Detecting and Explaining Potential Financial Fraud Cases in Invoice Data with Machine Learning

Lieke Hamelers

A thesis presented for the degree of Business & IT with the specialisation Data Science & Business

Auditdienst Rijk

Marije den Ouden

Vincent Gorka

UNIVERSITY OF TWENTE.

University of Twente

Dr. M. (Mannes) Poel, EEMCS

Dr. ir. G. (Guido) van Capelleveen, BMS

Prof.Dr. J. (Jos) van Hillegersberg, BMS

January 16, 2021

Summary

The financial losses due to fraud keep increasing over the years. A possibility for minimising errors and potential fraud is auditing. Auditors review a sample of invoices to give an indication of the quality of the entire population. This process is time-consuming and could be optimised. Outlier detection is an area that has already shown promise in fraud detection. Different algorithms can identify outlying invoices and mark these as potential fraud. However, it is important that these algorithms are transparent and thrust worthy. Auditors are accountable for their decisions and cannot make decisions based on a black box algorithm. To enable auditors to use outlier detection, some explanation mechanisms need to be used. Little research has been done on explanation mechanisms in the financial sector.

The goal of this research is to investigate how financial auditors can use machine learning in a legitimate manner in their work. This contribution of this research is five fold and each result is described separately below. First, an overview of the state of the art research to unsupervised outlier detection algorithms and explanation mechanisms in the financial sector is provided. A literature review was performed to analyse the current body of literature and an assessment matrix was created to analyse the different combinations of unsupervised outlier detection algorithms and explanation mechanisms. Based on this matrix, it was decided to include an Isolation Forest, Local Outlier Factor and an One Class Support Vector Machine in this research.

Second, this research provides an indication of which features are useful for detecting potential fraud. Based on academic and professional literature, a survey was distributed among auditors. The results of this survey show which features are useful for detecting errors and potential fraud. It is found that the highest scoring features are often based on information about the business partner.

Third, this research compared three different unsupervised outlier detection algorithms that are applied on invoice data from the public sector. The performance of the three algorithms are compared by using a small test set of identified outliers. The results show that the Isolation Forest outperforms the other two. Next, the predictions of the Isolation Forest were manually labelled by three auditors to see its performance. The Isolation Forest was able to correctly identify 72% of the invoices according to the auditors.

Fourth, this research has designed an explanation facility for the financial auditors. The facility is a web application that provides explanations on the input data, model and output of the algorithm. The facility was designed according to the design science methodology in two iterations and validated through usability tests with multiple financial auditors. The

feedback on the facility was positive and all indicated that they would use the facility again.

Finally, this research has identified how the outlier detection algorithms can contribute to the work of the financial auditors and where in the process it can be used. It was found that the algorithm can be used when making identifying the risks for the coming year, during the process of reviewing of the invoices or afterwards to identify and correct errors.

Acknowledgement

Dear reader,

With this thesis, I will finish my master Business&IT and with that my time as a student at the University of Twente. I could not have imagined before that i would have to carry out my final project from home. Not being able to meet my supervisors in person, or only visiting the Ministry three times. However, despite these limitations, I was able to execute my research and I am proud of the end result.

First, I would like to thank the Dutch Central Governmental Auditing Service for enabling me to do the research within their department. The willingness of everyone to help me conduct my research was very pleasant. Above all, I would like to thank my two supervisors, Marije den Ouden and Vincent Gorka that helped me out with a lot of practical issues. From finding people that are willing to participate in my research to helping me deploy my prototype on the server. I really appreciated your effort of trying to involve me as much as possible in the team while working from home.

Next, I want to express my gratitude to my supervisors from the University of Twente: Guido van Capelleveen, Mannes Poel and Jos van Hillegersberg. In the last six months, all our communication has been online but the meetings have helped me to improve this research significantly. Due to your supervision, ideas and input, I have been able to focus on novel and relevant topics.

Finally, I would like to thank all the participants that took part in this research. Without your input and feedback, I would not have been able to produce these results.

The efforts of the last six months have resulted in this thesis that lies before you. I hope you will read this with joy and discover something new from it.

Lieke Hamelers

Contents

List of Figures

List of Tables

Nomenclature

Auditor	A specialist that executes the audit.
DCGAS	The Dutch Central Governmental Auditing Service which acts as the internal auditor of the Dutch Government.
Explanation mechanism	A method that can provide insight or give an explanation about an algorithm or their outcomes.
Financial audit	An independent, objective evaluation of an organisation's financial reports and financial reporting processes to give the assurance of a correct and complete financial statement.
IF	Isolation Forest; an unsupervised outlier detection algorithm that isolates outliers by building many decision trees.
Internal controls	The plan of the organisation and all the co-ordinate methods and measures adopted within a business to safeguard its assets, check the accuracy and reliability of its accounting data
Invoice	A time-stamped commercial document that itemises and records a transaction between a buyer and a seller.
LOF	Local Outlier Factor; an unsupervised outlier detection algorithm that detects outliers by clustering data points and reviewing distant data points from clusters.
OCSVM	One Class Support Vector Machine; an unsupervised outlier detection algorithm that detects outliers by creating a hyper plane that encompasses all normal data points.
Outlier detection algorithm	A piece of software that identifies rare observations which raise suspicions by differing significantly from the majority of the data. Also called anomaly detection.
SHAP	(SHapley Additive exPlanations; a game theoretic approach to explain the output of any machine learning model

1 Introduction

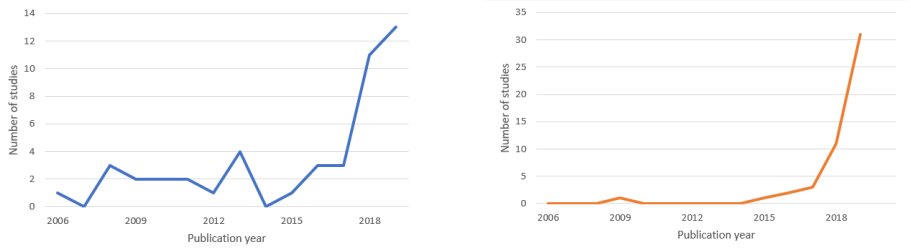
1.1 Problem description

Fraud in the financial world is an important theme and the losses involved in fraud keep increasing. A recent report of KPMG has indicated that in 2019 fraud has reached over one billion British pounds [120]. Auditing is the process for reviewing systems and processes to minimise the possibilities of errors and potential fraud. Auditors usually review samples of the complete set of transactions to give an indication about the quality of the entire population. The process of reviewing this sample can be time-consuming and could potentially be more efficient. Therefore, auditors are looking for a solution to ease the process of auditing, increase the efficiency and include suspicious transactions in the sample.

