthcare provider, the insurance user, or even both. This work attempts to connect both to rank the daily incoming encounters checked by the auditing department.

Overall, the goal is to detect suspicious behaviour and increase knowledge about behaviour patterns in healthcare. Concerning this goal, the project covers three main approaches. The first approach is clustering to understand the different behaviour groups of patients and if insurance might have odd behaviour. The second approach is the core part of this work, namely anomaly detection. In this part, the goal is to detect whether an insurance user, a healthcare provider, or the encounter itself is anomalous. The final approach is using a supervised learning model to predict if an insurance user with anomalous behaviour will cancel the insurance within the first year after contracting. The ultimate objective of the final part is to build a proof of concept to determine if it is possible to detect insurance users who have an odd behaviour and are more likely to cancel in the first year, after three months of insurance utilization. The primary approaches, however, are the unsupervised learning approaches of clustering and anomaly detection.

[1] Future Healthcare Group, https://www.future-healthcare.pt/en/

The project's methodology is based on the CRISP-DM process, a well-known methodology in data science projects (Chapman et al., 2000). However, at the moment of writing this document, the project has not yet reached the deployment phase. The current project phase is the evaluation which is why not all results are covered. The document addresses readers with basic knowledge in machine learning and data science, even though some of the main concepts and algorithms are explained in chapter 2.

The content is divided into six main chapters. Chapter 2 contains the main theoretical framework, definitions, and related work in fraud and anomaly detection. Chapter 3 introduces the methodology, which is based on the first steps of the CRISP-DM process. Chapter 4 summarises and discusses the main results. Chapters 5 and 6 conclude the work and give ideas for future works.

As mentioned above, this project deals with healthcare data, and the data set is confidential. For this reason, variables are anonymized. Due to company secrets, detailed descriptions in feature engineering cannot be disclosed. However, this work contains all necessary information and does not have a blocking notice to inhibit further research.

2. LITERATURE REVIEW

This chapter includes the main theoretical framework in 2.1, where the leading machine learning concepts related to this work are briefly introduced. However, this is not the main scope of this thesis. Sub-chapter 2.2 contains the introduction of the algorithms used inside modelling (not including those used for model selection; a list can be found in the appendix). Sub-chapter 2.3 deals with the main topic of this work. Here, literature related to fraud- and anomaly detection in healthcare is reviewed. This chapter is the basis for the methodology of this work and justifies the chosen techniques.

2.1. MACHINE LEARNING

The field of machine learning contains various concepts. In this chapter, the main ideas tackled in this work are introduced.

2.1.1. UNSUPERVISED LEARNING VS SUPERVISED LEARNING

Primary concepts in machine learning are those of unsupervised and supervised learning. In this context, supervised means providing a learning algorithm with more information about the target. For example, in data sets for fraud detection, historical data can contain information on whether a historical transaction was fraudulent. In this case, each observation has a label about the target for the chosen algorithm to learn. Having a label is called supervised learning (Han et al., 2012). Subclasses of supervised learning are regression, which involves a numeric target, and classification. Classification deals with a categorical target that could have binary or multiple labels. Alternately, unsupervised learning models do not include information about the target. Algorithms are not able to learn from a given label. Nevertheless, an unsupervised model can also be used for fraud detection to disclose new or rare patterns. Typical unsupervised learning tasks are clustering or anomaly detection (Provost & Fawcett, 2013).

Other concepts that are not covered in this work are semi-supervised learning, which requires both labelled and unlabelled data to learn, and reinforcement learning.

2.1.2. MODEL EVALUATION AND METRICS

Evaluating a model is a challenging yet crucial task in machine learning. Due to the lack of labels, unsupervised models are challenging to evaluate. Alternately, based on the existence of true labels, the performance of supervised learning models is easier to measure. One goal in classification tasks is to develop a model that learns from the given data but, in the same way, can generalise well for unseen data. It is vital to put aside minor data, which is not used to train the model to evaluate the performance to generalise new instances. This data set is called the test set, whereas the more extensive data set is called the training set. If the model is too complex, it tends to overfit, which means evaluating the training set is important.

In contrast, the performance on the test set is inadequate (high generalisation error). Working with a confidential data set is crucial in both supervised and unsupervised tasks. When comparing different models, it is recommended to implement another data separation. Another data set can be used just to select the best working model. This set is usually called a validation set (Géron & Safari, 2019). The terminology of validation and test set are sometimes used interchangeably (Kelleher et al., 2015).

As mentioned, the evaluation for unsupervised tasks, such as clustering, is not easy. However, in clustering, one can measure the stability of a cluster run. Another method is the computation of the silhouette score, which measures how well the clusters are separated from each other and how close the instances within one cluster are (Han et al., 2012). On the contrary, several defined metrics exist in supervised learning, computed based on the true labels. In this work, one supervised binary classification task is introduced, which means training a model with observations and corresponding true labels to predict these target labels for new instances (Provost & Fawcett, 2013).

In classification, the results depending on the predicted value and the actual value. For example, in a classification model to predict fraud, fraud will be the positive class (or 1), whereas no fraud is defined as the negative class (or 0). A predicted fraud observation that is confirmed as fraudulent is called true positives (TP). If the prediction is incorrect, these cases are false positives (FP). Negative predictions that are true are called true negatives (TN), and their corresponding error is false negatives (FN). The confusion matrix summarizes the results, usually considering the test set (Figure 1). Depending on the literature, the structure of the confusion matrix varies, having the predicted values as columns and the actual values as rows (Kelleher et al., 2015).

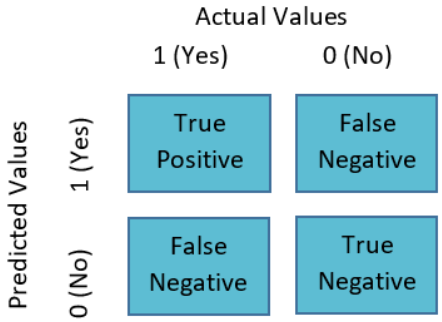


Figure 1: Confusion matrix

Based on this matrix, several metrics can be computed. For the classification task of this work, the most common ones are considered (Han et al., 2012).

(1) $$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

(2) $$Precision = \frac{TP}{(TP + FP)}$$

(3) $$Recall = \frac{TP}{(TP + FN)}$$

(4) $$F1 = \frac{(Precision\ x\ Recall)}{(Precision + Recall)}$$

The evaluation metric has to be chosen wisely according to the task and the target class distribution. In fraud and cancellation prediction, the target class is most likely not equally balanced. Usually, far more instances are not fraudulent or have not been cancelled, called an imbalanced target class (Géron & Safari, 2019). In this case, one is not interested in TN since they are the majority.

For this reason, rather than accuracy, recall and precision are suitable evaluation metrics for imbalanced data sets. The F1 score combines the two metrics and gives a good first overview.

However, there is a precision/recall trade-off: increasing the recall leads to a lower precision and vice versa (Powers, 2015). This trade-off can be depicted in the form of the PR-curve with the corresponding metric PR-AUC. The PR-AUC is a metric similar to ROC-AUC. However, instead of plotting FP (x-axis) and TP (y-axis), the PR-Curve displays recall (x-axis) and precision (y-axis). Thus, PR-AUC is an appropriate metric and visual tool for classifications of imbalanced data sets to set the threshold in predictions to obtain the desired outcome (Davis & Goadrich, 2006). Figure 2 shows the comparison of both curves. In this study, the PR curve is used as an evaluation metric.

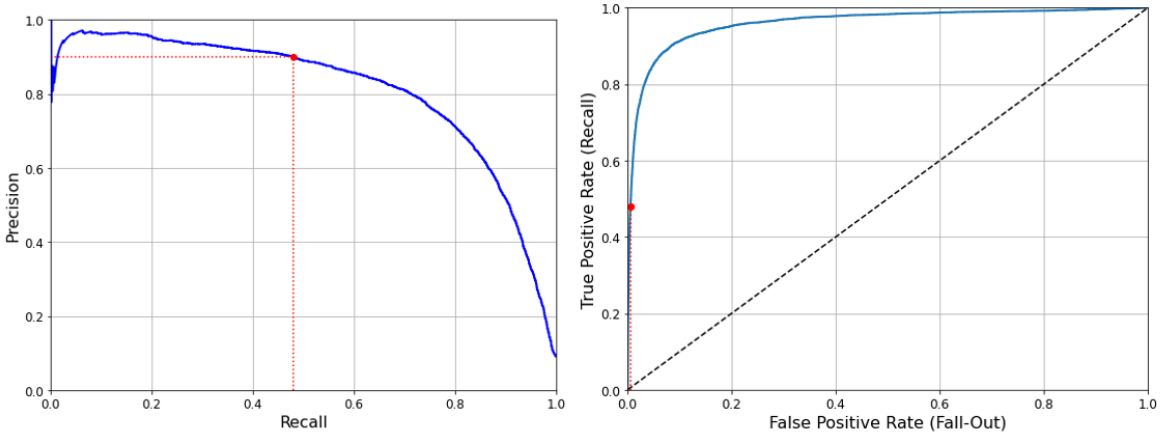


Figure 2: Comparison between PR-Curve (left) and ROC-Curve (right) from (Géron & Safari, 2019)

2.1.3. MODEL EXPLAINABILITY

When it comes to predictive models, it is challenging to interpret results quickly. Especially in ensemble models and deep learning environments, the gap between accuracy and interpretability is increasing. In a business context, the explainability of a model's output is vital to implement a data science solution and generate trust among all the stakeholders. Shapley Additive explanations (SHAP) is a novel method to overcome these issues and was first introduced in 2017 by Lundberg and Lee. Essentially, the method is based on Shapley values in game theory, and SHAP values compute how much each feature contributes to the outcome (prediction) of a machine learning model. Two main advantages of SHAP values are local and global interpretability. The global interpretability shows how much each feature contributes to the model overall.

In contrast, the local interpretability is the SHAP value computation for each observation (Lundberg & Lee, 2017). Both local and global interpretability is crucial in a business context. For example, applied to a fraud- and anomaly detection model, it might be helpful to know which variables to which extent have an impact on an anomaly. The local analysis explains which specific values to which extent caused the result for one observation only. This work uses SHAP values to explain the feature importance of the models but does not fully cover the theoretical approach. For further information, the consultation of the original paper is recommended. Nevertheless, it is crucial to read the SHAP visualizations and calculations correctly to understand the SHAP results.

SHAP is implemented in a library in Python and contains several methods and visualizations. In this work, two visualizations techniques related to global interpretability are used to explain the output of the models in anomaly detection and cancellation prediction, which are explained following (Lundberg et al., 2019; Lundberg & Lee, 2017).

Summary plot (Feature importance):

Figure 3 displays an example of a summary plot (feature importance plot) with SHAP. It lists essential features and shows the positive and negative relationship with the target. Here the target is the prediction, whether it is anomalous or normal. This plot is helpful to obtain a general overview of the feature importance (ranked in descending order) and their impact on the target. The colour represents the value of the feature. Red are high values, whereas blue values are low. For instance, a high value in variable "var15" in Figure 3 negatively impacts the target. Meaning, the higher the value in "var15", the more likely an observation is anomalous (higher negative anomalous score). The corresponding Python library provides more similar plots than those mentioned above. Nevertheless, the summary plot provides the most information within one visualization (Lundberg et al., 2019).

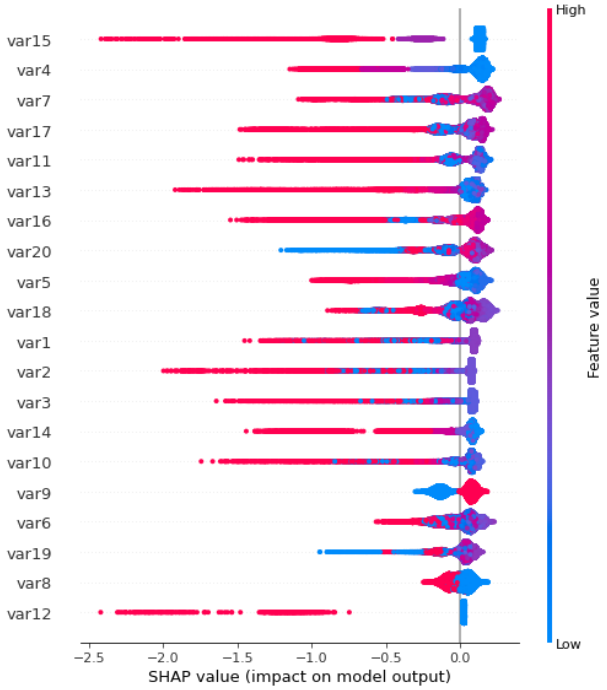


Figure 3: Example of feature importance plot with SHAP

Partial dependency plot:

Additionally, Figure 4 shows a dependency plot, which shows the relationship between a chosen variable and the target. The quoted implementation colours another feature automatically, which interacts the most with the chosen feature. As an example, Figure 4 points out two main insights. Firstly, "var3" interacts most with "var2", and secondly, higher values of "var2" have a higher impact of being anomalous. Compared with the summary plot, the dependency plot is a more detailed view towards one or two features but still considering all instances (Lundberg et al., 2019).

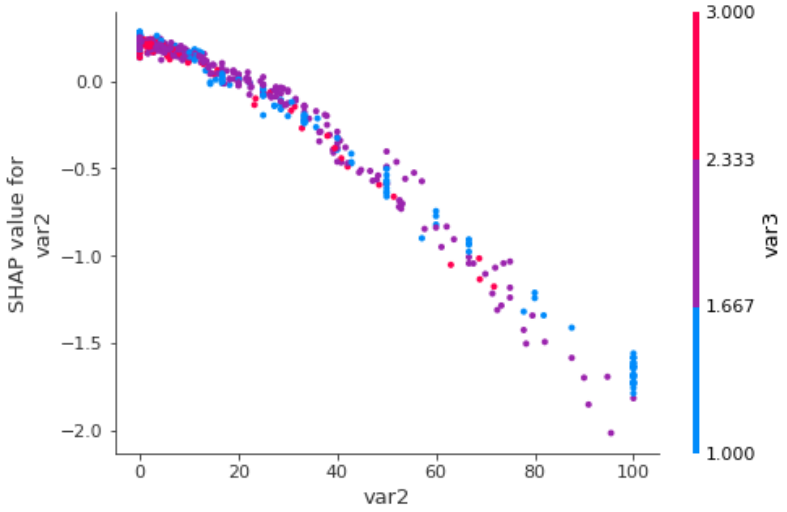


Figure 4: Example of a partial dependency plot with SHAP

2.2. ALGORITHMS

The following chapter deals with the algorithms implemented in modelling. It does not contain algorithms that are chosen for model selection within the supervised learning task. The rationale behind this selection can be found in related work is based on the best performance in the classification task.

2.2.1. ISOLATION FOREST

The isolation forest (iForest) was first described in 2008 and is one of the main approaches of this work (F. T. Liu et al., 2008). In other anomaly detection models (distance- and density-based methods), regular instances are usually defined. Based on this normality, anomalies are detected. Alternately, iForest focus on anomalies instead of profiling regular instances. In this case, the model is trained to detect anomalies and does not profile normal points, which might detect too many regular instances as anomalies or the other way around — not detecting anomalies if they are anomalies. This algorithm, which works with sub-sampled data, is applicable for imbalanced and has linear time complexity. Due to the corresponding paper, iForest performs well in high dimensional data with many irrelevant variables and is more skilful than other techniques, such as local outlier factors.

The base assumption of this method is that anomalies are few, and the corresponding values are separate from others. This approach uses a tree structure (isolation trees or iTrees) where anomalies end up in pure leaf nodes closer to the tree root to isolate instances. Regular points, on the other hand, need more data split to be fully isolated. The iForest is an ensemble of iTrees based on randomly sub-sampled data, which are both the main parameters of this algorithm. Because of its ability to separate anomalies close to the root, the algorithm converges quickly and does not need a large sub-sample size. Each iTree does random partitions until instances are separated. The path length, telling how many partitions were done from the root node until termination in a leaf node, of each iTree is calculated. The next step is to summarise the path lengths of all and divide them by the number of iTrees (number of estimators).

The main output of this algorithm is the anomaly score for each instance, which is computed as such:

(5) $$s_{paper} = 2^{-\frac{E(h(x))}{c(n)}}$$

$E(h(x))$: Average number of edges to be separated for an instance x.

$c(n)$: Normalisation constant for a data set with n instances.

An anomaly score of one gives a high probability of being an anomaly, whereas instances with a score lower than 0.5 are usually considered regular points. Anomaly scores around the threshold of 0.5 are on the edge of being anomalous or normal.

In this work, the implementation of iForest is done with the Python machine learning library scikit-learn (Pedregosa et al., 2011), available since 2016. This implementation uses ExtraTreeRegressor as the base estimator (iTree) and splits the data randomly on chosen variables and values. The output can either be a binary label (-1's are anomalies and 1's are regular points) and an anomaly score. Unlike described in the paper, the anomaly score in scikit-learn is calculated as such:

(6) $$s_{scikit} = \left(0.5 - s_{paper}\right)$$

More minor scores are considered anomalies or more anomalous instances.

2.2.2. DBSCAN

DBSCAN, short for Density-based spatial clustering of applications with noise, is a clustering algorithm based on density. Unlike distance-based clustering algorithms, DBSCAN can detect different shapes that contain outliers and noise (Ram et al., 2010). Initially, the algorithm was introduced in 1996 (Ester et al., 1996). Its main parameters are the distance measure epsilon and the number of minimum samples within the distance epsilon. For each data point, DBSCAN counts the number of instances within the defined distance epsilon. If the number of points is equal to or higher than the minimum samples number, the given point is called a core point. Core points are located in dense areas, and all instances in a neighbourhood of a core point form a cluster. Any given point which is neither a core point nor has a core point in the neighbourhood (border point) is called an anomaly or noise point. DBSCAN is not only a clustering algorithm but also a method to detect anomalies (Géron & Safari, 2019). Unlike K-Means, one does not have to define the number of clusters because the number is based on the hyperparameter settings of epsilon and the minimum number of samples (Abdulraheem et al., 2015). A main advantage is detecting clusters with arbitrary shapes (Z. Chen & Li, 2011). However, the main drawbacks of DBSCAN are the determination of epsilon and the minimum number of samples in advance. Finding the appropriate values for both is a challenging task and can lead to suboptimal performance (Abdulraheem et al., 2015).

Again, the implementation is done with scikit-learn. Anomalies or noise points obtain the label "-1", and each cluster gets a label from zero onwards. The counting of labels can give a notion about the distribution of each cluster and the anomalies. The computational complexity is around O(n log n) and slightly higher than linear. In scikit-learn, the complexity depends on the value of epsilon and can rise to $O(n^2)$ (Géron & Safari, 2019).

2.2.3. K-MEANS

K-Means is an established unsupervised machine learning algorithm clustering proposed mainly by James MacQueen and Stuart Lloyd (Lloyd, 1982; MacQueen, 1967). The algorithm divides the data set into a given number of K clusters. In K-Means, the number of clusters (K) needs to be set in advance, which can be an issue according to the given problem. First, the (Euclidean) distance from each data point to one of the randomly generated seed points, so-called centroids, is computed. The number of centroids is equal to the number of selected clusters (K), and each data point is assigned to its closest centroid. Next, the centroids are re-calculated. The new centroids are the mean of the point within one cluster, considering the sum of the squares of the distance to its assigned centroid. Finally, the distance of each instance to each (new) centroid is calculated and probably reassigned. The process is iterative, and the algorithm converges with no reassignment of data points and their corresponding centroids (Bishop, 2006). The steps of iteration are displayed in Figure 5.

The main advantage of K-Means is its ability to converge fast and the scalability with large data sets. However, it is necessary to run K-Means several times to obtain stable solutions. As mentioned, defining the number of clusters in advance can be complex and arbitrary. To overcome the pre-selection of K, one can use methods such as the elbow-graph method and the silhouette score. The elbow method computes the inertia for each K, where usually low inertia is desired. The silhouette score considers the mean intra-cluster distance and the mean distance to the instances of the next cluster. The score varies from -1 to +1, where +1 is the perfect score (Géron & Safari, 2019). Another drawback is the sensibility towards outliers (Han et al., 2012).

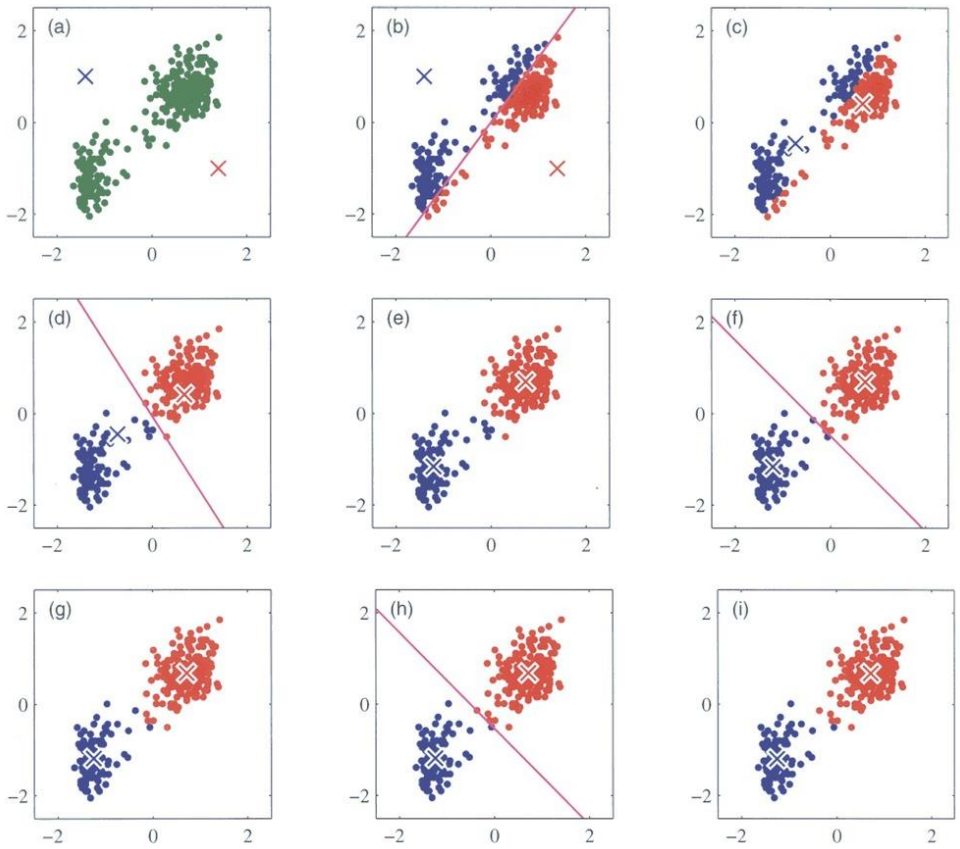


Figure 5: Iteration overview of K-Means algorithm from (Bishop, 2006)

2.2.4. XGBOOST CLASSIFIER

Current machine learning competitions show that tree boosting is an effective algorithm to use for classification problems. More precisely, XGBoost (short for extreme gradient boosting), a scalable ensemble method for tree boosting, was introduced by Chen and Guestrin in 2016 and since then has often achieved good results in classification tasks. (T. Chen & Guestrin, 2016). The XGBoost algorithm is an optimized implementation of the Gradient Boosting algorithm. Gradient Boosting adds predictors to an ensemble model, improves each previously made prediction, and minimizes overall prediction error (Géron & Safari, 2019).

2.3. RELATED WORK

In this section, some of the research related to this study is summarized and evaluated. This review gives an overview of the current state of the art and justifies the methods chosen in the methodology part of this work. In addition, it contains the work related to anomaly detection in general, applied to healthcare, and clustering techniques as a special section for anomaly detection.

Anomaly detection, sometimes referred to as outlier detection or novelty detection, is a process to detect observations that are far apart from other instances. Over the last decades, anomaly detection and its increasing demand in various domains is a common subject of research (Pang et al., 2021).

This work overviews unsupervised machine learning algorithms' general performance and characteristics applied for anomaly detection, including fraud detection. It introduces the possibility of using an anomaly score which acts as a label to show if an instance is an anomaly. In contrast, the label is based on a certain threshold of the anomaly score. They are pointing out the significant challenges of unsupervised anomaly detection. These are class imbalance, class overlap, and the speciality for which the unsupervised detection model is created. Anomaly detection models suffer from imbalanced classes because, per definition, an anomaly is a rare case. Due to the imbalance, the evaluation metrics are recall and precision, in addition to the memory and runtime complexity and robustness (Domingues et al., 2018).

Anomalies might have similar values in some variables and are overlapping with no anomaly class. According to Domingues, these methods can be differentiated into probabilistic methods, distance-based methods, density-based methods, neighbour-based methods, information theory, neural networks, domain-based methods, and isolation methods. The authors conclude that iForest is an excellent method to detect anomalies efficiently and with high scalability on extensive data compared to good memory usage. Due to a more general perspective considering a large sample of models, this work acts as a good starting point to implement iForest as the main algorithm in this study. When sampling a small proportion of anomalies from a classification data set, applying density-based methods or gaussian mixture models (GMM) can work well. However, GMM methods are time and resource-consuming without outweighing models such as iForest.

The authors use a generative adversarial network (GAN) to detect fraudulent health care providers. The GAN's model identifies the labels distinguished into normal and anomalous labels and generates new labels. It turns an almost unsupervised problem into a supervised classification model. Based on these labels, standard classification algorithms such as Logistic Regression, XGBoost, Random Forest, and Decision Tree are used to predict whether a healthcare encounter is fraudulent. To explain the

model, SHAP is implemented. Here, one-hot encoding for categorical features is applied, and highly correlated features were removed. Despite having good results, to train a GAN, one needs to have a few true labels.

For this reason, this might instead be applicable as a semi-supervised problem when having a few labels. Nevertheless, using SHAP as an explanatory model is good practice to understand which features have a higher impact on anomalies. Lastly, applying a similar method might improve the performance after obtaining true labels in business tests (compare chapter recommendations and future works) (Naidoo & Marivate, 2020).

Thornton et al. (2014) applied an anomaly detection model on healthcare data with healthcare providers, encounters, and patients as entities. They provide a process, a roadmap for anomaly detection in healthcare in which an anomaly score is computed.

Zhang et al. (2020) consider 10 per cent anomalies to fraud and abuse and mention the main problems to develop and apply anomaly detection systems in healthcare. Firstly, without an exact rule, it is challenging to differentiate abnormalities from regular treatments. Secondly, fraudsters cover their behaviour behind a high number of regular transactions in an already imbalanced environment. Finding one solution that fits all is problematic due to various diseases, patient characteristics, and doctors' specialities. Lastly and because of a high number of transactional data, updating anomaly detection systems can cause issues.

If the patterns are well known, supervised learning techniques promise to achieve better performance. However, supervised learning approaches mean that the data is trained on specific patterns and has issues in detecting novelties. Secondly, in healthcare, the dataset is usually not comprehensive, which increases the chance of an overfitted model. In contrast, they point out that most unsupervised learning anomaly detections are based on clustering with K-Means and DBSCAN and anomaly detection with iForest. Based on Ikono (2019) meta-analysis, in which 88 articles were studied, traditional fraud detection is difficult to apply in healthcare because new fraud patterns occur frequently. Zhang et al., 2020 introduce a group of features related to monetary, frequency, and information regarding the person and the medical treatment. The authors tested traditional rule sorts, K-Means, DBSCAN, iForest, and local outlier factor, with iForest having the best detection rate. The result is justified by the ability of iForest to deal with irrelevant features. This article is one of the main justifications behind using iForest and DBSCAN in this work (Pourhabibi et al., 2020).

Another anomaly detection approach is the creation of a network, also called a graph. According to J. Liu et al. (2016), graph analysis is applied to detect suspicious entities, their relationships, geospatial dispersion, and anomalous network structure. A network is a visual tool that gives a good overview of the data. Suspicious entities and relationships can be found with filtering, selection, and zoom into the network according to the user's needs. Transactional healthcare data usually contains various encounters with many patients and healthcare providers. As mentioned previously, this variety exacerbates the implementation of a good-working fraud detection model and requires fundamental domain knowledge. The detection of suspicious connections among healthcare providers, graph-analytics methods can help to overcome these issues. In their network design, they include geospatial data to recognise fraudulent patterns in a specific region. Once a suspicious connection of providers is found, the final step is applying the iForest algorithm on this sub-sample to detect anomalies within one group.

In a broader perspective, Pourhabibi et al.(2020) analysed 39 academic papers related to graph-based anomaly detection among various areas. Graph-based anomaly detection enables one to find new network designs according to the area of implementation. Even though it works as a roadmap for future network designs, another finding is the dependency of research studies on domain knowledge. Without this knowledge, the finding of new patterns is not likely (Pourhabibi et al., 2020). These authors examined different credit card fraud techniques and pointed out that fraudsters change their behaviour over time. Even though the majority of credit card fraud detection is supervised, supervised learning methods have drawbacks. Besides, they require labels, which are challenging to get in anomaly detection, mainly when humans perform it as class imbalance affects the efficiency of supervised algorithms. Based on a public data set on Kaggle, the authors compare the unsupervised learning anomaly detection algorithms local outlier factor (LOF), One-class support vector, K-Means, and iForest with each other. Eventually, the comparison shows that iForest outranks the other methods by far in each metric (F1 score, accuracy, and AUC score). Even though the paper deals with credit card fraud, the results can be adapted towards anomaly detection in healthcare data and are helpful for this work (Ounacer et al., 2018).

Also, John & Naaz (2019) studied fraud and anomaly detection related to credit card transactions. The authors studied several techniques applied for anomaly detection, such as genetic algorithms and support vector machines and found the effectiveness of random sampling. The experimental part deals with credit card data with labels and compares iForest and LOF among other supervised methods. The usage of iForest is justified with the small memory requirement and low linear time complexity. Applied on the example data set, LOF has the best accuracy, but iForest has higher precision and recall. According to the authors, accuracy usage is critical since it might skew results from data set with imbalanced targets. Due to these results and the imbalance target in the studied data set, the application of iForest is appropriate for this work. This paper points out the advantages and disadvantages of iForest. The advantages, as previously mentioned, are the reduction of flooding and masking, the ability for high scalability, and the low computational overhead and linear time complexity. On the other hand, iForest introduces some random factors that might lead to lower accuracy and stability. The authors describe a combination of iForest and LOF in the related work section, which improves the accuracy and stability but leads to reduced efficiency. The proposed algorithm uses the simulated annealing (SA) method to optimize iForest in terms of generalization and reduces the redundant iTrees and time complexity and is called SA-iForest. This paper compares SA-iForest, iForest and LOF on different data sets. The improved iForest (SA-iForest) performs best in accuracy and execution by the standard implementation of iForest. SA-iForest is an excellent method to test the studied data set after finalizing the test scenario (Xu et al., 2017).

The research from Tang et al. (2011) deals with the medical healthcare data in Australia and unsupervised techniques to tackle fraud and abuse, especially targeting the abuse of prescription. The authors proposed a system containing unsupervised sub-processes, underlying components called feature extractor, cluster builder, model constructor, and outlier detector. According to the authors, the model's modularity increases flexibility and decreases the maintenance complexity due to better localisation. The feature extractor is comparable with the process step of feature engineering and feature selection. The aim is to extract and select a few features to get as much information possible without having too many dimensions (curse of dimensionality). The cluster builder groups the consumers based on their activities which are the input for the model constructor, which builds Hidden Markov Models as an unsupervised learning technique. N-dimensional vectors represent the consumer

patterns, and each dimension implies a pattern encoded as a Hidden Markov Model. The outlier detector generates an n-dimensional score vector compared by the distance against the vector of the same pattern group to compute a final outlier score. This paper shows the advantages of building an ensemble model and considering several unsupervised learning methods to detect anomalies. Based on these insights, this project is structured into parts such as clustering and the utilization of several sub-processes within the implementation of iForest (Tang et al., 2011). The authors implemented a two steps iForest model to detect anomalies to monitor gas turbines. Like healthcare and credit card fraud, anomalies in gas turbines are a rare event and represent an imbalanced data set. After grouping the given data by time series, the first step is to run an iForest with low contamination. The observations labelled as anomalies within the first iForest run are the input for another iForest iteration. In this step, the contamination parameter is high. This filtering can help to detect real anomalies and decrease the false positive rate. Anomaly detection in a supervised context usually works quite well. However, in practical engineering, anomalies are less common and have often not occurred before, making unsupervised learning techniques such as clustering more suitable for this problem. According to the authors, models with unlabelled data have fewer issues with overfitting. They point out that unsupervised techniques based on distance, density, and clustering depend highly on prior knowledge, making it sometimes challenging to achieve high accuracy. iForest overcomes these issues and has other advantages, as mentioned above. In the second step, iForest show promising results in precision, recall, and F1 score and outweigh the one-class support vector machine results. This paper gives values insights to use iForest for the data of the studied data set, even though it was applied in another context. Secondly, a two-way filter of running the iForest algorithm might be a suitable process for future works and the application in healthcare data (Zhong et al., 2019). The study gives a more general overview of unsupervised anomaly detection algorithms without considering iForest. It deals with 19 different unsupervised learning anomaly detection algorithms classified into nearest neighbour-based, clustering-based, and statistical methods evaluated on ten public data sets from several domains. The evaluation criteria are accuracy, the stability of the scoring, the sensitivity to parameters, and the computation time. In general, for global anomalies, the authors recommend nearest-neighbours based methods and LOF for local anomalies.

However, to a certain extent, each studied algorithm requires prior knowledge (Goldstein & Uchida, 2016). This paper introduces several data mining tools and techniques to detect fraud in healthcare data. According to this work, the most common technique is anomaly detection, which aims to point out outliers with a data pattern far apart from other observations. These data mining techniques disclose patterns among large data sets. Kirlidog and Asuk introduce valuable data mining techniques for fraud and anomaly detection: associations, classification, clustering, modelling, and sequential patterns. The types of fraud differ with their type of service, making it more challenging to detect fraud in healthcare data. However, they point out some behaviour that is commonly seen and generally applicable. For example, this can be an unusually high number of encounters for a specific insurance user in general or in a short time frame or excessive prices for medical treatments or consultations. This explanation is a valid rationale to divide features into groups of recency, frequency, and monetary values. In conclusion, the authors attest to techniques such as clustering, anomaly detection, and classification as valuable tools to detect fraudulent encounters and patterns in large data sets. With these three techniques, insurance companies can detect fraudulent encounters and find new fraudulent patterns and knowledge about the behaviour. This article is one of the fundamental bases for the structure and methodology of this study (Kirlidog & Asuk, 2012).