One of the more promising areas for efficiently detecting fraud is machine learning. Machine learning is a subset of artificial intelligence (AI) that encompasses algorithms that can learn from data without relying on rule-based definitions [94]. Machine learning-based financial fraud detection is one of the few applications of AI that is applied in real-world problems. Under the assumption that fraudulent finances are abnormal from regular finances, these algorithms are able to detect outliers through patterns and statistical learning.

Within the scope of financial fraud detection, there has been increased attention to unsupervised learning techniques [44, 86]. Figure 1 shows the publication dates of the studies on unsupervised learning in the financial domain included in this research. Research in the financial fraud detection area has increased over the years but became more apparent from the year 2000 onward. Unsupervised learning methods can predict outliers without labelled data. Usually, an algorithm needs to learn by including fraudulent and non-fraudulent cases in the training data. Unsupervised learning can learn without these labels and thus requires less manual work beforehand. Unsupervised learning is fit for financial fraud detection as usually no data of identified fraudulent activity is available or only in low quantities. Unsupervised learning is a method that can possibly identify potential fraudulent activity more efficiently than sampling a selection of invoices. Furthermore, unsupervised learning techniques are able to detect unknown, and new forms of fraud [44, 86].

However, unsupervised learning for detecting fraud has met some societal resistance in recent years. An example of this is the SyRI case. SyRI is a Dutch system that was used for providing a risk indication of fraud. The system was able to connect all kind of information sources to provide a fraud indication without any explanation. However, Dutch court has decided that SyRI is in violation with the European treaty for Human rights as it is not transparent and auditable [100]. This case highlights the need of transparency and explainability of an algorithm. Without these, organisations are unable to use algorithms for vital processes. For financial auditing the need for trust and transparency is especially high.



(a) unsupervised algorithms for detecting financial fraud

(b) explanation mechanisms for fraud detection methods in all domains

Figure 1: Publication dates of included studies on

Financial auditors are required to document their work and activities and they need to be accountable for their choices. When considering using unsupervised outlier detection for potential fraud detection, the transparency and trust are needed to review the decisions by the algorithm. The auditor needs to review the outcomes of the algorithm and by knowing why a certain invoice was labelled as outlying, further research can be specified.

Situations like the SyRI case have fuelled research on explanation mechanisms for algorithms. Figure 1 shows that the number of papers concerning explanation mechanisms for fraud detection in all domains has grown explosively from 2017 onward. However, to our knowledge, there is no clear assessment of the different existing mechanisms that can be used to increase the trust and transparency of unsupervised algorithms in the financial domain. Some research has been done on specific algorithms and explanation mechanisms in the financial fraud detection domain, however these studies are scattered and sparse. This lack of research could potentially prevent organisations from applying explanation mechanisms in their own outlier detection algorithms.

1.2 Purpose and contribution of Research

The purpose of this research is to investigate the opportunities for financial auditors to use machine learning in a legitimate manner in their work. There are enough opportunities to apply outlier detection, however the auditors need to be able to trust and explain the outcomes. The goal is to research how financial auditors can trust and begin using outlier detection algorithms in their work of analysing financial statements. The financial statement is a record that summarises the business activities and financial performance of an organisation. These are audited to review the accuracy of their financial reporting [124]. Therefore, in this research we have researched how outlier detection techniques can be applied on invoice data from the public sector and how these can be explained. The public sector encompasses all public good and governmental services and goods such as infrastructure, military, law

etc. [61]. An invoice is a document that documents all the details concerning a transaction between a buyer and a seller. The invoices of the public sector thus concern the transactions that are paid by the government for services for the public sector.

The contribution of this thesis is fourfold. First, it provides an overview of the state of the art research to unsupervised outlier detection algorithms and explanation mechanisms in the financial sector. Second, it provides an indication of features that could indicate errors and potentially fraud in transactional data. Third, it evaluates different unsupervised outlier detection algorithms and their performance on the data of the Dutch public sector. Fourth, it designs an explanation facility according to the needs of the financial auditors. Lastly, it provides recommendations on how to use these processes to support their work.

1.3 Research questions

To achieve the goals of this research, the following research questions are answered.

1. To what extent can unsupervised outlier detection algorithms help to identify potential financial fraud in invoices of the public sector?
    1.1. What features are important for financial fraud detection (identified by domain experts)?
    1.2. What unsupervised outlier detection algorithm delivers the most promising results on the invoice data of the public sector?
2. How can an explanation facility, aimed at explaining the algorithm and its outcomes to a financial auditor, be structured?
    2.1. What is the purpose of the explanation facility for the financial auditor?
    2.2. What explanation mechanisms should be included in this facility?
3. How can the models and explanations contribute to the assessment of the reliability of the financial statement?

1.4 Organisation of thesis

This paper is structured as follows. Section 2 discusses some background on the topics of this area. Section 3 describes the state of the art concerning the existing body of literature on this topic. Section 4 describes the research approach of this thesis. Section 6 introduces the results obtained and discusses them. Section 7 discusses the implications and contributions of the outcomes, the limitations and the recommendations for future work. Section 8 concludes this thesis. Section 9 contains the appendix.

2 Background

The background provides some information related to the topics that are studied in this research. Section 2.1 elaborates on what a financial audit is and how fraud is related to auditing. Section 2.2 provides some information about the unsupervised outlier detection algorithms that are used for fraud detection in the financial sector. Finally, section 2.3 provides some background on explanations of algorithms and what has already been studied.

2.1 Financial audits

Financial auditing is performed to review the financial statements from different organisations and check the reliability of their financial reporting. At the DCGAS, they perform audits for all departments of the government. The entire audit process happens each year and can be divided into three stages.

First, at the beginning of the year all assigned financial audits are collected and planned. In this phase, the different risks are appointed to which the auditors will pay attention. These risks can be based on the financial impact, political impact or risks associated with previous audits. Second, the reviews of the financial statements are executed through audit procedures. For example, samples of financial transactions are being reviewed. If the results of the procedures are found to be correct, the auditor can approve of the financial statement. Finally, once all financial statements are reviewed, the auditors will reflect on the processes and do additional analysis on the data to review the entire year. Some corrections can be made in the final phase.

2.1.1 Fraud within auditing

A common framework that is used to explain the different factors of fraud is the fraud triangle, as shown in Figure 2. The standard for accountants on dealing with fraud uses this triangle as a basis [1]. It states that there are three possible causes of fraud: opportunity, pressure and rationalisation. Opportunity is generally provided through weaknesses in the internal controls. Pressure may be caused by personal financial problems. Rationalisation is a crucial component of most fraud because most people need to reconcile their behaviour with the commonly accepted notions of decency and trust [68].