Samriya (2016) describes the clustering of healthcare data in his work. In summary, clustering with the k-means algorithm is recommended. Also, Li & Wu (2012) say k-means performs well, especially brevity, efficiency and speed. The authors examine the power of unsupervised learning techniques to

reveal an unrecognised pattern. Mainly, clustering was one of the main methods used (Ogbuabor & F. N, 2018). In clustering, several different methods occur. The standard methods can be grouped into fuzzy clustering, density-based clustering, hierarchical clustering, and partition clustering. In the latter, one needs to define the number of clusters in advance. In contrast, hierarchical and most density-based clustering does not require such a definition, which is advantageous for these algorithm types. The choice of the proper framework in healthcare is challenging. Their work compares clustering methods, precisely K-Means and Density-based spatial clustering of applications with noise (DBSCAN). The performance evaluation is measured with the silhouette score that measures the separation distance between the resulting clusters. Regarding the results, both clustering algorithms show a high intra-cluster cohesion and an excellent inter-cluster separation. In terms of the silhouette score and runtime, K-Means outweigh DBSCAN. In addition, the authors suggest a model of clustering in comparison with a classification model to predict the activity of instances. The last article is the primary motivation for clustering and the following prediction model.

To summarise the related work, most of the research for anomaly detection deals with supervised data. Nonetheless, unsupervised learning techniques are improving and outweigh supervised techniques depending on the context. The most significant advantage of unsupervised anomaly detection methods is finding anomalies that have not occurred yet, unknown anomalies. In contrast, in supervised methods, the algorithms learn only from known anomalies. Since fraudsters adapt their behaviour and patterns, unsupervised methods help to detect also new fraudulent patterns. Around these methods, it turns out that algorithms such as iForest, DBSCAN, and to a certain extent also LOF have the best results overall studies and broadest context of implementation. Due to its ability to scale on large data sets without requiring much prior knowledge, iForest is the first choice to apply to this study. Based on the literature review, clustering is an appropriate method to understand abnormal behaviour and act as an additional method.

3. METHODOLOGY

The methodology is based on the CRISP-DM process. However, according to this work, not all CRISP-DM output described in the paper is shown here. The phase of deployment, for instance, will just be briefly emphasised. This approach is distinguished into clustering, anomaly detection, and prediction, whereby the central part is anomaly detection.

3.1. TOOLS AND TECHNOLOGY

The following sub-chapter introduces the main tools and technologies used in this work. Python version 3.8 is the primary tool used for all the data preparation and visualisation. Python is an open-source, object-oriented programming language commonly used in data mining and data science projects (Van Rossum & Drake, 2009). Python provides community-contributed packages which support the user with specific tasks such as data preparation, visualisation, and the implementation of machine learning models. While several packages are utilised, the primary two are “Pandas” and “Scitkit-Learn”. “Pandas” is a package built on top of the library “NumPy” that supports multidimensional arrays and is specialised for tasks from reading the data until data preparation. For modelling and the implementation of machine learning algorithms, Scikit-learn provides state-of-the-art implementations. The main algorithm used in this work, isolation forest, is implemented with Scikit-learn. A complete list of packages applied can be found in the appendix. Jupyter Notebook, a web-based Python development tool, is used as an integrated development environment (IDE) (Kluyver et al., 2016). In addition to Python, the programming language R and its corresponding IDE RStudio is used (R Core Team, 2017).

The later introduced network of healthcare providers is done with “Gephi”, open-source software to create, visualise, and explore graphs and networks. To build networks, the input data sets (nodes and edges) need to be in a particular format to display the source-target connection, which is done with Python. Besides creating a graph, Gephi allows computing graph statistics such as centrality measures to fulfil complete graph analysis. The software uses different layout algorithms to spread the network and make it more readable (Bastian et al., 2009).

3.2. BUSINESS UNDERSTANDING

The initial phase of the CRISP-DM process deals with understanding the objectives from a business perspective transformed into a data mining/machine learning goal. The following sub-chapters introduce the general business field and the current situation, the business objective, and the data mining goal.

Since this work discusses anomaly- and fraud detection, defining and differentiating these subjects is vital.

Abuse in healthcare is a direct or indirect treatment or indication without the medical necessity or failure to meet professional standards that end up preventable cost for the payer. As an example, an abusive encounter would be billing for services that were not medically needed. In comparison, health care fraud is an intentional action with the full knowledge of ending up in an unauthorised benefit. Unlike abuse, fraud is officially defined as a crime, for instance, the billing for services that have not been performed (Joudaki et al., 2014).

Anomalies, on the other hand, are instances that have different behaviours than the majority. These behaviours pertain to fraud- and abuse, but also for typical cases that are legitimated. In other words, not every anomaly is fraud or abuse. For example, more severe medical cases that do not occur often might be classified as anomalies. Nevertheless, they are entirely appropriate. Fraud and abusive behaviour can be done by the healthcare provider, the insurance user, or even both.

Figure 6 shows the different kinds of subsets. It is challenging to get accurate estimations regarding the percentage of anomalies, specifically fraud and abuse. According to the literature and figures published by insurance associations, around 6 to 12 per cent can be considered fraud and abuse. In this work, the model assumption works with a value (contamination parameter in isolation forest) of 8 per cent for anomalies.

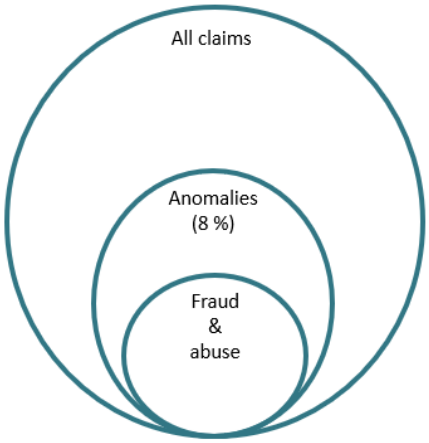


Figure 6: Subsets of the entire population. Fraud and abuse as a subset of anomalies and the entire population

3.2.1. BACKGROUND AND CURRENT SITUATION

According to the Insurance Europe's report, fraud and abuse represent up to 10 per cent of all encounters expenditure in Europe over all types of insurances. [2] Since the dark figure of undetected cases is even higher, fraud and abuse cause impactful damage to insurances and increase premiums. For those reasons, it is crucial to investigate and identify possible fraudulent or abusive behaviour: verify the medical necessity, suitability of services provided, coverage eligibility, and the billing process.

It is a common industry standard to have cost containment rules implemented. These rules automatically check and block the billing of encounters when specific rules are triggered. For instance, an encounter could be refused if medical services were requested, not captured by a specific insurance policy. In addition, fraud detection is cumbersome and involves many human resources and business knowledge — fraudulent and abusive behaviour changes over time. One disadvantage of fixed cost containment rules is that it is less likely to detect new suspicious patterns. Even though business knowledge exists and encounters are examined thoroughly, checking daily encounters can be overwhelming without the notion that cases might be more suspicious.

One technique to overcome this issue is implementing data-driven fraud- and anomaly detection models based on machine learning. These models aim to detect encounters with different patterns

[2] Insurance Europe aisbl, https://www.insuranceeurope.eu/impact-insurance-fraud/

than similar cases. They aim to detect anomalies with suspicious circumstances. As mentioned above, this does not necessarily mean it is fraud. However, it represents rare encounters wherefore fraud and abuse also count (around 8-10 per cent). Fraud- and anomaly detection in machine learning is commonly solved either with supervised or unsupervised learning techniques. Supervised learning requires labels that historical flag encounters fraud or no fraud, which usually requires humans to do so. In many cases, these flags do not exist, and unsupervised techniques have to be applied. Since this study deals with unlabeled data, unsupervised algorithms such as isolation forest are implemented.

To better understand the modelling, it is vital to point out the main entities and the general appearance of encounters. The data contains insurance users and healthcare providers, which can bill encounters on a provided web-based platform.

Before continuing, it is vital to deal with the business terminology briefly.

Terminology	Meaning
Health plans	Healthcare products with fixed conditions (for instance, how many service treatments) that do not contain risks for the insurance company.
Health insurance	A broader health coverage than health plans which covers more and has risks.
Insured person	A person who has valid health insurance.
Insurance user	An insured person who utilises the insurance.
Provider	A healthcare provider for certain medical acts which can access the web-based platform to bill their encounters.
Encounter	Whenever an insured person visits a healthcare provider, one encounter includes all the items associated with one specific indication. One encounter has one person and one provider. Sometimes also called a claim.
Item	One encounter can contain several medical items.
Service	Each encounter can be associated with one medical service, such as treatments or consultations.

Table 1: Most essential terminologies

3.2.2. BUSINESS AND DATA MINING GOAL

This sub-chapter is divided into business and data mining goals. The initial business goals are subdivided into specific data mining goals.

Business goals:

The main business goal is based on the three entities, insurance users, healthcare providers, and encounters. The business goal is to obtain answers to the following questions:

Persons:

1. What kind of behaviour does the insurance users have?
2. Do the behaviour change within the first years of insurance?
3. Are abnormal and suspicious insurance users detectable?
4. Is it possible to predict whether an abnormal person cancels the insurance after the first year of insurance, right after contracting the insurance (3 months)?

Provider:

5. Are abnormal and suspicious providers detectable?
6. How is the connection between the providers? Are there suspicious provider networks?

Encounter:

7. Which of the daily incoming encounters is anomalous?
8. What are the most anomalous encounters for each medical service?

From these questions, this work's most critical and most focused business goal is to rank the daily incoming encounters to support the auditing department and understand the most important influences. Instead of randomly picking encounters, the goal is to choose the ones with the highest probability of being fraudulent or abusive.

Data mining goals:

In addition to a deep understanding of the data through an exploratory data analysis, a more general data mining goal is to build a pipeline with applicable code for similar problems. To develop prototype models as a proof of concept to evaluate if further investment in deployment promises a return on investment. Following the business goal, the interpretability, such as feature importance or impact, of each developed model's output is crucial. The prototype models are:

1. Customer segmentation/clustering model,
2. Anomaly detection model for each entity, person, provider, and encounter,
3. Cancellation prediction model.

Each of the models has its own specific data mining goal. The goal of clustering is to obtain groups of instances with interpretable centroids that are stable over different runs and filtering.

The objectives in anomaly detection are versatile and different according to the corresponding entity. However, obtaining an anomaly score and label for each entity is a common goal to flag and rank

anomalies. For encounters, the goal is to build an ensemble model based on all anomaly scores and labels to consider the encounters themselves and the persons and providers and run each model separately for each medical service to get both a more relevant result and the most anomalous encounters of each service. The model output should be a ranked list of encounters sorted by descending probability of being fraudulent. The final output should include all scores and labels that summarises all daily encounters with a default sorting. However, at first, the sorting should be adjustable for further tests.

In order to detect persons with anomalous behaviour, the goal is to create new features and build a model with the best working set-up (desired F1-score of 80 per cent). Besides implementing a similar anomaly detection pipeline for providers, the aim is to create a visual tool, a provider network, to detect suspicious connections among anomalous providers. This visual tool will be a network/graph with edges and nodes and filters for specific features and values.

3.3. DATA UNDERSTANDING

This chapter deals more precisely with the data itself. It is divided into the data collection, which shows the ETL process (extract, transform, and load), its general description, and a brief report regarding the data quality.

3.3.1. DATA COLLECTION AND DESCRIPTION

The anonymised data analysis requires the data to be stored in an ordered database, for instance, a data warehouse or data mart (Giudici & Figini, 2009). The data source is an internal data mart in which all medical encounters are stored that were submitted on the platform. From the data mart, monthly transactional data was extracted for three years. However, merging and pre-processing the monthly data is separated by year to differentiate according to the model design (Figure 7). Due to the COVID-19 pandemic, which might skew the data, 2020 data were not used for modelling. Furthermore, the project started in October 2020 and the business decision was to use only anonymised data from whole years.

In general, the data contains medical encounters and their corresponding medical items. Each item represents one row and belongs to one unique encounter. The item itself has no unique identifier yet belongs to an encounter with a unique encounter identification number. One encounter has at least one item associated but might have more items that can also be duplicated within one encounter.In total the data contains 5,613,478 items, 1,811,423 encounters, and 62 features.

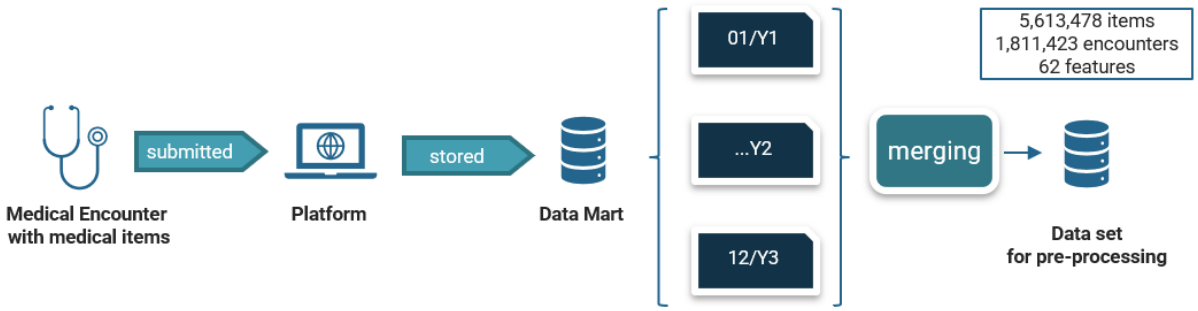


Figure 7: Data collection and preparation at a glance

3.3.2. DATA QUALITY

Since the data source is a data mart with the related advantage of the data schema, the available quality for analytical purposes can be considered reasonable. Regarding the missing values, solely features regarding the location have a missing ratio higher than 3 per cent. The location in each of the models plays a subordinate role and does not affect the performance.

Due to transactional data, only persons and providers who had encounters in the time studied can be considered. Features such as the whole number of persons in one policy are not accurate because people without any encounter are not considered. This issue has to be considered when calculating new features.

3.4. DATA PREPARATION

3.4.1. DATA CLEANING AND FILTERING

The data set includes data from health plans and health insurance. Based on their characteristics, health plans do not contain a high risk for insurances. However, actual health insurance products that are not only products for reimbursements have risks. For this reason, health plans and insurance products with only reimbursements were filtered out. Besides that, transformations such as changing the data types or removing unnecessary features were made. The exploratory data analysis uncovers a few errors in the age, which were discarded. Since the main goal is to detect anomalies, and those are most likely less common, no outliers are removed.

Figure 8 illustrates the time filtering at a glance, and each section has different filtering according to the model goals. Clustering divided the data by the corresponding years of insurance into the first, second, and third-year insurance users. In anomaly detection, all the data until the chosen date is selected. The prediction model takes only the first year of insurance (from clustering), get a label if one cancelled after or within the first year, and filter only the utilisation of the first three months of insurance.

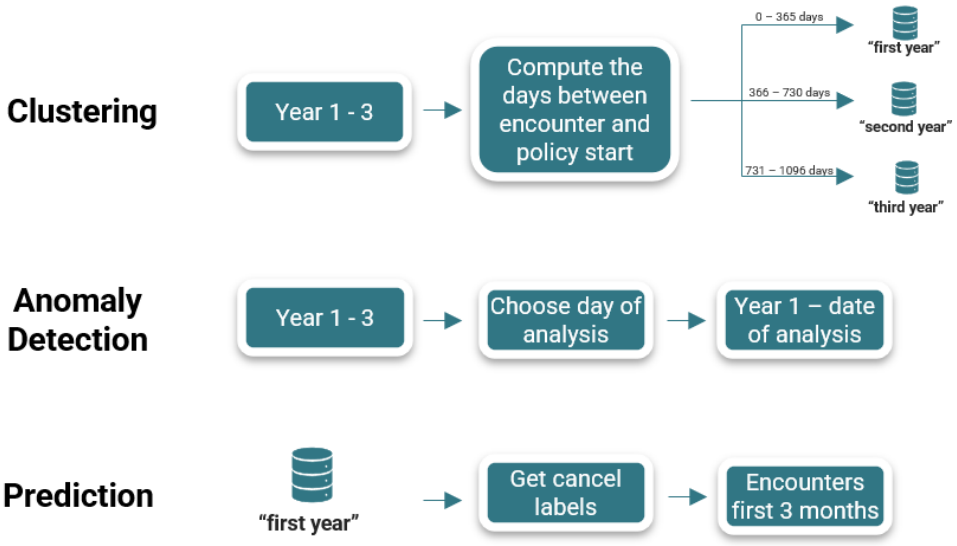


Figure 8: Time filtering for each of the main models

3.4.2. FEATURE ENGINEERING AND SELECTION

This sub-chapter describes the process of feature engineering and its corresponding selection. For each model, the initial data is aggregated by either person, provider, or encounter. Due to various entities, it is vital to consider normalised values and be aware of outliers within one feature. Categorical features are encoded with one-hot-encoding.