2.1.2 Process

The process and its components are depicted in Figure 3. The government has a business partner that provides services to the government. This can be a public service, employee etc. The government pays for these services to the business partner. The services delivered and the amount of money together make up the transaction. The business partner, government and transaction together create the process in which fraud can be committed. It should be noted

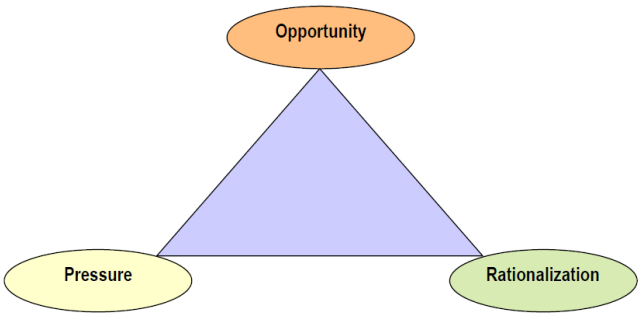


Figure 2: Fraud triangle

that auditors are not actively looking for fraud. Once they suspect fraud, they are obligated to look into this. They must report this back to the client and give them the opportunity to come up with a strategy to fix the mistakes. If no plan is devised, they will reject the audit and give it back to the client. It is assumed that 3% to 5% of fraud is committed in all organisations. 5% is the accepted boundary of receiving an audit certificate.

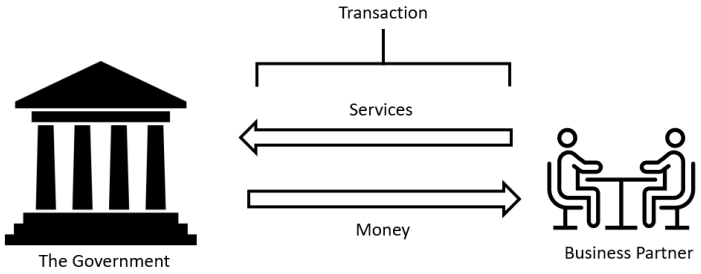


Figure 3: Components of the process

It is relevant to discover what features of transactions can lead to potential fraud. The features of fraud have been researched through interviews with domain experts and by consulting literature. Interviews were held with two employees of the Dutch Central Government audit services. The first person was specialised in fraud within European Union funding and the second in fraud within operational auditing. The interview was semi-structured. The domain experts recommended some sources and protocols within the accounting domain that entail fraud and red flags. Furthermore, a small literature review has been performed to see which factors indicate financial fraud. It is important to research these features from a practical area as well as an academic perspective to include all relevant features. Table 1 shows all the found features. The most popular features are elaborated on in the subsections.

2.1.3 Transactions

There are quite some different features that can flag potential fraud in transactions, as shown in Table 1. Not all features can be implemented to be used for machine learning as they cannot be represented in data. For example, odd behaviour and private problems of the director cannot be translated to a feature. However, a part of the features described in Table 1 can be implemented. A check on double and overdue payments is easy to implement. High volume of purchases from vendors, large, complex transactions at the end of the year and Benford's law can be implemented with some effort. Odd journal entries and odd ledger accounts do not show objectively any requirements and are thus harder to implement in machine learning. An example of an odd journal entry is that a period of days is negative. This should technically not be possible and thus makes it odd. An indication by a financial auditor is needed to indicate this.

2.1.4 Date and Time

Date and time are a characteristics of the transaction. However, also some features can be applied to the business partner. For example, it can be reviewed whether the time of the transaction was during business hours, but also whether the time of creation of the business partner was during business hours. Date and time are mentioned separately as it is a rich source of information. Sources report peak moment analysis, different periods and business hours as indicators of outliers.

2.1.5 Business Partner

The business partner which is involved in the transaction can have many features which could indicate potential fraud. Most features indicate behavioural change such as different activities, high staff turnover, odd behaviour and private problems, insufficient division of functions etc. However, only some features have the potential to be translated into data features. Risky/non-complying countries can be listed. Many entries and corrections can be found in the journal entries. Also sudden activity can be measured such as the lack of a physical address.

2.1.6 Other

There are a few other findings in the literature that do not occur within the transaction or business partner. The authors of [79] have researched the perception of the effectiveness of red flags for detecting fraud. Their study found that some flags are perceived significantly different between internal and external auditors. The relevant flags that are mentioned contribute mainly to management behaviour. Next, the authors of [54] have researched whether the red flags as mentioned in Statement on Auditing Standards No.99 (SAS99) are useful for external auditors. Their research confirms that multiple flags have been identified

as useful and they rank these in their research. SAS99 is an older standard for handling fraud from the United States. These findings cannot contribute to the selection of features but do influence them.

2.1.7 Discussion and Conclusion

Many indicators that can flag potential fraud are described in professional literature. As shown in Table 1, many features concern behavioural changes. However, some can be translated into data features to use for machine learning. The following will be considered in this research.

- Invoice not paid in time
- High volume of purchases from new vendor
- Large, complex and unusual transactions at the end of the year
- Benford's law
- Peak moments
- Quartiles
- Business hours and days
- Risky/non-complying countries of origin of business partner
- Many journal entries and corrections of business partner
- Sudden activity of business partner while being dormant
- No physical address recorded of business partner

2.2 Unsupervised outlier detection algorithms

Within this section, we aim to give some more background on the different domains of unsupervised outlier detection algorithms and how these work. In this research, we focus on the following categories: decision trees, support vector machines, local outlier factors, neural networks, clustering, Bayesian networks and others.

2.2.1 Trees

Decision trees consist of nodes and paths [44, 83, 84]. Each node represents a test on a certain feature. The classification rule concerns the entire path from root to the leaf . The design of the decision tree depends on the information gain of each attribute. The attributes

Subject	Features	Source(s)
Transactions	- Double payments - Paid through offshore - High brokerage fees - Payment in natura - High travel and phone costs - Anonymous deposits - Odd journal entries - Not paid in time - Odd ledger account - Excessive number of voids, discounts and returns - Abnormal number of expense items, supplier, or reimbursement to employee - High volume of purchases from new vendor - Large, complex and unusual transactions close to end of year - Benford's law	Interview DCGAS, workshop DCGAS, [4, 30, 54, 33]
Date and time	- Peak moments - Periods - Business hours	Interview DCGAS, workshop DCGAS
Business partner	- Non-existent - Different activities than registered - Foreign holding of a Dutch partnership - Family ties - Risky/Non-complying country - Director in India - Excessive or unjustified amount of cash - Past of money laundering - High staff turnover - Odd behaviour and private problems director - Administration not up-to-date - Insufficient division of functions - Many journal entries and corrections - Complaints of suppliers - Small organisation - Corporate business - Staff lacks training - Sudden activity while being dormant - Vendors with no physical address - Improper recording of sales	Workshop DCGAS, [4, 30, 54]

Table 1: Features of fraud detection

with the higher information gain, meaning that they separate instances often, are placed in the beginning of the tree. More specific attributes are placed in the last few splits.