In clustering and classification, features can be chosen based on the correlation or other feature selection techniques such as feature importance or recursive feature elimination. However, selecting the correct variables for anomaly detection is challenging and is mainly done based on business knowledge.

The features extracted follow four main groups:

1. General,
2. Recency,
3. Frequency,
4. Monetary.

One of the rationales behind this is the known RFM-Analysis. General features are directly related to the entity itself (for instance, age or location for persons). Some socio-demographic features are enriched with external data, such as census data. As far as possible, categorical features are replaced by continuous values.

The goal of clustering is to divide the data into reasonable clusters to detect patterns and behaviours. For this reason, specific features are selected for clustering, whereas others are used for the following profiling to obtain a better description of each cluster. The feature selection is based on the correlation among the features as well as business knowledge. Features with a correlation coefficient above 0.85 were dropped unless the variable is vital according to the business insights. Features for anomaly detection and the corresponding sub-models are firstly chosen based on business knowledge and secondly validated by their feature importance with SHAP values.

In the provider network, each node needs a unique number to be identified. The features further used are dependent on the network design. In this work, two network designs are created, which are described in the modelling chapter. Besides the geographical location and the service, they provide the most (mode), the most critical feature for providers is the anomaly score and label, which is the output of the described underlying sub-processes in anomaly detection. The edges represent a source-target relationship and list all the unique combinations of providers and persons once. Because both network designs are undirected, it does not distinguish having one entity as a source or target. In network 2, the weight of the edges is another feature. A higher edge weight means more shared distinct patients between two providers. If one provider has no patients in common, there is no edge (connection) between them in the network.

For the cancellation prediction, mostly person-related features are used that are similar to those used for clustering. The results of the previous tasks incorporate as features in this model. Precisely, the anomaly score of a person and the rate of anomalous encounter one person has are two essential features. Highly correlated features (above a coefficient of 0.85) were dropped, and according to the feature importance of the XGBoost algorithm chosen.

3.5. MODELLING

This chapter deals with the modelling technique or the multiple applicable techniques of each section described in the previous chapters.

Clustering:

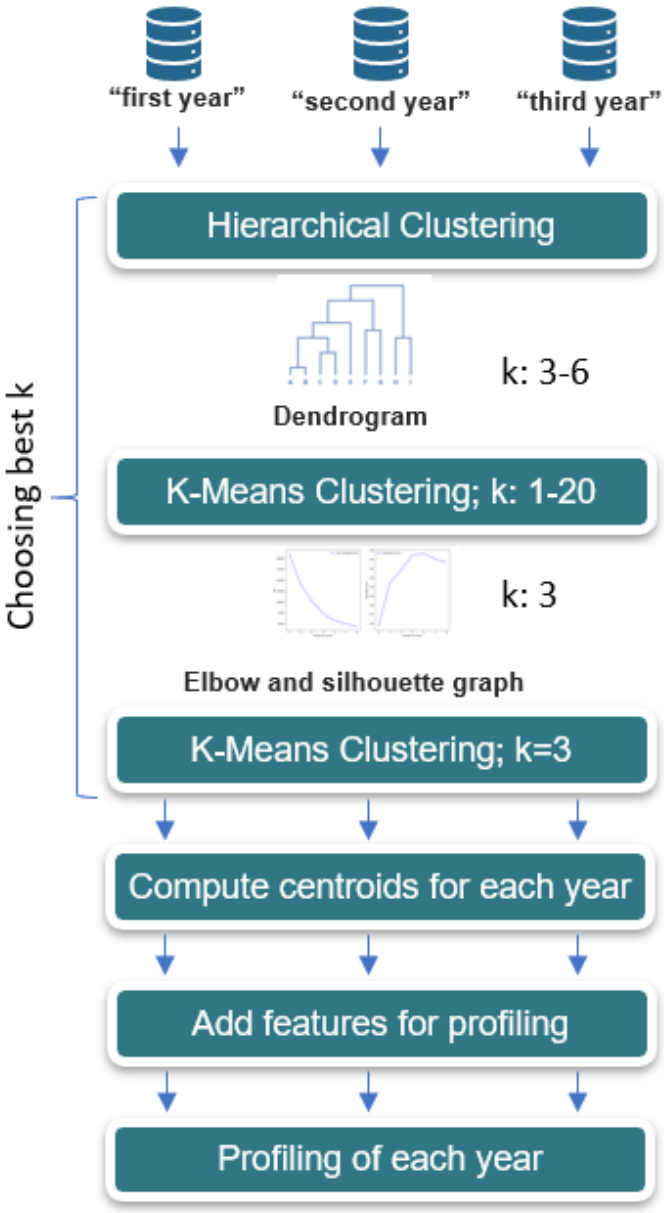


Figure 9: Clustering process overview

The pre-processing output is grouped as datasets by person identification number for their first, second, and third year of insurance. The overview in Figure 9 shows the entire modelling process at a glance.

The modelling technique used for clustering is the K-Means clustering algorithm. Due to the difficulties in defining the parameters of DBSCAN, K-Means is preferably used. The reasoning for this decision is based on the work of Samriya (2016) and Li & Wu (2012), who both recommend k-means for clustering extensive medical data.

The limitation of the number of clusters for interpretation purposes is vital. The focus is instead on understanding the groups of behaviour than detect outlier clusters. However, DBSCAN is considered a possible future improvement to be used for anomaly detection.

Three steps for each data set (first, second, third year) are preceding to determine the best number of clusters. At first, hierarchical clustering is performed from its output, a dendrogram. Hierarchical clustering gives an overview of the entire groupings and helps to define the number of clusters. The dendrogram displays the hierarchy of groups with all the instances on the x-axis. The distance among all of them is illustrated on the y-axis (Provost & Fawcett, 2013). One can get a first idea of how many clusters are suitable (Figure 9). In this case, the cut-off could be either for three or six clusters, visualised by the two dashed red lines. Clustering aims to form groups of instances that are close together and far apart from other groups.

The measure this, the so-called elbow graph and the silhouette score method are applied. The elbow graph measures the inertia (how far away the instances within a group are) for each k in the k-means algorithm. Since low inertia is desired, a k with the lowest inertia or the highest drop (the elbow) in inertia is the right choice. The silhouette score, however, describes how far apart the points in one group are from another. A possibly high silhouette score (ranges from -1 to 1) is desired. Based on the results for each year (Figure 11) and the business perspective, the number of clusters is three.

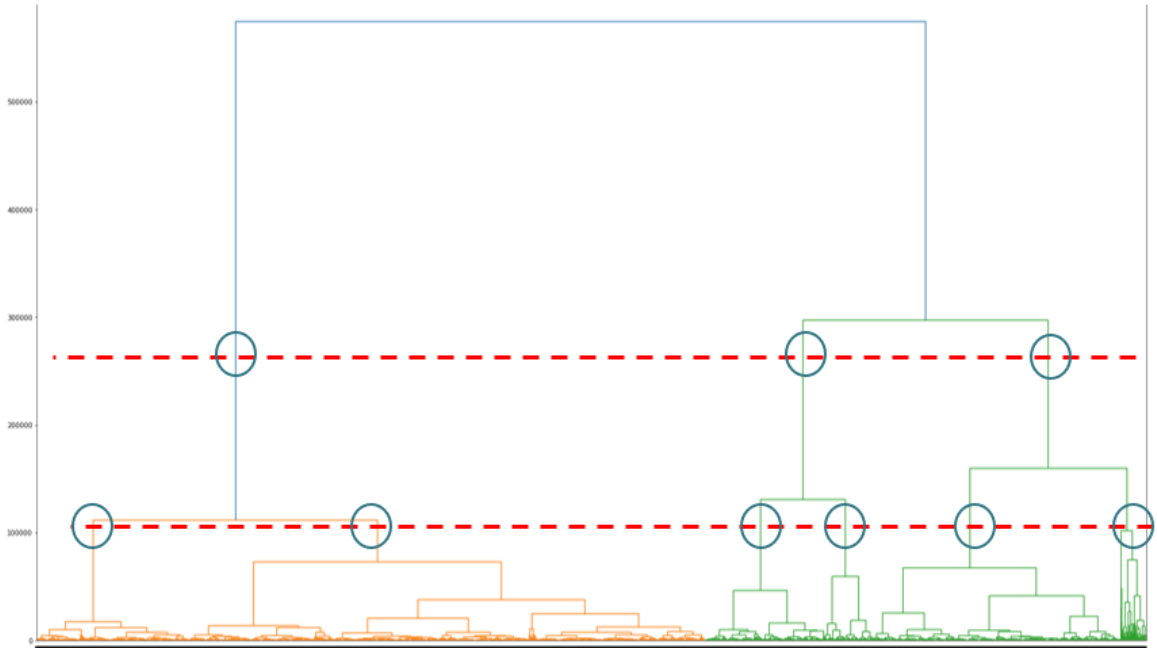


Figure 10: Dendrogram for hierarchical clustering of insurance users. Here for the first year of insurance

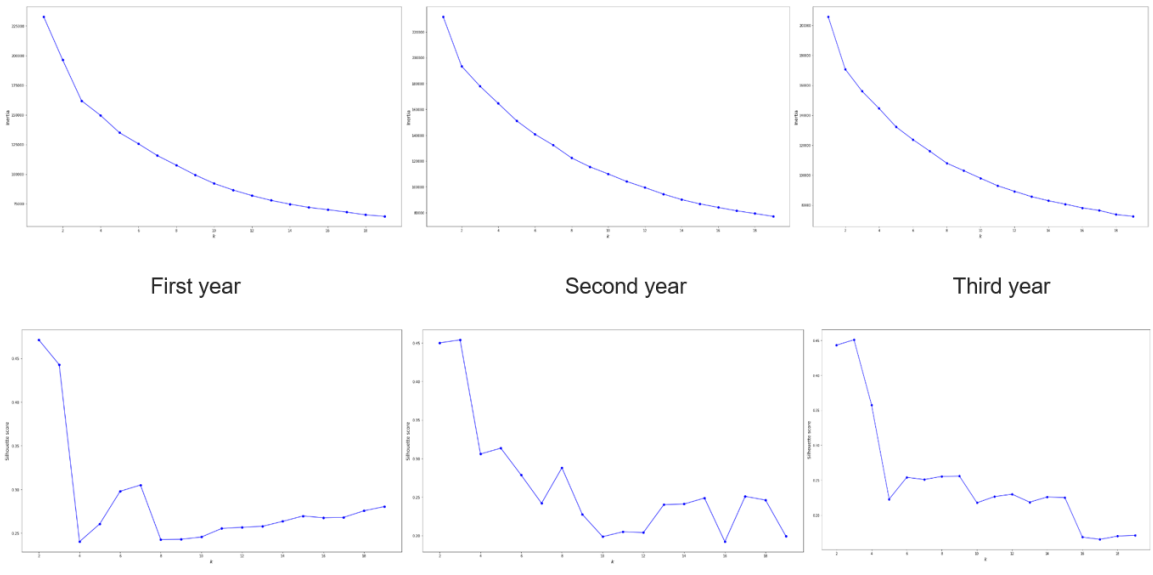


Figure 11: Elbow graphs and silhouette score graphs for each year of insurance

After running k-means for each year of insurance, the computation of the centroids, which represents each cluster, is made and enriched with other features, not used for clustering to better profile the clusters.

The insurance users are distinguished if they used the insurance within one year and ran the clustering for those who used it. If people cancelled in one year, a separate clustering is performed to see what kind of profile they are (Figure 12). Based on this approach, one can see the behaviour movement of people over time.

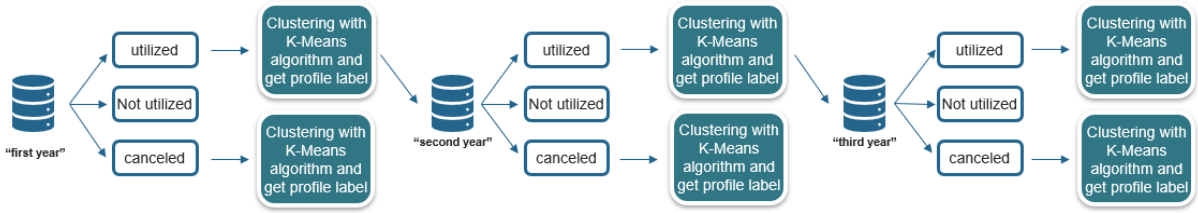


Figure 12: Clustering workflow for each year, including persons who cancelled

Even though it is an unsupervised learning technique without labels, cluster stability can be tested. For this reason, several iterations of clustering are performed to see whether the clusters are stable and instances are assigned to the same cluster.

Anomaly detection:

Anomaly detection is the core part of this work. As mentioned, the goal is to rank the daily incoming encounters by their probability of being an anomaly so that the auditing department can focus on the correct encounters. Figure 13 illustrates the entire process and model. For example, one encounter contains three entities: the insured person, the healthcare provider, and the encounter itself, containing all information.

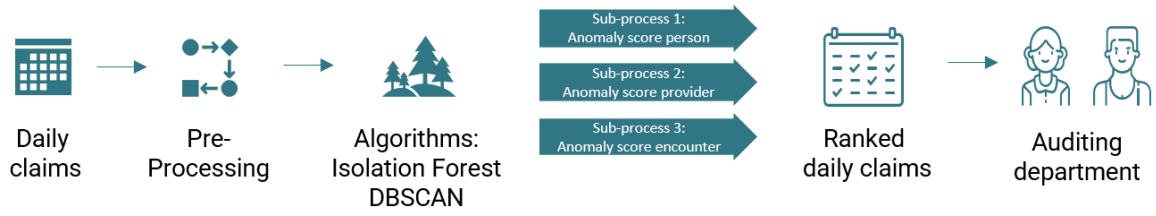


Figure 13: Anomaly detection process overview

The model's core is split into three sub-processes to compute an anomaly score for each, the person, the provider, and the encounter itself. Each sub-model is explained in the next section.

The algorithms used are isolation forest and DBSCAN (for persons) implemented with the python library scikit-learn. The reasoning for using both algorithms is illustrated in chapter 2.3, related work. Primarily, the algorithms are chosen due to Zhang (2020) and Zhong (2019), showing excellent results for iForest and considering DBSCAN to detect medical fraud and abuse. However, predominantly because of Zhang (2020) and the results for iForest to detect medical fraud and abuse. In addition, John & Naaz (2019) describe the advantages of iForest and its good results in the detection of credit card fraud. Even though credit card fraud is not related to medical fraud and abuse, the described modelling technique is applicable for fraud and anomaly detection in healthcare. Lastly, Zhong (2019) obtained good results in applying iForest twice to detect anomalies in gas turbines. Like credit card fraud, this domain can be applied to medical fraud and anomaly detection. The approach to combine

sub-process is based on Tang (2011), who achieved an excellent result with the ensemble model. For all sub-models, the contamination in isolation forest is 8 per cent.

The hyperparameters are, according to the main paper, the default settings:

- Number of estimators: 100,
- The maximum number of samples: 256 (F. T. Liu et al., 2008).

The anomaly score is the main output of all models. Unlike the original isolation forest paper, the scikit-learn implementation of isolation forest returns 0.5 minus the anomaly score. When the contamination parameter has a specific value, anomalies have a score below 0, which means that the decision threshold considering an instance as an anomaly is 0.

Regarding the general test design and model evaluation, a specific date is chosen, and model training is based on all encounters until this date. Due to the lack of labels, no automatic evaluation can be fulfilled. Samples of persons, providers, and encounters are examined by the business. More precisely, by the auditing department as well as clinical staff.

Sub-process 1 - Anomaly score person:

Computing the score and label for persons, isolation forest and DBSCAN are used to compare both algorithms performance. The input is the grouped data set with information up to the test date. The output is an anomaly score and labels for each person. In contrast, scikit-learn labels anomalies as -1, a mapping is applied to label anomalies with one and regular observations with 0.

Sub-process 2 - Anomaly score provider:

Providers differ according to their size and speciality. The output of this sub-process is an anomaly score for each provider, considering their specific characteristics. The following chart displays the calculation for each provider.

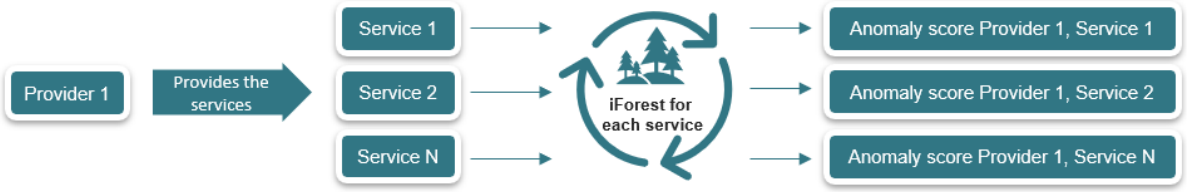


Figure 14: Anomaly score calculation for the healthcare provider

Provider network:

Besides calculating the anomaly score for providers within the sub-process, the following paragraph describes the approach to detect anomalous entities in a network of providers. For the creation, the open-source software for exploring and manipulating networks “gephi” is used. A network is an appropriate tool to visually explore connections and detect suspicious inactions based on specific filtering. Besides, this allows performing a complete network /graph analysis. One can compute various statistical measures of centrality, such as degree centrality, which computes how many edges (connections to other providers) one provider (node) has. However, the focus is not on the statistical

approach, instead of on the exploratory part. Based on the business goal to specialise in suspicious connections, the correct filtering of the variables within the network is vital.