There are different variations in this category e.g. random forests construct many decision trees on the same data and use the average prediction. Decision trees are favourable as they can handle numerical and categorical data, are easy to interpret and can deliver good results [44]. However, overfitting and bias are risks of the decision tree.

The first model that is used in this research is the Isolation Forest. Therefore, below is a more elaborate explanation of this algorithm.

2.2.1.1 Isolation Forest

Isolation Forest (IF) aims to isolate anomalies and distinguishes itself by not focusing on the normal exemplars [65]. It is based on the assumption that outliers are few and different. IF builds an ensemble of trees. The outliers have on average shorter path lengths. IF goes through two stages for detecting anomalies. First, it builds the trees using the training data set. It does so by selecting a subset $X' \in X$. Then, it recursively divides X' by randomly selecting an attribute q and a split value p until either (1) the node has only one instance or (2) all data at the node have the same values. Once the trees are build, all instances go through the trees and get assigned a proper anomaly score.

$$c(m) = 2H(m-1) - \frac{2(m-1)}{n} m > 21 m = 20 otherwise$$

where n is the testing data size, m is the size of the sample set and H is the harmonic number, which can be estimated by $H(i) = ln(i) + \gamma$, where $\gamma = 0.5772156649$ is the Euler-Mascheroni constant.

The estimation of the anomaly score of X can be given by:

$$s(x,m) = 2^{\frac{-E(h(x))))}{c(m))}}$$

2.2.2 Support vector machines

Support vector machines (SVM) seek out a hyper plane in an N-dimensional space that has the maximum margin [44, 98]. The SVM searches for a plane that maximises the distance between the points of both classes of data points. The loss function that attributes to this is hinge loss. The loss function helps to detect the gradients and update the weights. The dimension of this hyper plane depends on the number of attributes.

SVMs are known for their robust outcomes and they can deal with very high dimensional data. However, SVMs require a lot of memory and CPU time [44]. Further, the regular SVM requires both positive and negative examples for training. The one-class SVM does not require both examples to be included [50].

2.2.2.1 One class support vector machine

One class support vector machine (OCSVM) is a machine learning technique optimised for novelty detection. OCSVM intends to separate the origin from the data instances in the kernel space which results in some complex hulls describing the normal data in the feature space [44]. The OCSVM is first trained using the data set and afterwards, all instances are scored by a normalised distance to the determined decision boundary.

As the regular SVM uses a hyper plane with the largest possible margin, the OCSVM uses the smallest possible hyper sphere to encompass all of the normal instances.

2.2.3 Local outlier factor

Local outlier factor (LOF) is an algorithm based on the neighbours of a certain data point [39, 44, 101]. LOF measures the local deviation of density of a given sample with respect to its neighbours. A comparison can be made of the density of the cluster to the densities of other neighbours. If the density is significantly lower, the instance is considered an outlier. LOF works for local examples but is also suitable for detecting global outliers.

The process can be summarised in three steps [44]:

1. The k-nearest neighbours have to be found for each instance of X.
2. The local reachability density (LRD) is computed to estimate the local density for an instance:
$$LRD_k(x) = 1/\frac{\sum_{o \in N_k(x))} d_k(x,0)}{|N_k(x)|}$$
3. The LOF score is calculated by comparing the LRD of one instance to its k-nearest neighbours:
$$LOF(x) = \frac{\sum_{o \in N_k(x))} \frac{LRD_k(o)}{LRD_k(x)}}{|N_k(x)|}$$

2.2.4 Neural networks

Neural networks are modelled after the human brain. They exist of multiple layers of neurons (nodes) that make the predictions [44, 62]. Neurons are interconnected and data travels from the input to the output through the neurons. The building blocks of the neural network are the connections between so-called weights living inside the neurons and the neurons themselves. The neurons include a bias term and an activation function. These are used to calculate the output of the neural network. Finally, a cost function is used and minimised by using gradient descent optimisation.

It is important to note that forward propagation is used to calculate the activation functions. However, backward propagation is used to reconstruct the error and to fine-tune all the neurons and their connections. This means that each neuron is their own miniature model with its own features and weights, allowing for high accuracy. This makes neural networks complex to understand but also robust [83].

2.2.5 Clustering

Clustering is the task of dividing individual and similar data points into groups [40, 44]. These groups posses similar traits and are called clusters. Hard clustering specifies that each data point belongs to a cluster or not. Soft clustering calculates the probability that a data point belongs to a certain cluster. The clustering can be based on different means. This explains the large amount of clustering algorithms that exist.

Clustering is simple and easy to scale. However, it can be sensitive to noise and outliers. Furthermore, it is not fit for high dimensional spaces [83].

2.2.6 Bayesian algorithms

Bayesian algorithms are based on Bayes' theorem [15]. The theorem assumes conditional independence between different pairs of features. The algorithms differ mainly on the assumption they make regarding the distribution. Despite the simplicity of the algorithms, they perform generally quite well and can be extremely fast. Its simplicity also provides clarity to the users.

The "others" category contains different algorithms that are based on different assumptions and methods.

2.3 Explanation mechanisms

To identify the scope of this research, we elaborate first on what a trustworthy algorithm constitutes. The European Commission has published the Ethics guidelines for trustworthy AI to give some guidelines on trustworthy AI and and how to create such an AI. [104]. The unsupervised outlier detection algorithms that this research targets are part of these AI methods. The realisation requirements discussed in the report are summarised in Figure 4.

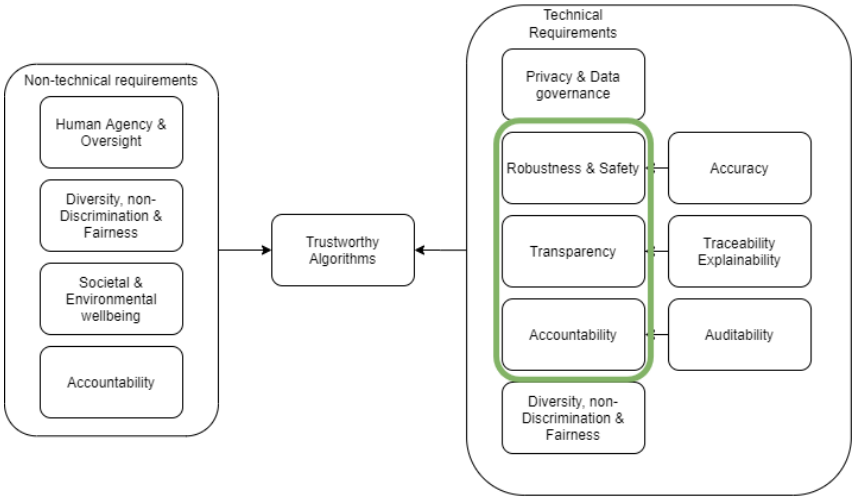


Figure 4: The different realisation requirements for a trustworthy algorithm

2.3.1 Trustworthiness

The green circle highlights the requirements on which this research focuses. These three are selected as they are heavily influenced by the algorithms themselves. To realise these requirements, different aspects have been proposed, as shown in Figure 4 on the right of the green circle. Figure 4 specifies on the accuracy, traceability, explainability, and audibility. The accuracy concerns the ability of an algorithm to make the correct decisions. Traceability pertains the process of documenting the use of data and algorithms as well as the decisions of algorithms. It facilitates auditability and explainability and is also called interpretability by others. Explainability is the ability to explain both the technical processes of the algorithm as the human decisions related to it. Auditability entails that algorithms should be able to be assessed including the data and their design processes and are therefore reproducible.