A network design usually contains nodes that are connected with edges. These edges could either be directed or undirected. In this work, two main undirected network designs are applied that consider providers and encounters' anomaly scores and labels. The first network is an example of one specific medical service. The nodes are specific providers as well as patients, coloured by geographical location. The edges are encounters between a patient and provider that could either be anomalous (coloured in red) or normal. Force Atlas 2 is the layout algorithm used.

	Network 1	Network 2
Nodes	Specific providers and patients	All providers of all medical services
Node colours	Geographical location	Geographical location
Edges	Encounters (anomalous encounters are coloured in red)	Shared patients among the providers
Edge weight	uniform	Number of shared patients
Layout algorithm in gephi	Force Atlas 2	Open Ord

Table 2: Main characteristics of the two network designs

The second network design for exploratory purposes contains all providers as nodes coloured by geographical location. The edges are patients shared among the providers, where the edge weight is the number of shared patients. Since this includes all providers and is a complex network, the layout algorithm is "open ord" to increase the performance. Table 2 shows the main characteristics, although the tool allows changing according to the needs.

Sub-process 3 - Anomaly score encounter:

Like providers, the data set is split by their service associated with an isolation forest for each service. The first output of this sub-model is an anomaly score and the binary label for each encounter of the date chosen. The encounter anomaly score is the most important. To better understand the magnitude, the min-max normalisation score is normalised, obtaining values ranging from 0 to 1 in each service. The reception of high values close to one for more anomalous encounter, the normalised score is subtracted from 1, where 0 is absolutely no anomaly, and 1 means a high anomalous encounter.

$$s_{new} = 1 - \frac{s - s_{min}}{s_{max} - s_{min}} \quad (7)$$

As a result, each service will have encounters with an anomaly score of 1, which is the most anomalous encounter for one service and is vital to get a notion of the most anomalous encounters within one service.

Cancellation prediction based on anomalous behaviour:

After feature engineering and selection, the first step is to split the data into a training (70 per cent of the data) and test set (30 per cent). Then, a function compares the most popular classification

algorithms in default settings with each other. The list contains linear classifiers, support vector machines, kernel estimations, neural networks, decision trees, and so-called dummy classifiers. The dummy classifier gives a first notion of how well the default models perform with the given data. Two set-ups of the dummy classifier (scikit-learn implementation) are considered: “most_frequent”, which consistently predict the majority class, and “stratified”, which generates predictions by the distribution of the class. The entire list can be found in the appendix.

Around 10 per cent of the entire data set cancelled and get the label one. Based on this distribution, the prediction class is called imbalanced. As already mentioned, an imbalanced data set needs different treatments in terms of metrics. In addition, undersampling the majority class and oversampling the minority class increase the model performance. This model contains both under- and oversampling. At first, the majority class instances are reduced by about 40 per cent, followed by oversampling the minority class by about 50 per cent. The final distribution of the target is 60 per cent negative class and 40 per cent positive class, which turns out to get better performance.

Algorithms based on decision trees, such as random forest or XGBoost, are powerful algorithms. However, they tend to overfit. For that reason, plotting the learning curve (error and chosen evaluation metric) with many estimators (for instance, the number of trees in a random forest) helps choose the correct number of estimators. With this number fixed, the other hyperparameter can be improved by using the grid search optimisation algorithm. The following figure displays the process at a glance.

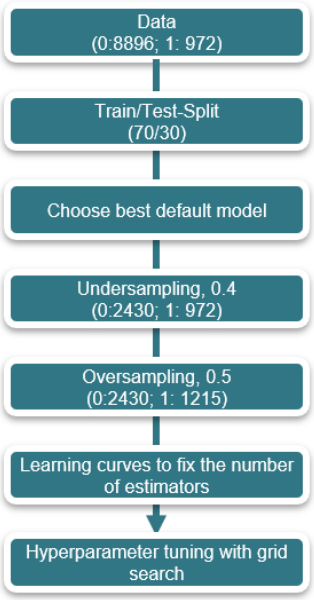


Figure 15: Process steps within the classification. Zero stands for the number of people with no cancellation and one for people who cancelled

When considering both the test and the training set, learning curves give insights into whether the model will over-or underfit. The following graphic shows an ideal learning curve plotting the model error. Although the error of the training set is still decreasing, the optimal number of estimators is the turning point (minimum) of the validation error (see dashed line), which is called the bias-variance trade-off point. If both lines run parallel without a drop in error, this implies too little data to learn.

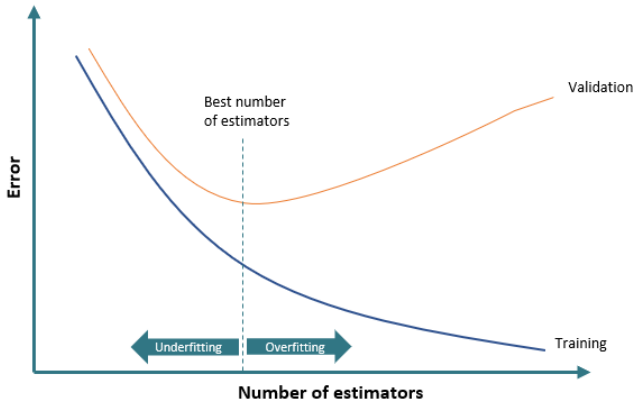


Figure 16: Idealised learning curve to explain the best number of estimators (Source: own elaboration based on (Géron & Safari, 2019)

In most real-life data set, however, the learning curve is not always ideal. The following figure shows the learning curves (error and the desired performance metric PR-AUC) for the best working model (XGBoost Classifier) with over-and undersampling for 2000 and 60 estimators. Based on this visual approach, the number of estimators chosen is 36. This number is a fixed hyperparameter to optimise the hyperparameter with the grid search algorithm.

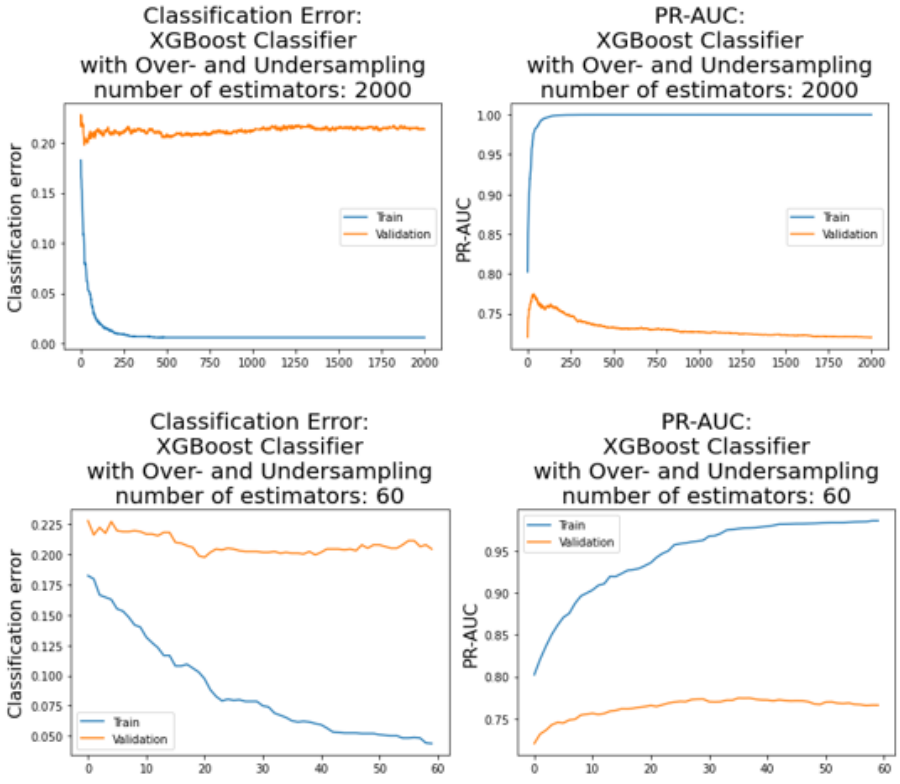


Figure 17: Learning curves (error and PR-AUC) for the best working algorithm (XGBoost Classifier) with over- and undersampling for 2000 and 60 estimators

4. RESULTS AND DISCUSSION

This chapter contains the results and evaluation of each model described in the previous chapters, which conforms with the evaluation phase in CRISP-DM. The results are split into the sub-chapters of clustering, anomaly detection, and cancellation prediction based on anomalous behaviour. In contrast, the anomaly part is further divided into its corresponding sub-processes.

4.1. CLUSTERING

The result for clustering is the profiling, or that is to say, the profiles of the insurance users within their first three years of insurance. The features used can be grouped into recency and general, frequency, and monetary, based on the RFM model, known from marketing analysis. The following radar plot (Figure 18) is an example for the second year of insurance and visualises all characteristics of each cluster. Their core points, called centroids, represent each cluster and could be either the mean or mode for all instances within one cluster, depending on the feature. The plot illustrates the variables used for clustering and additional features that are solely used for profiling.

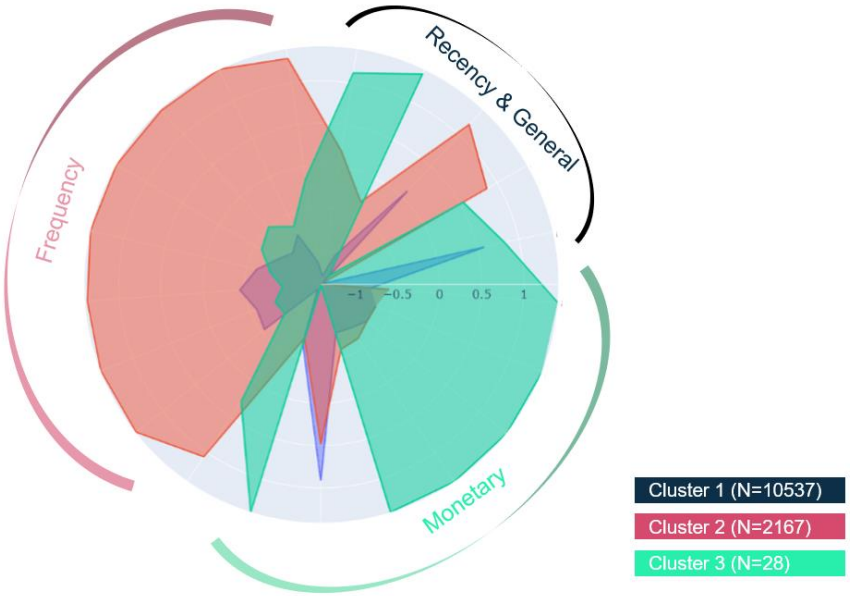


Figure 18: Radar plot to profile the insurance users' clusters

The clusters have evident peculiarities in each group of features, but mostly in frequency and monetary features. Based on this radar plot and the centroids representing the clusters, the three profiles can be described.

With around 80 to 82 per cent, cluster 1 represents the most frequent pattern of insurance users each year. They do not have outstanding characteristics to be associated with "regular" users. Cluster 2 is the second most frequent cluster and have striking characteristics within the group of frequency features. The smallest cluster each year is cluster 3, which are highly related to monetary features.

According to the distribution of the profiles among all years, the second and third years are similar. The first year, in contrast, contains significantly fewer instances of cluster 3 but more from cluster 1. This difference might be explained by the waiting period an insured person has.

	Total	Profile 1st year	Profile 2nd year	Profile 3rd year
Cluster 1	26112 insurance users	10537 (82.76 %)	10426 (80.91 %)	8773 (80.84 %)
Cluster 2		2167 (17.02 %)	2091 (16.23 %)	1790 (16.50 %)
Cluster 3		28 (0.22 %)	369 (2.86 %)	289 (2.66 %)

Table 3: Distribution of insurance users among each year of insurance

The number of insurance users per year differs because persons cancelled the insurance or had no encounter within one year.

Behaviour over time:

The reception of a better understanding of how people's behaviour changes over the time studied. The Sankey diagram gives a good overview and insights into the years' movements between clusters, cancellation, and no utilisation.

When insurance users had no usage in a year, they are more likely to be in cluster 1 in the following year. Vice versa, if one started being in cluster 1, the chance of not using the policy in the second and third years is higher. Within the cancellations of each year, many persons did not use the insurance in the previous year, and this number remained constant in the studied period. The proportion of the cluster profiles seems to be roughly the same whether people cancelled or not. However, none of the persons in cluster 3 during the first year withdrew in the following year.

Cluster 3 during the first year did not use the insurance in the following years, except for one person who moved to cluster 2. For both cluster 1 and cluster 2, a proportion of subjects keep their profile through all the years. Although this proportion is small, the absolute number, especially for cluster 2, can be relevant in the context of abuse detection.

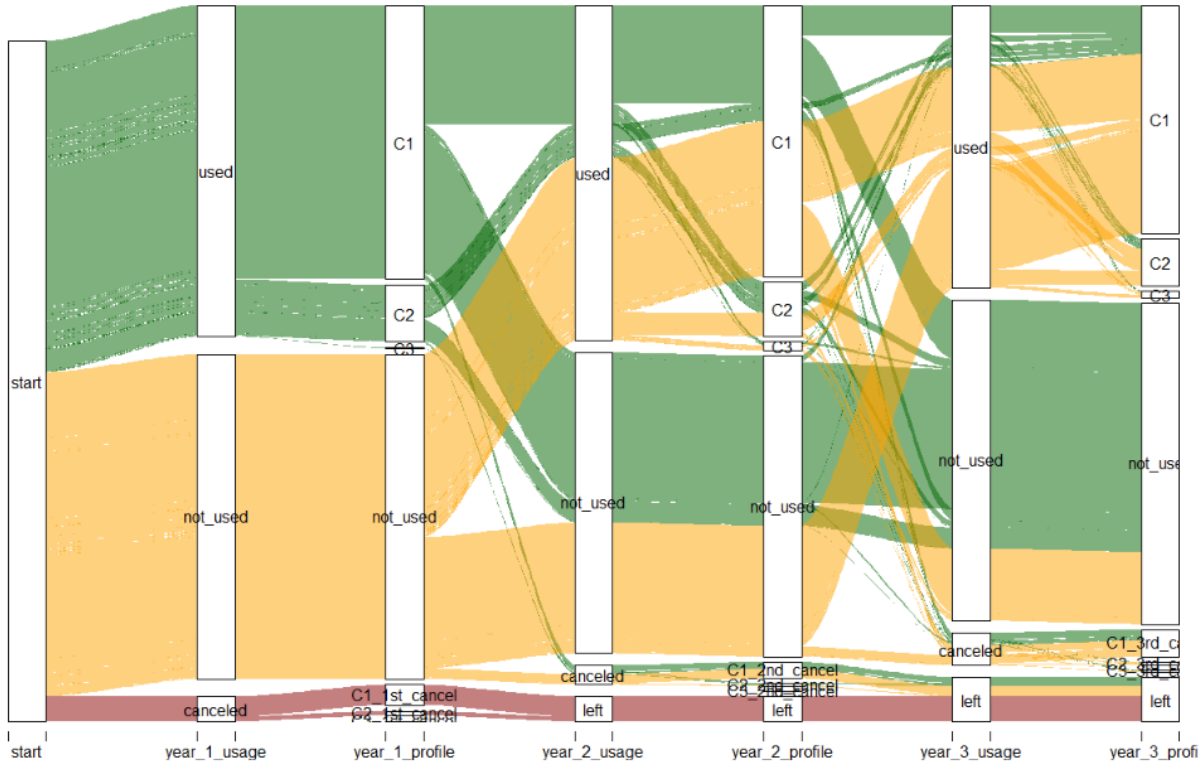


Figure 19: Sankey diagram to show the movement among the profiles of the insurance users in the first three years of insurance

Who cancelled in between?

Separate clustering analysis is performed to detect what kind of behaviour they had before their cancellation for those who cancelled each year. The first perception is that the clusters can also be distinguished into the exact characteristics of clusters 1,2, and 3 for the second and third years. Given that the characteristics of the centroids only consider cancelled insurance users, one can say that the second and third-year look like the characteristics of the insurance users who stayed in the policy in the corresponding years. Nevertheless, the pattern in the first year deviates. There is no clear pattern that is outstanding for cluster 3. It turns out that insurance users represent a cluster with characteristics from both cluster 2 and cluster 3. In other words, these insurance users spent and used a significant amount in the first year and cancelled their insurance afterwards.

Nevertheless, the second and third-year behaviour is comparable with the characteristics of those who did not cancel. It stands out that in the first two years, fewer people of cluster 3 cancelled the policy, whereas the share in the third year is higher than the years before.

4.2. ANOMALY DETECTION

The core results of this work are the results of the anomaly detection model and its underlying sub-models, which are discussed in this chapter. Due to the lack of labels, the testing of each model is based on manual check-ups by the auditing department based on provided, anonymised files.

Sub-process 1 - Anomaly score person:

The evaluation of both algorithms' performance requires the actual labels for each person, whether anomalous or normal. For this task, a sample of 90 instances is chosen after running an isolation forest for the first time. Due to this approach, samples are picked based on their anomaly score. To better understand the error the model makes, the exact size of samples (30 each group) is chosen with a high and low anomaly score and a score around the decision score of being anomalous or not. Every instance is examined by business experts and labelled as anomalous or normal without seeing the predicted label.

According to the literature review, two algorithms are implemented: DBSCAN and isolation forest. In terms of accuracy, it turns out that isolation forest (82 per cent accuracy) outweighs DBSCAN (60 per cent accuracy) for this sample (Figure 20).

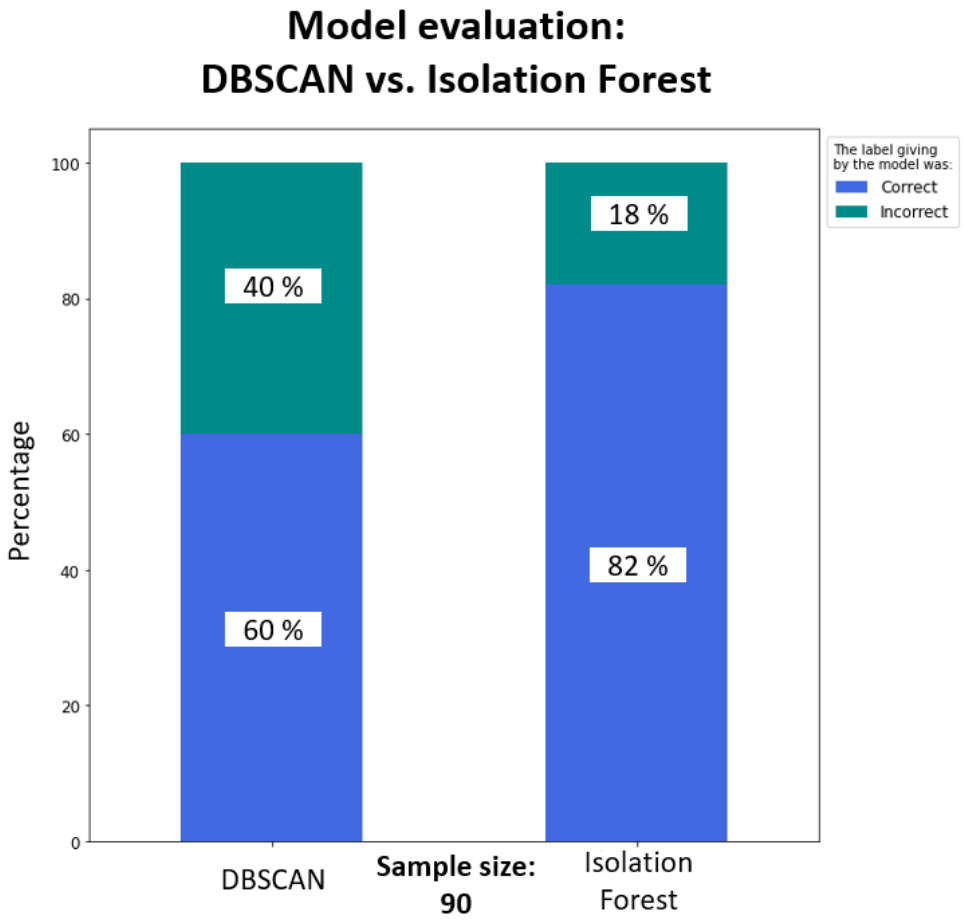


Figure 20: Model evaluation for persons: Isolation Forest vs. DBSCAN

A deeper examination of the actual results in a confusion matrix and the metrics for each algorithm (Figure 21 and Table 4) show that in other, isolation forest performs better than DBSCAN in all metrics.

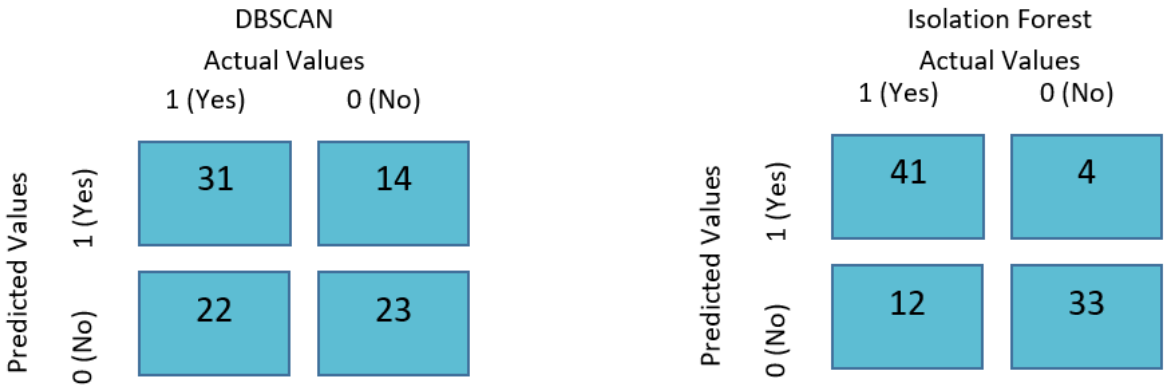


Figure 21: Confusion matrix for anomaly score person: DBSCAN (left) vs Isolation Forest (right)

Metrics	DBSCAN	Isolation Forest
Accuracy	0.60	**0.82**
Recall	0.58	**0.77**
Precision	0.68	**0.91**
F1	0.63	**0.83**
Specificity	0.62	**0.89**

Table 4: Metrics for anomaly score person: DBSCAN vs Isolation Forest

The results for DBSCAN are relatively balanced, around 0.6 in each metric, whereas the performance of isolation forest deviates more. It has higher precision (0.91) than recall (0.77), meaning it does a better job classifying anomalies than normal instances. This insight is reinforced with the analysis in which scenarios the model performs poorer. Figure 22 displays the distribution of correctly classified instances around each score. The accuracy in classifying instances with a high anomaly score, hence a higher probability of being an anomaly, is higher than for those with a low score (93 per cent to 87 per cent). When it comes to instances with an anomaly score around the decision threshold, meaning persons without an outstanding probability being either an anomaly or normal, the models perform worse (accuracy of 67 per cent). The manual adjustment of this decision threshold score can drive better results in detecting anomalous persons. This adjustment needs to be aligned with the business goal, whether it tends to have a higher precision (persons who are predicted anomalous) or recall (of all anomalies, how many were captured).

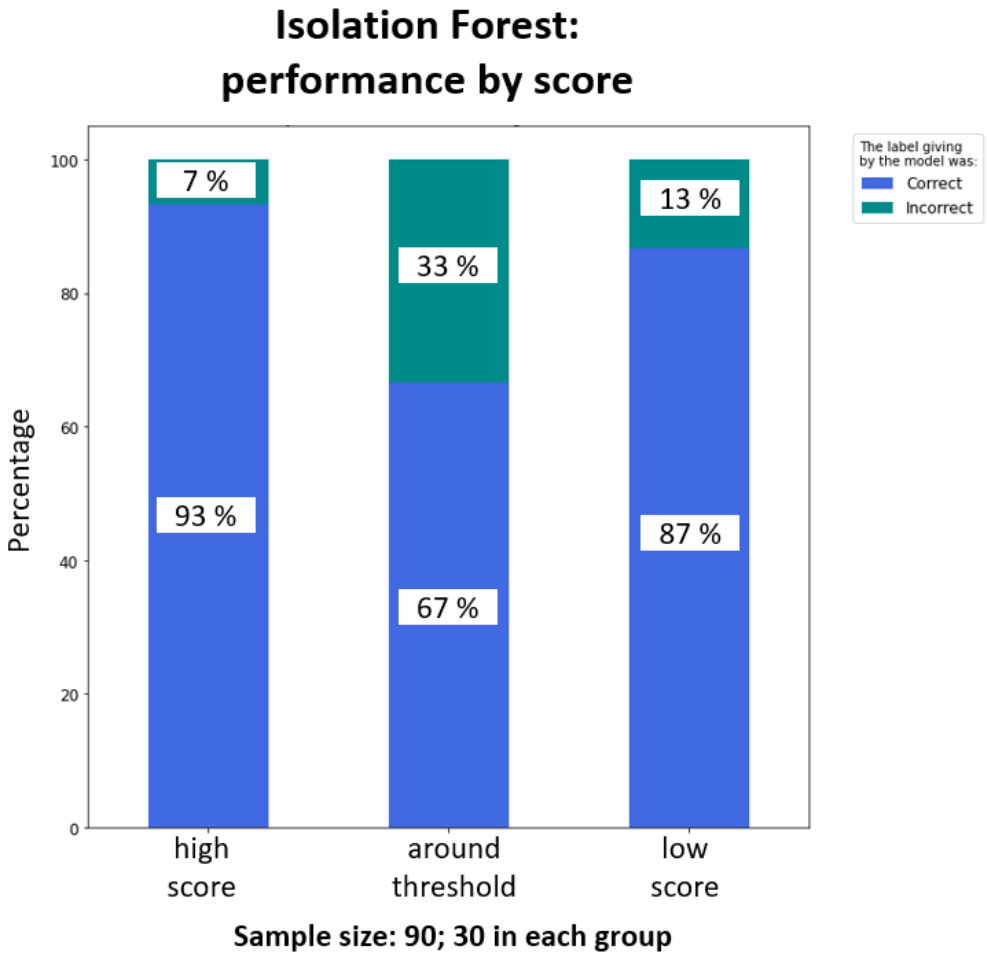


Figure 22: Error analysis isolation forest: evaluation around each anomaly score group

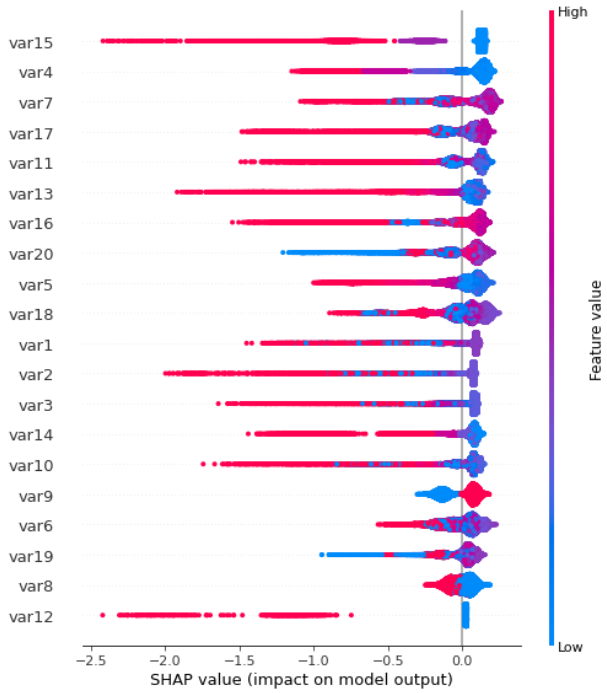


Figure 23: Feature importance for anomaly score persons with SHAP values

Overall features, the frequency features have the highest impact on the model output. Figure 23 displays the feature importance of this model (the most important feature is on the top). The lower the SHAP value, the higher the impact of being an anomaly. The more frequent certain features are (higher values are shown in red), the more likely the instance is classified as an anomaly.

Sub-process 2 - Anomaly score provider:

It is more challenging to define objective criteria to label a whole provider, especially when performing more than one service, as an anomaly. Because of that, a random sample is chosen to do a reasonable check for specific medical services.

Rather than calculating the metrics for this sub-process, the focus is on understanding which features plays a significant role to classify a provider as an anomaly. The connection between anomalous providers raises intelligence.

Like the persons' anomaly score model, a technique to get more insights into which features impact the model output. Figure 24 shows the 20 most important features that have an impact on the anomaly score. Another insightful method is the SHAP dependency plot, which illustrates interdependencies between chosen variables. The dependency plot in Figure 25, for instance, displays the connection between var2 and var3. One thing to point out is that a higher value of var2 significantly affects the anomaly score. In addition, providers with a high value in var2 and a high anomaly score have lower numbers in var3.

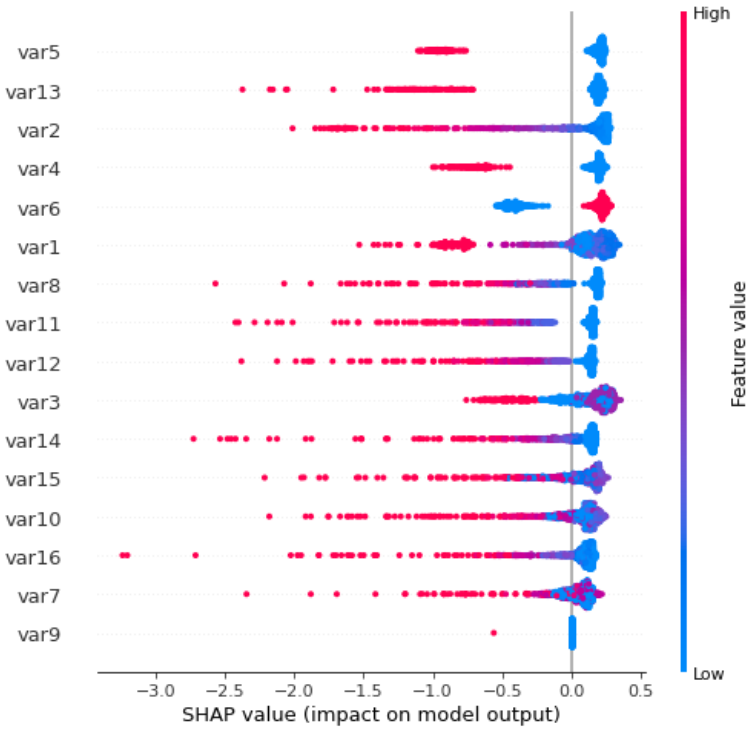


Figure 24: Feature importance for anomaly score provider with SHAP values (here: only service dental care)

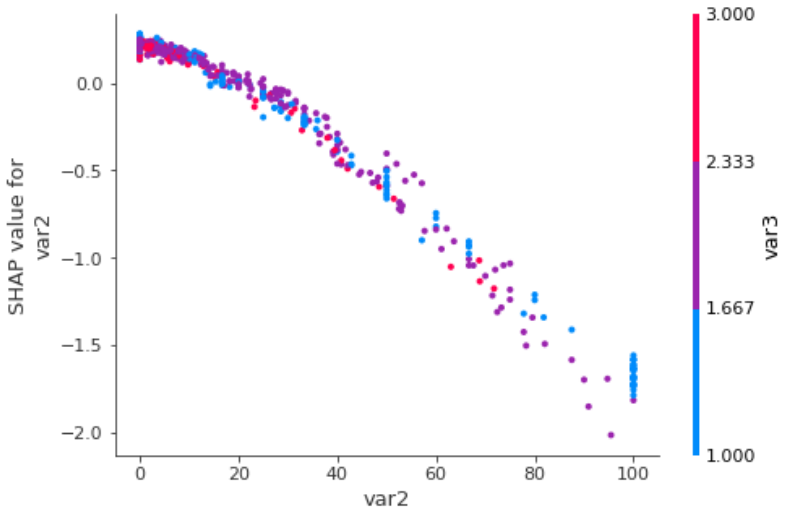


Figure 25: SHAP dependency plot for provider anomaly score

Provider network:

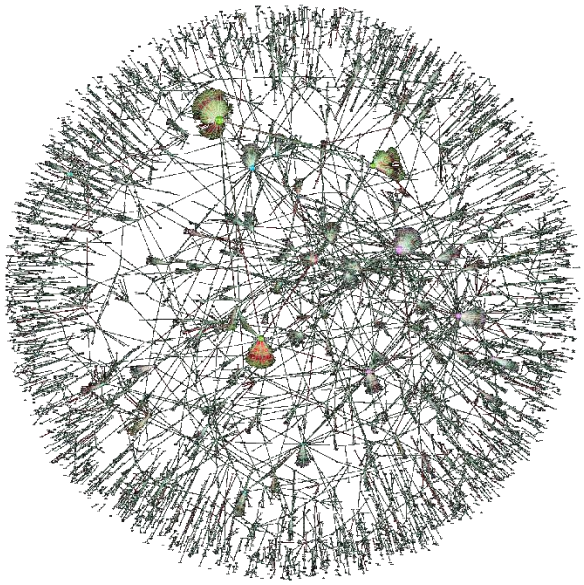


Figure 26: Overview network 1

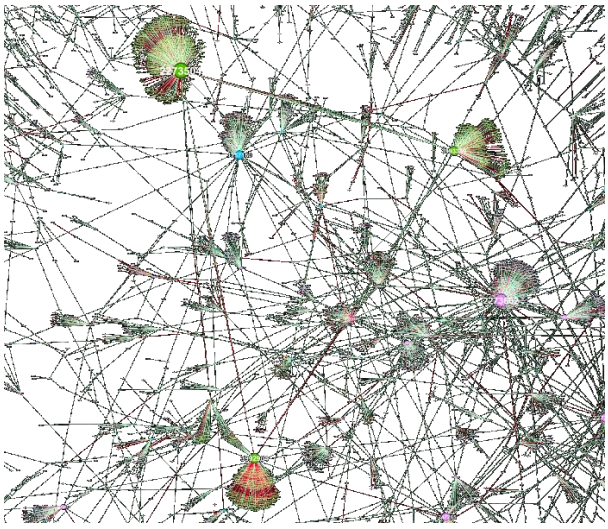


Figure 27: Enlargement of network 1

On the other hand, network 2 (Figure 28) is a network with all healthcare providers as nodes connected to the patients they have in common (edges). Unlike network 1, this network has an edge weight, represented by the number of patients the providers share. The nodes are coloured by location, and the seize stands for the anomaly score. A higher anomaly score means a greater node in the network. Although not recognisable in this figure, the edges are coloured

Based on the different network designs and layout algorithms, each network overview in gephi differs. Figure 26 shows network one, which connects specific providers and persons (nodes) by their encounters (edges). Anomalous encounters are coloured in red during their region colours providers and persons. Due to the layout algorithm, three connected providers strike with the same region and a significantly high number of anomalous encounters. Nodes at the outer edge have a less outstanding connection and many anomalous encounters.