Term	Definition
Transparency	Explain how the system works
Scrutability	Allow users to tell the system it is wrong
Trust	Increase users' confidence in the system
Effectivenes	Help users make good decisions
Persuasiveness	Convince users to try or buy
Efficiency	Help users make decisions faster
Satisfaction	Increase the ease of usability or enjoyment

Table 2: The aims of explainability according to [114]

2.3.2 Explainability

As many aspects and terms are proposed, this leads to a confusion. The aspects are used interchangeable as they overlap while each having their unique properties. The authors of [114] have distinguished different terms that constitute explainability. These are shown in Table 2. The author of [64] states that interpretability consists of transparency and post-hoc explanations. These two elements determine the explainability of an algorithm. This research focuses on transparency, trust, effectiveness, and efficiency captured in post-hoc explanations.

We refer to the trust increasing mechanisms as explanation mechanisms. These mechanisms influence the interpretability, transparency and traceability of the algorithms and therefore influence the trust.

The authors of [115] state that there are six different types of explanations for recommender systems. Case-based explanations show similar instances, collaborative explanations show similar users, content-based explanations apply previous behaviour on new instances. Conversational explanations show them in textual form. Demographic explanations reason from the demographics of the user and knowledge-utility based explanations reason from the

experience of an user. The first four are applicable on explanations for invoices as they are based on instances. The latter two are based on the user and their experience which is not applicable to invoices.

The authors of [121] categorise explanations as causal attribution which explain why events and behaviours occur. This provides broad information from which the user can judge and identify potential causes. The other explanation is causal explanation which is the explanation of the internal physical mechanism of a phenomenon. It focuses on selected causes to interpret the observation with regard to existing knowledge.

The information commissioner's office of the UK has developed a guideline together with the Alan Turing Institute to correctly implement explainable AI [49]. They distinguish between process-based and outcome-based explanations. For accountants, outcome-based explanations are most interesting as they want to know why a certain invoice was flagged as outlying. Furthermore, the guideline reports six different types of explanation dependent on the stakeholders:

- **Rationale explanation**: the reasons that led to a decision, delivered in an accessible and non-technical way.
- **Responsibility explanation**: who is involved in the development, management and implementation of an AI solution, and who to contact for a human review of a decision.
- **Data explanation**: what data has been used in a particular decision and how.
- **Fairness explanation**: steps taken across the design and implementation of an AI solution to ensure that the decisions it supports are generally unbiased and fair, and whether or not a stakeholder has been treated equitably.
- **Safety and performance explanation**: steps taken across the design and implementation of an AI solution to maximise the accuracy, reliability, security and robustness of its decisions and behaviours.
- **Impact explanation**: steps taken across the design and implementation of an AI solution to consider and monitor the impact that the use of an AI solution and its decisions has or may have on a stakeholder, and on wider society.

Mainly the first explanation is important within this research. The others can also be important for financial auditors, however are not considered in this scope.

Other research mostly indicates the different explanation mechanisms that can be implemented directly but do not categorise these in types of explanations [17, 41, 45, 48, 116, 117]. Especially the distinction made by the authors of [117] can be useful and has been incorporated in Figure 5. This figure has combined the existing categories of explanation mechanisms as mentioned in section 3. An important distinction that is made here is between

process-based and outcome-based as described in [49]. There are slight deviations as this part of the research focuses mainly on the format of the explanation mechanism. E.g. SHAP is a model-agnostic mechanism, but shows the feature importance of all models. Therefore it is now grouped in the feature importance category.

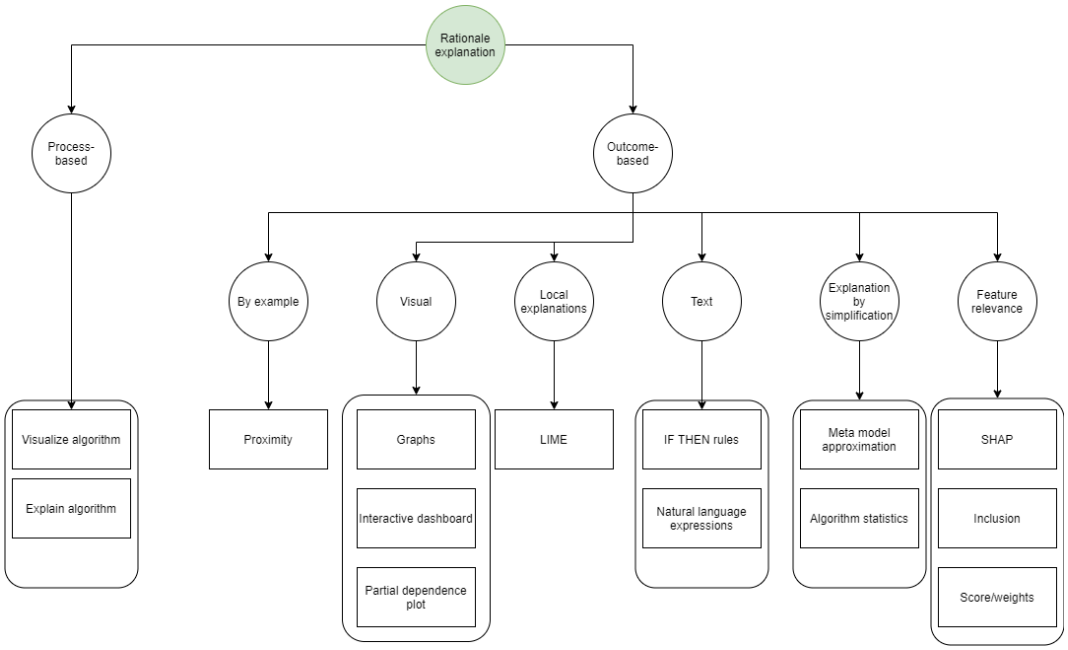


Figure 5: Overview of different types of explanations connected to explanation mechanisms.