An increase towards the centre spells a better view of how the providers and persons are connected Figure 27. It shows another underlying provider with a high rate of anomalous encounters. With the support of the software gephi and the creation of features, applying filtering based on the business needs to show only the nodes and edges of interest. For instance, a possible scenario would be to filter and display only providers with more than 20 per cent of anomalous encounters. Visually it is less challenging to detect the label of each. The guidance could be to do this analysis every once and focus on these providers for further investigations.

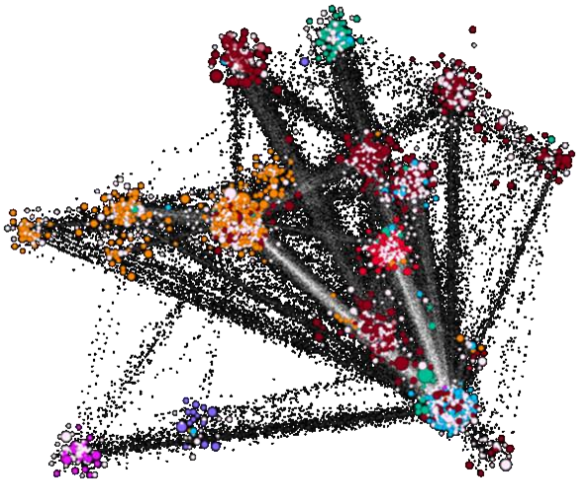


Figure 28: Overview network 2 (all providers)

by their weight in three groups: red edges are a high number of shared persons, orange means medium, and black are just one or a few. The layout algorithm shows the graph and grouped them by their similarities in the connection features and strength (edge weight). The overview shows that mostly the geographical location orders the clusters. An enlargement into specific areas points to those providers with a high anomaly score or many patients in common are close.

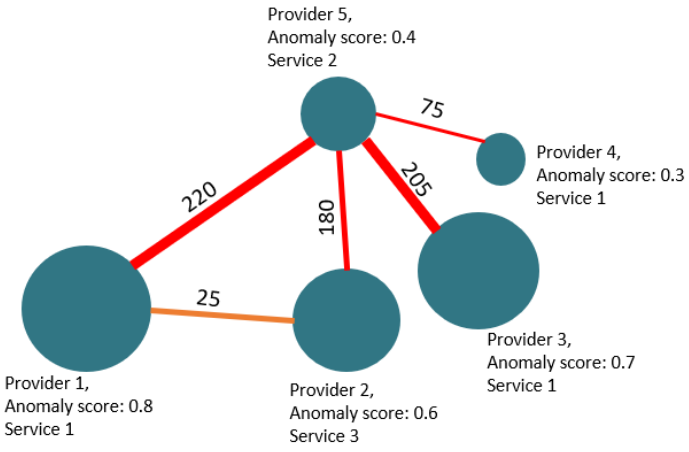


Figure 29: Example of a smaller suspicious network of providers

When applying filtering for an edge weight above 25 shared patients and an anomaly score of 0.3, a minor connection of providers occurs like in (Figure 29). Even though Provider 5 has no high anomaly score, it has strong connections with other providers with relatively high anomaly scores. Analysing the primary (mode) service they provide, most providers performing Service 1 are connected with Provider 5 that performs primarily Service 2. One possible inference would be that an encounter in Service 1 usually indicates Service 2. For example, an exam at one provider can cause a consultation with another provider to discuss the exam results. Another more abusive pattern can be that providers get a commission when sending their patients to Provider 5. As mentioned in the definition of anomaly and fraud, this does not mean undoubtedly fraud or abuse. However, it is a suspicious network to do further investigations. Based on the business needs, the filtering and the caption of the nodes and edges can be adjusted in the gephi software. The guidance is to use this visual analytics tool regular to define a focus when auditing providers to detect fraud and abusive patterns rather than implementing this daily.

Sub-process 3 - Anomaly score encounter:

Based on the business test, the performance varies depending on their specifics. As a first result, it can be said that building a generally applicable anomaly detection model for all encounters does not work well. Due to that complexity, an ensemble of models needs to be adjusted according to the specifics of the entities. The SHAP feature importance plot can be found in the appendix.

As mentioned, without labels, the testing task is cumbersome and binds human resources. Even with methods explaining the model, such as SHAP values, the interpretation is not always a simple task. In addition, anomalies and so fraud and abuse can be just minor differences to a regular observation. Fraudsters usually try to make the encounter appear normal. For that reason, the testing of this sub-process is an ongoing process.

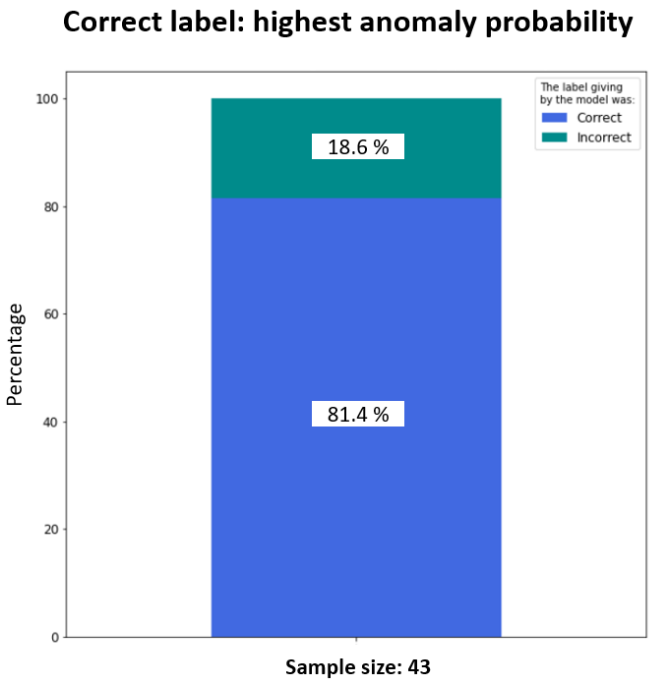


Figure 30: Evaluation of anomaly score encounters

Nevertheless, a sample of 43 is tested. In this test scenario, the encounters with the highest anomaly score for anomalous persons and providers are tested by business experts. However, the evaluation will continue after the generation of this work. The test design will take more instances into account that are also more distinguished to get a more balanced result.

Out of this sample, which contains a sample of all medical services, 81.4 per cent were labelled correctly. The lower the anomaly score for encounter, the worse is the performance. As guidance, it is vital to do continuous tests to get more data. Another advantage of manual testing and a detailed look into the data is the gain in knowledge. Business experts auditing encounters in their day-to-day business can detect new patterns based on the model that might be fraudulent or abusive. These patterns might point out some findings, as, for example, exams where more tests are performed than in comparable conditions and more than medically necessary.

4.3. CANCELLATION PREDICTION BASED ON ANOMALOUS BEHAVIOR

Often, cancellation prediction models suffer due to an imbalanced target class. Hence, only considering accuracy as the primary evaluation metric can skew the results. Usually, the negative class (no cancellation) represents the vast majority. The initial data set before over- and undersampling contain around 10 per cent of the positive class (cancellation). A dummy classifier, which always predicts the majority class, will achieve an accuracy of 0.90 without fulfilling the actual purpose nowhere near enough. For that reason, the metrics precision, recall, f1, and PR-AUC are the primary evaluation metrics.

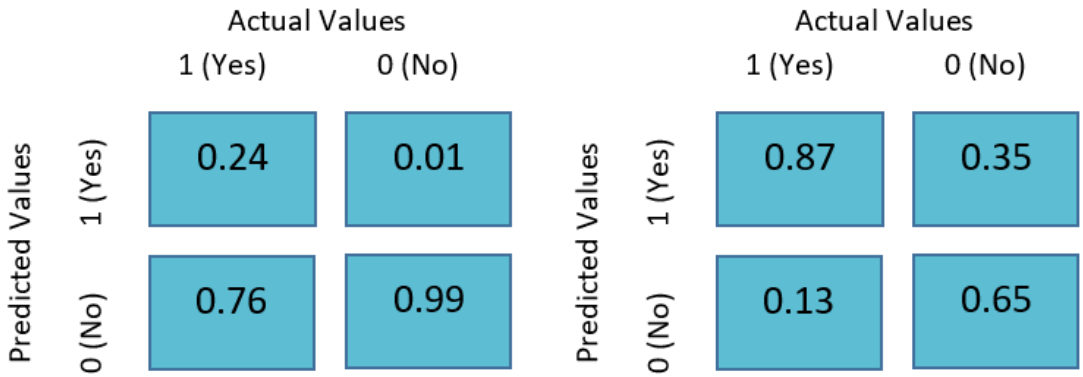


Figure 31: Confusion matrix normalised (validation set), without (left) and with (right) over-and undersampling

To overcome the imbalance of the target class, over-and undersampling is performed. Comparing the results in a confusion matrix (Figure 28), over and undersampling the target class has a better performance. Besides accuracy, which is not an appropriate metric, in this case, over-and undersampling has better results (Table 5).

Metrics (validation set)	Without over- and undersampling	With over- and undersampling
Accuracy	**0.92**	0.80
Recall	0.24	**0.65**
Precision	0.71	**0.72**
F1	0.36	**0.68**
PR-AUC	0.49	**0.77**

Table 5: Comparing metrics between a model with and without over- and undersampling

The area under the precision-recall curve (PR-AUC) is, like the more commonly used ROC-AUC, a score to evaluate whether the model is skilful or not skilful. The PR curve is a line that plots the recall (x-axis) and the precision (y-axis) for different probability thresholds. In a ROC curve, plotting a diagonal line (equivalent to random guessing) represent the threshold if the classifier is skilful or not. However, for PR curves, this threshold is a horizontal line with a precision proportional to positive instances (positive class). In this dataset, a threshold value of 0.10 was applied for the model without sampling transformations and 0.33 for the model with over-and undersampling.

Based on this definition, both models are skilful. The PR curve in Figure 32 plots the values for different thresholds and shows the trade-off between precision and recall. Even though the precision can be improved by adjusting the threshold, it causes a decrease in recall and vice-versa. The adjustment is not only a model-driven decision. Instead, it is a business decision since it depends on the business strategy. Higher precision in the cancellation prediction means more people who are predicted positive are positive.

On the other hand, a higher recall means more people who will cancel are captured. Nevertheless, more people predicted positive who will not cancel. Based on this interaction, a strategy to adjust the threshold needs to be developed.

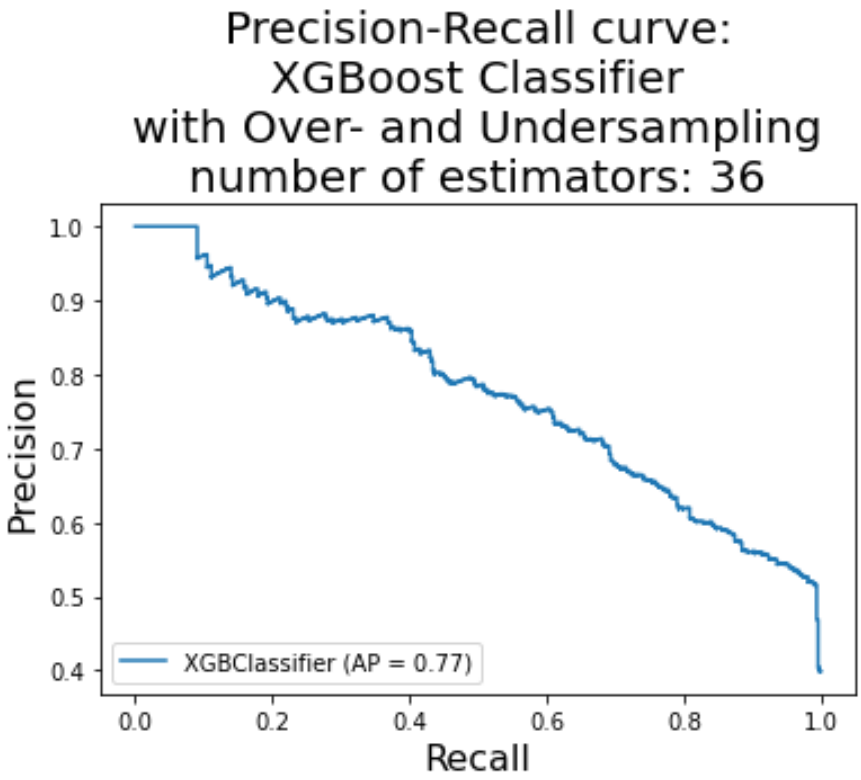


Figure 32: Precision-Recall-Curve

According to the goal to predict after three months whether insurance users cancel the policy, it can be said that it is generally possible. However, to obtain better results, more data is needed. More personal data is known in advance, or the prediction time is closer to the end of the first year of insurance.

5. CONCLUSIONS

The literature review shows several supervised and unsupervised techniques to detect fraud and anomalies. This work combines unsupervised methods and one supervised approach applied on healthcare encounters: clustering, an ensemble of anomaly detection models, and a predictive model. Although evaluation in unsupervised learning is more challenging and binds more human expertise, not having labels opens a new opportunity to detect new fraudulent patterns. In contrast, this might not be the case when using a supervised learning approach.

An initial clustering is a suitable method to get more knowledge about the behaviour of the entities, such as providers and insurance users. It gives the first notion of how these entities behave and clusters further apart from others that might be anomalies. A mixture of hierarchical clustering and the k-means algorithm is easy to implement and fast solution. However, as a complete and independent fraud- and anomaly detection model, solely using clustering is too general. The implementation of other clustering algorithms, for instance, DBSCAN, might improve the model. The variety of medical encounters is too high to implement clustering as a single tool. However, it is a vital step to understand the behaviour.

The core part of this work is an anomaly detection model. This model ensemble of different scores and labels connects all instances (provider, insurance user, and the actual encounter) and examines one encounter solely. In healthcare, the behaviour and consumption vary where factors such as medical service, type of provider, or type of insurance policy impact whether an encounter is anomalous or not. These circumstances prevent a so-called one-fits-all solution and need to be differentiated in more detail. On the other hand, too granular data means less data for model training, which leads to issues to separate anomalies. This unsupervised approach is a suitable add-on to static cost containment rules. However, it is crucial to further test and continuously update and evaluate the model. Due to the requirement of expert knowledge for evaluation, more testing, especially in the anomaly score for encounter and the final sorting, is necessary to obtain a better performance. Overall, this model shows good results, and with future improvements, it is ready to test in deployment.

Although not an unsupervised machine learning technique, the predictive model can hint at who is more likely to cancel the insurance within the first year. Nevertheless, the rate of false positives and false negatives are relatively high to deploy fully or to use as a tool to detect fraudsters and abusive people in advance. For this reason, more data, such as a medical questionnaire or socio-demographic data, is needed. Also, before further improving this model, the business actions that follow its results need to be defined. For example, a more promising model can be a model to predict churn for loyalty campaigns rather than use it for fraud detection.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

No labels cause the exclusion of supervised learning techniques, which is one of the main limitations to detect fraud and anomalies. Although unsupervised learning approaches helps to detect new patterns, the evaluation binds more resources to get the ground truth. For these reasons, the test samples are relatively small concerning the entire data set. However, the test and evaluation is not terminated and will continue after this work. This project is a project from scratch without knowing the data and established data science processes and infrastructure. With general fraud and anomaly detection, the project goal is rather broad, and the development of pipelines to do the proof of concept takes time. Although good expertise is present, the business knowledge is spread and obtained with a time offset.

Consequently, this can require changes even in the first process steps, such as data preparation. Eventually, this leads to a longer cycle time. Encoding features with high cardinality is computationally expensive. It is highly recommended that the deployment use a big data environment or cloud computing services such as AWS by Amazon or Microsoft Azure.

One of the main recommendations and improvements for future works is the implementation of a supervised learning pipeline. The purple lines in Figure 33 show a process draft. Whenever the auditing department sees a ranked encounter, it will label them as fraud, no fraud, or anomaly. Then, a supervised learning model can be implemented based on these labels, giving additional input to rank the daily incoming encounters.