3 State of the Art

In this section, we present and discuss the results of a literature study to the state of the art concerning unsupervised outlier detection in transactional data and their potential to integrate different explanation mechanisms. Section 3.1 describes the methodology approached for this literature review. Section 3.2 discusses the different algorithms found. Section 3.3 elaborates on the different explanation mechanisms found and what these entail. After, the assessment matrix is discussed in section 3.4. Two determining factors are found for the results which will be discussed in section 3.5. This section concludes with some promising areas for research in section 3.6.

3.1 Literature review approach

The methodology is based on the systematic quantitative literature review developed by Griffith University and the systematic literature review as described by Kitchenham, and Webster and Watson [59, 92, 122]. This literature review focuses on unsupervised outlier detection algorithms that have been applied in the financial domain that concerns transactions and statements, as shown in Figure 6. When referenced to the financial domain in this literature review, it only includes these two areas. The entire financial domain is not accounted for. The explanation mechanisms that are included are applied in all domains on anomaly or fraud detection. The found mechanisms are thus mechanisms compatible with outlier or anomaly detection, independent of the domain. A selection of 106 articles was included in this literature review. The goal of this literature approach is to provide insight into (1) what unsupervised, outlier detection methods exist for detecting fraud in the financial sector, (2) which mechanisms are described that can increase trust in the outlier detection methods and to what extent are these applied in the financial world, and (3) what the opportunities and gaps are in the application of explanation mechanisms on the outlier detection methods in the financial fraud domain.

3.2 Algorithms

The literature review provides an overview of the different unsupervised algorithms found in the included studies concerning financial fraud detection. These are listed in Table 3. A total of forty-five algorithms are found and divided into seven categories which are described below.

3.2.1 Trees

Trees are often used as outlier detection methods. Decision trees have been around for some time but its simplicity and good accuracy still make it attractive to use [7, 36]. The random forest is a collection of many decision trees. This makes it an ensemble method and has

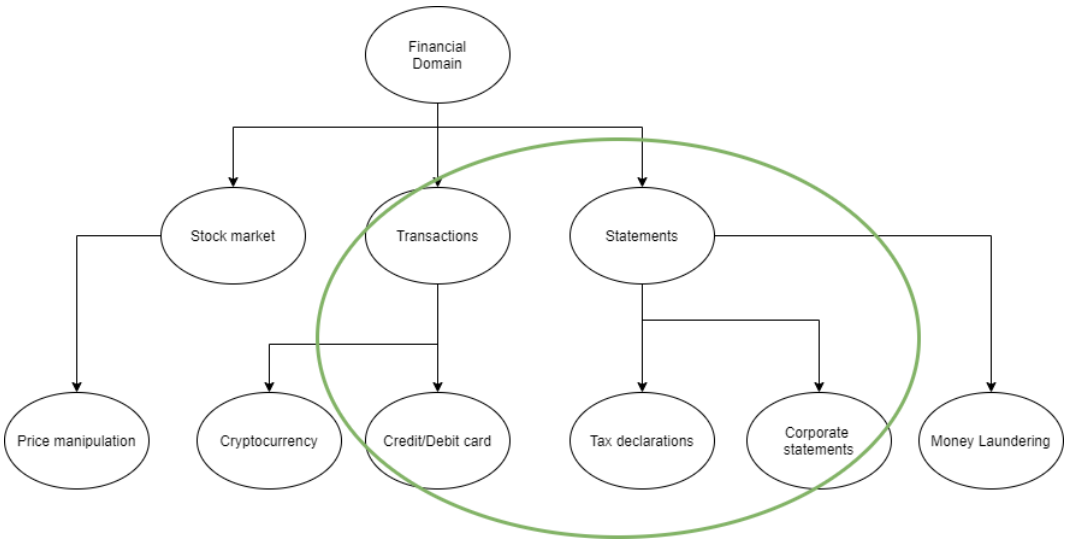


Figure 6: The different areas of the financial domain

thus in general a higher accuracy [10, 22, 112, 130]. The Isolation forest is based on random forests to isolate the anomalous points from the normal ones. The Isolation forest outranks others as it has many advantages. It only needs a small training set and does not require example anomalies [31, 37, 38, 40, 75, 84, 111].

3.2.2 Support Vector Machines

Support vector machines are a popular category as well. The support vector machine has been popular from early discovery and has seen some improvements over the years. Its ability to deal with highly imbalanced data makes it a likely candidate [98, 22, 7, 83]. The one class SVM (OCSVM) only needs a training set with instances of one class. This is ideal in fraudulent cases as the number of known, fraudulent instances is usually very low. This makes the OCSVM particularly fitting for unsupervised learning in the financial domain [31, 37, 50, 83, 84, 101, 111].

3.2.3 Local Outlier Factor

As discussed in Section 2.2, the local outlier factor (LOF) method uses the local neighbours of a data point and calculates the deviation and outlier score [31, 38, 75, 84, 118]. LOF is used to identify data points that have a bigger distance from the others. Within financial fraud, these data points are then flagged as potential fraud. The authors of [118] found that LOF is still one of the most state of the art methods for detecting potential. The advantage is that LOF is specialised in recognising local outliers but also works on a global level, making

it possible to recognise different types of fraud. [44].

3.2.4 Neural Networks

Neural networks, auto-encoders and self-organizing maps are used frequently in the category of neural networks. Neural networks have a high accuracy when configured properly and have a wide array of functionalities [22, 26, 53, 70, 91, 106]. The authors of [31] use a neural network in combination with two thresholds to detect closer and outlying nodes that can identify outliers and consequently fraud. Autoencoders have proven to be fast in comparison to other methods while still detecting subtle anomalies [50, 86, 89, 93, 101, 106, 127]. Self-organizing maps are a type of neural networks that configure their neurons according to the input data. They are easy to interpret and while processing the training data can take some time, new data is mapped immediately [13, 53, 62, 75, 95, 129]. This is very useful in the financial domain where transaction are created continuously.

3.2.5 Clustering

Clustering methods find groups of data points with the same characteristics [83]. These clusters are used to identify data points that do not belong to any cluster. These are labelled as outliers and thus potential fraud. The core of the clustering technique can be based on different metrics. K-means is easy to scale to large data sets and adapts to new examples [7, 75, 76, 84, 111]. Gaussian models are used frequently as they are simple and flexible [10, 37, 53, 86, 96].

3.2.6 Bayesian Models

Different Bayesian models have been studied in the area of fraud detection [7, 15, 22]. The authors of [15] use the probability that transaction A is paid by account B and create two different thresholds for this Bayesian model. Although included in this literature review, there is not much research describing the use of Bayesian models. This can be due to its preference for labelled data and that it is often outperformed by more complex algorithms.