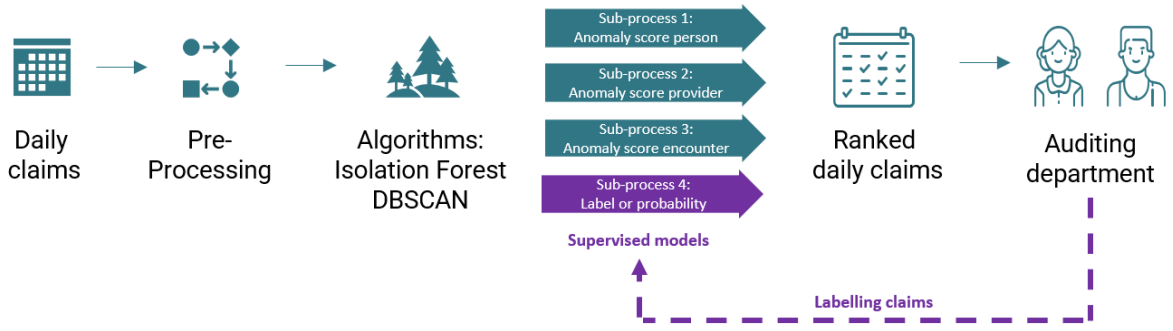


Figure 33: Reinforcement learning as a recommendation for future works in anomaly detection, marked in purple

One of the main improvements for this work is extending and finalising the test scenario to obtain more real labels. With these labels, one can evaluate used models more accurate. Based on these adjustments, further models, such as SA-iForest in Xu (2017) and the two steps iForest, are proposed in Zhong (2019).

The implementation of SHAP values for single observations (local interpretability) to explain which features impact the observation to be normal or anomalous would help the auditing department understand the predicted label.

Lastly, other anomaly detection models, such as Gaussian mixture models or deep learning models, can be implemented and tested. According to Naiboo (2020), GANs can be developed to improve the supervised models with some labels.

7. BIBLIOGRAPHY

Abdulraheem, A., Abdul-Hadi, J., Akintola, A., & Ameen, A. O. (2015). Anomaly Detection in Dataset for Improved Model Accuracy Using DBSCAN Clustering Algorithm. *African Journal of Computing & ICT*, *Vol 8. No.1*

Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: An Open Source Software for Exploring and Manipulating Networks*. https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag.

Carvalho, L. F. M., Teixeira, C. H. C., Meira, W., Ester, M., Carvalho, O., & Brandao, M. H. (2017). Provider-Consumer Anomaly Detection for Healthcare Systems. *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 229–238. https://doi.org/10.1109/ICHI.2017.75

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Cross Industry Standard Process for Data Mining. In *CRISP-DM Consortium*.

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Z., & Li, Y. F. (2011). Anomaly Detection Based on Enhanced DBScan Algorithm. *Procedia Engineering*, *15*, 178–182. https://doi.org/10.1016/j.proeng.2011.08.036

Davis, J., & Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. *Proceedings of the 23rd International Conference on Machine Learning, ACM*, *06*. https://doi.org/10.1145/1143844.1143874

Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, *74*, 406–421. https://doi.org/10.1016/j.patcog.2017.09.037

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231.

Géron, A., & Safari, an O. M. C. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*. O'Reilly Media, Incorporated. https://books.google.pt/books?id=O2VJzQEACAAJ

Giudici, P., & Figini, S. (2009). *Applied Data Mining for Business and Industry* (2nd ed.). Wiley Publishing.

Goldstein, M., & Uchida, S. (2016). A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLOS ONE*, *11*(4), e0152173. https://doi.org/10.1371/journal.pone.0152173

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

John, H., & Naaz, S. (2019). Credit Card Fraud Detection using Local Outlier Factor and Isolation Forest. *International Journal of Computer Sciences and Engineering*, *7*(4), 1060–1064. https://doi.org/10.26438/ijcse/v7i4.10601064

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2014). Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature. *Global Journal of Health Science*, *7*(1). https://doi.org/10.5539/gjhs.v7n1p194

Kelleher, J. D., Namee, B. Mac, & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. The MIT Press.

Kirlidog, M., & Asuk, C. (2012). A Fraud Detection Approach with Data Mining in Health Insurance. *Procedia - Social and Behavioral Sciences*, *62*, 989–994. https://doi.org/10.1016/j.sbspro.2012.09.168

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & Team, J. D. (2016). Jupyter Notebooks - a publishing format for reproducible computational workflows. *ELPUB*.

Li, Y., & Wu, H. (2012). A Clustering Method Based on K-Means Algorithm. *Physics Procedia*, *25*, 1104–1109. https://doi.org/10.1016/j.phpro.2012.03.206

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17

Liu, J., Bier, E., Wilson, A., Guerra Gómez, J. A., Honda, T., Sricharan, K., Gilpin, L., & Davies, D. (2016). Graph Analysis for Detecting Fraud, Waste, and Abuse in Health-Care Data. *Ai Magazine*, *37*, 33–46.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2019). *Explainable AI for Trees: From Local Explanations to Global Understanding*.

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*.

Naidoo, K., & Marivate, V. (2020). *Unsupervised Anomaly Detection of Healthcare Providers Using Generative Adversarial Networks* (pp. 419–430). https://doi.org/10.1007/978-3-030-44999-5_35

Ogbuabor, G., & F. N, U. (2018). Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value. *International Journal of Computer Science and Information Technology*, *10*(2), 27–37. https://doi.org/10.5121/ijcsit.2018.10203

Ounacer, S., el Bour, H., Oubrahim, Y., Ghoumari, M., & Azzouazi, M. (2018). Using Isolation Forest in anomaly detection: the case of credit card transactions. *Periodicals of Engineering and Natural Sciences (PEN)*, *6*, 394. https://doi.org/10.21533/pen.v6i2.533

Pang, G., Shen, C., Cao, L., & Hengel, A. Van Den. (2021). Deep Learning for Anomaly Detection. *ACM Computing Surveys*, *54*(2), 1–38. https://doi.org/10.1145/3439950

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pourhabibi, T., Ong, K.-L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature

review of graph-based anomaly detection approaches. *Decision Support Systems*, *133*, 113303. https://doi.org/10.1016/j.dss.2020.113303

Powers, D. M. W. (2015). Visualization of Tradeoff in Evaluation: from Precision-Recall {\&} {PN} to LIFT, {ROC} {\&} {BIRD}. *CoRR*, *abs/1505.0*. http://arxiv.org/abs/1505.00401

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking* (1st ed.). O'Reilly Media, Inc.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Ram, A., Jalal, S., Jalal, A. S., & Kumar, M. (2010). A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases. *International Journal of Computer Applications*, *3*(6), 1–4. https://doi.org/10.5120/739-1038

Samriya, J. (2016). *EFFICIENT K-MEANS CLUSTERING FOR HEALTHCARE DATA*. 2393–8390.

Tang, M., Mendis, B., Murray, D., Hu, Y., & Sutinen, A. (2011). Unsupervised fraud detection in Medicare Australia. *Conferences in Research and Practice in Information Technology Series*, *121*, 103–110.

Thornton, D., van Capelleveen, G., Poel, M., Hillegersberg, J., & Mueller, R. (2014). 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.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

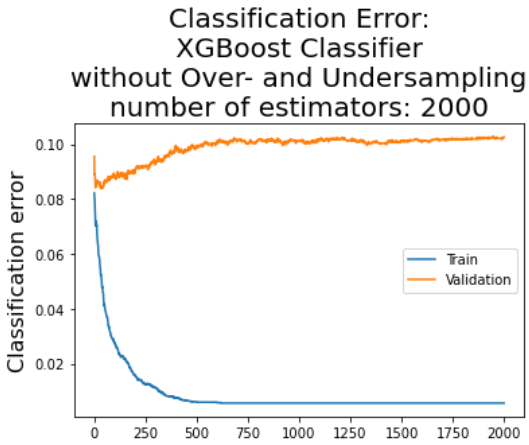
Xu, D., Wang, Y., Meng, Y., & Zhang, Z. (2017). An Improved Data Anomaly Detection Method Based on Isolation Forest. *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, 287–291. https://doi.org/10.1109/ISCID.2017.202

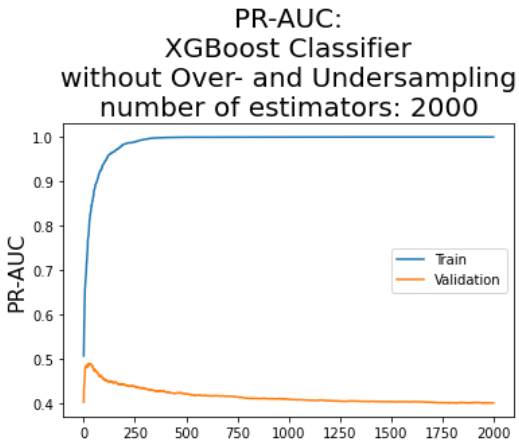
Zhang, C., Xiao, X., & Wu, C. (2020). Medical Fraud and Abuse Detection System Based on Machine Learning. *International Journal of Environmental Research and Public Health*, *17*(19), 7265. https://doi.org/10.3390/ijerph17197265

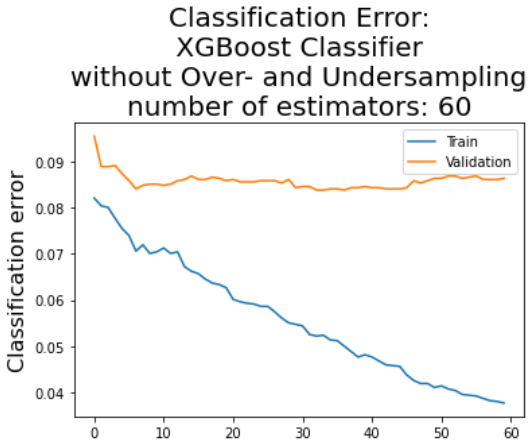
Zhong, S., Fu, S., Lin, L., Fu, X., Cui, Z., & Wang, R. (2019). *A novel unsupervised anomaly detection for gas turbine using Isolation Forest*. 1–6. https://doi.org/10.1109/ICPHM.2019.8819409

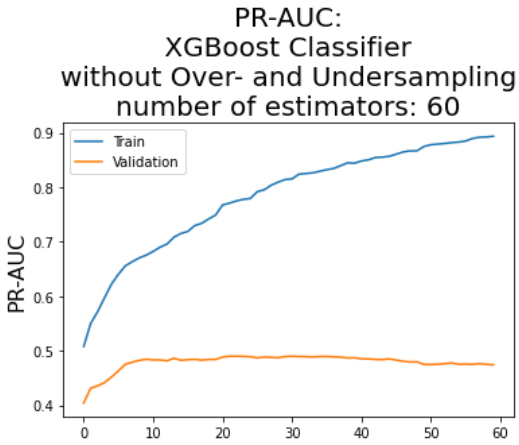
8. APPENDIX

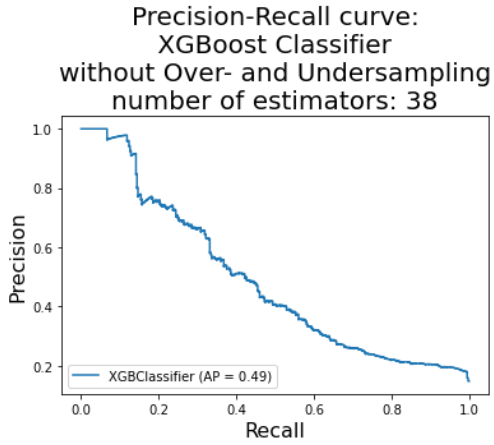
Cancellation Prediction:







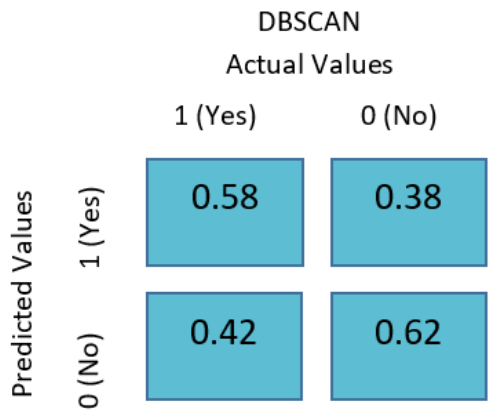


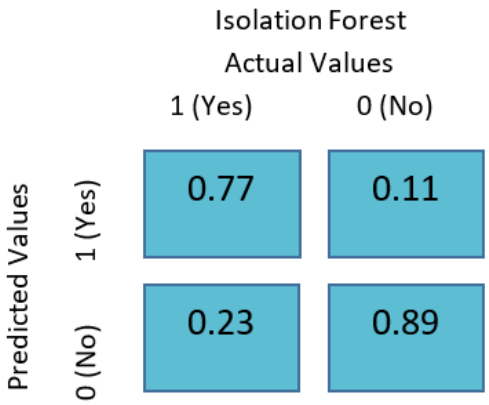


	model_used	recall_train	precision_train	f1_train	recall_validation	precision_validation	f1_validation	pr_auc_validation
6	XGBClassifier(base_score=0.5, booster='gbtree'...	0.83	1.00	0.91	0.30	0.60	0.40	0.56
8	DecisionTreeClassifier(random_state=0)	1.00	1.00	1.00	0.39	0.35	0.37	0.46
7	AdaBoostClassifier(random_state=0)	0.23	0.70	0.35	0.20	0.58	0.30	0.54
2	RandomForestClassifier(random_state=0)	1.00	1.00	1.00	0.17	0.61	0.26	0.50
4	GaussianNB()	0.99	0.11	0.20	0.97	0.11	0.20	0.54
0	MLPClassifier(random_state=0)	0.33	0.96	0.49	0.08	0.30	0.12	0.23
3	KNeighborsClassifier()	0.17	0.67	0.27	0.07	0.31	0.12	0.26
9	DummyClassifier(random_state=0, strategy='stra...	0.11	0.11	0.11	0.12	0.11	0.12	0.21
1	LogisticRegression(random_state=0)	0.01	0.60	0.02	0.00	0.07	0.01	0.21
5	SVC(probability=True, random_state=0)	0.04	0.99	0.08	0.00	0.03	0.00	0.17
10	DummyClassifier(random_state=0, strategy='most...	0.00	0.00	0.00	0.00	0.00	0.00	0.55

Anomaly Persons:

Relative numbers confusion matrix for anomaly score persons:





Anomaly Encounters:

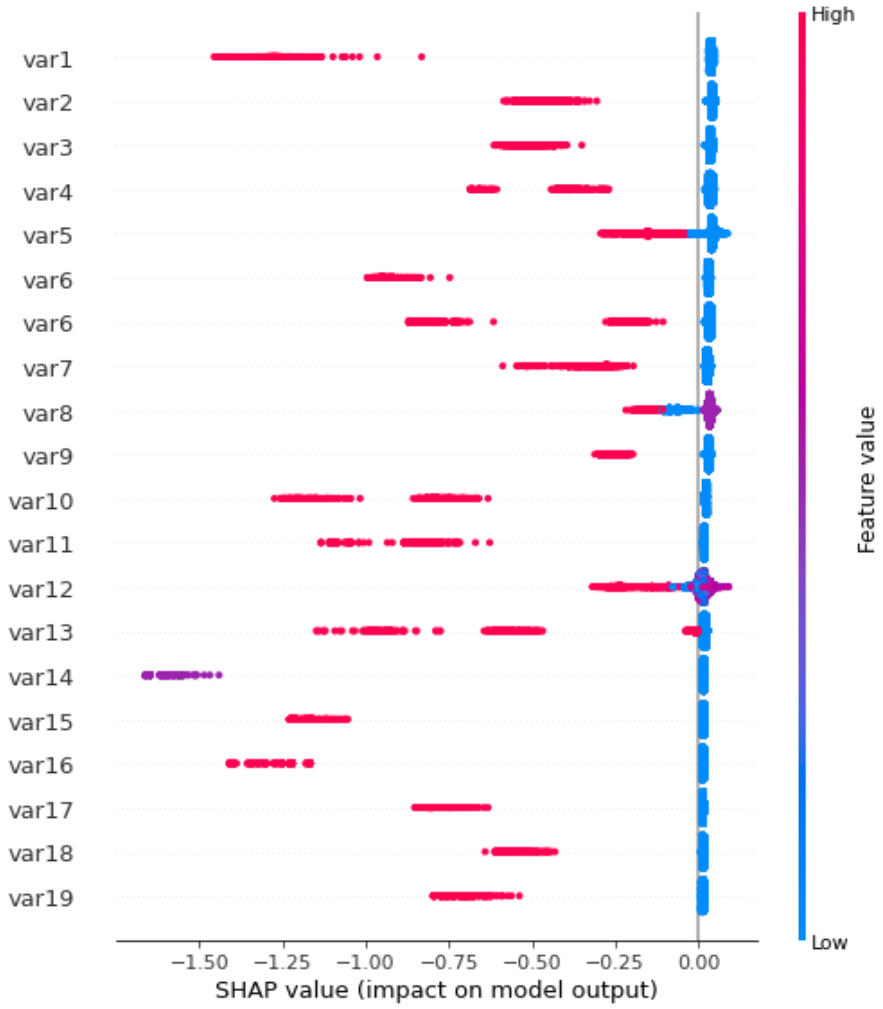


Figure 34: Feature importance for anomaly score encounters (here: service dental care) with SHAP values