3.2.7 Other

Several other algorithms have been studied within this domain which do not belong to the other six categories. Logistic regression is included in quite some studies [22, 83, 98, 106]. It is named as a classical and traditional method for binary classification. The others are used sparsely.

Overall, it seems that Isolation Forests, support vector machines, neural networks, self-organizing maps, auto-encoders and k-means clustering are most used in research within the financial domain, as stated in Table 3. The respective studies declare these algorithms as state of the art and being able to provide good results.

3.3 Explanation mechanisms

Second, an overview is established of the different explanations mechanisms found in in the included studies. In Table 4, the mechanisms are listed, including a small explanation and a reference to the original study. In total, twenty-nine explanation mechanisms are found and divided into six categories which are described below.

3.3.1 Feature-based explanations

Feature-based mechanisms are used in multiple studies. Showing whether features are included in the model is used often [2, 3, 5, 6, 12, 20, 43, 47, 78, 80, 81]. By stating how much influence a feature has had on a decision, transparency into the processes is created. The scores or weights are often readily available and give quite some insight into the process and decision of an algorithm. Relationship of features and their differences are used infrequently as these are often not available.

3.3.2 Semantic explanations

Semantic explanations use natural language with human meaning to explain the decisions. When available, a semantic meaning can be given [87]. This can be the relationship between transactions e.g. However, often the data does not include a semantic meaning. Other information can be incorporated into natural language expressions and rules [2, 5, 32, 98, 110]. Natural language expressions are flexible and thus allow for different sources of information to be communicated. However, it is challenging to communicate a lot of information in natural language. The overview is rapidly obscured.

3.3.3 Visualisation techniques

Visualisation techniques are a great tool for explaining different processes. Graphs are commonly used in all sectors to enhance understanding [8, 46, 58, 67, 81, 97, 106]. These graphs can be turned into an interactive explanation facility, making it more informative and tailored to specific needs [34, 67, 108, 119, 130]. In the financial domain, this could mean that one can filter on transactions for a certain department. Explanation facilities are able to contain a lot of information and give a clear overview. The deep Taylor decomposition can be applied to neural networks [55]. Saliency maps can be used for image processing and provide visual aid to understanding the anomalous nature of certain images [21, 73]. As these maps share the same dimensions as the input data, it is possible to maps the anomalous pixels to the input image. Partial dependence plots are able to plot the relation between input factors and their outcome. While being a great aid, the plots can only be applied to a small number of

features before losing the overview [73, 130].

3.3.4 Metrics

Metrics are a common aid for checking algorithms and their performance. The metrics can be used as explanation and to increase trust in the decision of the algorithm. Most metrics are common such as confidence, distance, mutual information, proximity, odds ratio, activation, algorithm statistics, and the Gini index [5, 29, 43, 52, 73, 74]. However, two metrics are specifically aimed at explaining the algorithm, namely permutation importance and normalcy exemplar. The first metric alters the input variables to show the change in outcome [73]. The second metric aims to show a normal example to explain why a certain data point is anomalous [63, 108]. These last two metrics are interesting as their sole intention is explanation.

3.3.5 Model-specific mechanisms

The model-specific mechanisms are aimed at two classes of algorithms, namely decision trees and clustering. Trees can easily be visualised when a feasible amount of trees is executed [18, 23]. A challenge arises when multiple trees are created and used as an ensemble, making it harder to visualise the trees in one overview. Clusters can easily show why a certain data point belongs to a cluster, making it a low effort mechanism to explain the decisions of an algorithm [57, 105, 131].

3.3.6 Model-agnostic mechanisms

The last category includes the model-agnostic mechanisms which can be applied to any algorithm. One approach is the approximation of the decisions by using a meta model [128]. Any kind of algorithm can be used to approximate but certain subsets can heavily influence the results. LIME and SHAP are two well known mechanisms that have been created to explain algorithms [42, 43, 102, 113]. SHAP uses methods from game theory to explain the output of the algorithm. LIME focuses on training local surrogate models to explain the predictions instead of a global surrogate model. LIME and SHAP do not depend on reconstruction and are therefore advantageous. Their efficiency and speed are disadvantages. Parzen and localization are two existing mechanisms that can also be used for explanation [58, 102, 108]. Localization is the process of identifying the location of an instance that influenced the decision. An example is using a box to place over the outlying element in a picture. Localization can be simple and thus easy to interpret. Parzen is based on taking the gradient of the prediction probability function. Parzen requires more fine tuning, but has a strong theoretical foundation.

3.3.7 Explanation mechanisms in the financial domain

The results in Table 4 show that there is a wide variety of mechanisms for increasing trust in fraud detection algorithms in all domains. The explanation mechanisms that have been applied in the financial fraud domain are coloured in Table 4. The number of described mechanisms in the financial fraud domain are limited compared to the total number of mechanisms. The authors of [2, 74] have used the feature score and natural language as explanation mechanisms. The use of IF THEN rules have also been discussed [98]. The authors of [56] use the semantic meaning of a graph database and the explainability of trees is investigated [37] Furthermore, the use of feature inclusion has been used in financial statement fraud [47]. The authors of [63] have used a normalcy exemplar to explain their outliers. The other explanation mechanisms have been not been studied in research on unsupervised outlier detection algorithms in the financial domain. Only six out of the twenty-nine described explanation mechanisms are used in the financial fraud domain.

Table 3 shows that most mechanisms have been described only once in the included literature on fraud detection in all domains. There are only a few that have multiple studies, these include feature scores/weights, feature inclusion, graph, interactive visualisations and SHAP. This could be due to their success or their ease of implementation. It can be noted that the use of different explanation mechanisms is very limited in the financial domain on unsupervised algorithms.

	Name	Studies
Trees	Decision tree	[7, 35]
	Random forest	[22, 10, 112, 130]
	Isolation forest	[31, 37, 38, 40, 75, 84, 111]
	Balanced random forest	[16]
	Minimum spanning tree	[53]
	Gradient boosted tree	[106]
	Frequent pattern tree	[51]
	C4.5	[22]
Support vector machine	Regular SVM	[98, 22, 7, 83]
	One class SVM	[31, 37, 50, 83, 84, 101, 111]
Local outlier factor	Regular LOF	[31, 38, 75, 84, 118]
	Clustering-based LOF	[39]
Neural network	Probabilistic NN	[98, 83]
	(Deep) NN	[91, 22, 26, 53, 70, 106]
	Recurrent NN	[38]
	Autoencoder	[50, 85, 89, 93, 101, 106, 127]
	Variational autoencoder	[99]
	Competitive learning network	[62]
	Multi-layer feed forward NN	[83]
	RBM	[93]
	Long short term memory	[20]
	CNN	[67]
	Self-organizing map	[13, 53, 62, 75, 95, 129]
	GAN	[108]
	Hierarchical cluster-based deep NN	[57]
Clustering	Cluster analysis	[91, 7, 44]
	Spectral clustering	[27]
	K-means	[7, 75, 76, 84, 111]
	Euclidian distance	[38]
	Expectation-maximization	[60]
	Gaussian mixture modelling	[10, 37, 53, 85, 96]
	DBSCAN	[60, 111]
Bayesian Algorithms	Naïve Bayes	[22]
	Bayesian model	[7, 15]
Other	Principal component analysis	[31, 91]
	(Multi-nomial) Logistic regression	[98, 22, 83, 106]
	Discriminant analysis	[83]
	OneR	[22]
	Hidden markov model	[91, 7]
	Graph mining	[91]
	Link analysis	[91]
	Peer group analysis	[91, 123]
	Angle based outlier detection	[31]
	Salient Object detection	[31]
	Kernel density estimation	[38]

Table 3: Studies per algorithm

	Mechanism	Study	Description	Strength	Weakness
Feature-based	Score/weights	[2, 3, 5, 12, 20, 43, 78, 80, 81]	Showing the importance of a feature to a certain decision through their score or weight in the process.	Easy to implement	Subjective interpretation
	Inclusion	[6, 12, 20, 47, 78, 80]	Showing which features were used for making the decision.	Easy to implement	Hard to give an overview
	Relationship	[12, 72]	Showing the relationship between the features that enabled the decision.	Applicable to different properties of features	Limited features
	Differences	[103]	Showing the difference of one feature to the others.	Intuitive visualisation	Too limited
Semantic explanation	IF THEN rules	[98]	Implementing rules interpretative for humans	Easy understanding	Too simple for more complex problems
	Natural language expressions	[2, 5, 32, 110]	Using natural language for explaining the factors leading to a decision.	Understandable for all users	Much manual work to set up
	Semantic meaning in relationships	[87]	Showing the relationship between data points in a human-interpretative, meaningful manner.	Much information gain	Only available for graph databases
Visualisation	Saliency maps	[21, 73]	Creating an image that shows the anomalous pixels and thus generates an explanation	Share dimensions of input data	does not necessarily imply importance
	Interactive explanation interface	[34, 67, 108, 119, 130]	Creating a dashboard that displays the decisions and factors in an interactive manner.	Many possibilities	Knowledge on data needed
	Partial dependence plots	[73, 130]	Depicting the functional relationship between the input variables and the prediction output.	Intuitive and easy to interpret	Assumption of independence
	Graphs	[8, 46, 66, 81, 58, 97, 132]	Using graphs and other simple visual aids to show the factors and features leading to a decision.	Applicable to all data	Shows only limited data
	Deep Taylor decomposition	[55]	Producing a decomposition of the neural network output in terms of the input variables.	Appropriate for complex structures	Limited to neural networks
Metrics-based	Confidence	[5]	Showing the quantification of the uncertainty of an estimate.	Good indicator of results	Little explanation
	Distance	[74]	Showing the distance to the kernel gives an indication of the outlier score.	Uncovers a degree of outlierness	Hard to interpret
	Mutual information	[52]	Calculating the gain of each variable in the context of the target variable.	Can compute optimal explanation	
	Proximity	[43]	Showing the similarity between two data points.	Intuitive	Only local
	Odds ratio	[43]	Showing the quantification of the strength of the association between two data points.	Objective measure	Shows no causality
	Activation	[43]	Showing the function that defines the output of that node given an input variable.	Easily available	Knowledge needed
	Algorithm Statistics	[29]	Showing the general statistics of the performance of an algorithm.	Always available	Only limited information gain
	Permutation importance	[73]	Showing the relationship between the feature and the target by shuffling values of features.	Compressed and global overview	Repeated results may vary greatly
	Normalcy exemplar	[63, 108]	Showing a normal-considered example to demonstrate the difference with an outlier.	Available to all data types	Needs some visualization to make it easy to interpret
Model-specific	Tree explainability	[18]	Showing the decisions of a tree by visualizing them.	Readily available	Not feasible for large and complex trees or ensembles
	Characteristics of cluster	[57, 105, 131]	Showing the characteristics of a cluster can help to explain why an instance belongs or does not belong to a cluster.	Readily available	
Model-agnostic methods	Meta model approximation (ACE)	[128]	Using a different and simple model to recreate the decisions and making the process more transparent.	Flexible	Results depending on subsets
	LIME	[102]	Tweaking the input variables and determining how these influence the outcome to detect the internal processes.	Human-friendly explanations	No correct explanation of neighbourhood
	(Kernel) SHAP	[42, 43, 113]	Showing the average marginal contribution of a feature value across all the possible coalitions of features.	Solid theoretical explanation	Slow to compute
	Localization	[108, 58]	Finding the location of the data points that make it an outlier and showing these.	Interpretive	Used only for images
	Parzen	[102]	Explaining individual predictions by taking the gradient of the prediction probability function.	Strong theoretical foundation	Hyper-parameters need to be tuned

Table 4: The different explanation mechanisms

3.4 Assessment matrix

The last result concerns a matrix that assesses which combinations of algorithms and explanation mechanisms have been studied. Furthermore, the matrix indicates whether it is feasible to use certain explanation mechanisms for certain algorithms. This is set out in Table 6. The green cells represent combinations which have been reported in academic literature and deemed possible. The scale used for the opportunities can be seen in Table 5 which includes the colour scale as well. The explanation of the scale is based on the feasibility of implementing an explanation mechanism and the technical limitations. Table 6 contains a lot of information so the highlights of the matrix are presented in the sections below.

Scale	Explanation
1	Very suitable
2	Suitable
3	Possible
4	Not preferable
5	Not possible

Table 5: Description of the scale

3.4.1 Feature-based mechanisms

The use of scores and weights of features has been studied for the classes of trees, support vector machines, LOF and neural networks. [3, 5, 12, 20, 43, 78, 80, 81, 88, 110, 131]. The inclusion of features has been researched in all categories except LOF and clustering [6, 47, 78, 80, 110]. As these explanation mechanisms are compatible with most algorithms, the feasibility of the combinations of these mechanisms and algorithms is high. The relationship of features has been studied in neural networks and Bayesian networks but is overall harder to achieve because of the complexity [12, 72]. The differences of features has not been studied yet but there are some possibilities depending on the transparency and complexity of the algorithm.

3.4.2 Semantic explanations

"If then rules" are easy for decision trees to implement, but harder for the other categories due to their design. Natural language expressions have been used for neural networks and clustering methods [2, 5, 32, 74, 80]. For the others, this can be implemented with ease.

Semantic relationships are only possible with data graphs [56].

3.4.3 Visualisation techniques

In the category of visualisation, substantial research has been done. Especially interactive explanation interfaces and graphs have been researched in most categories and are feasible for all algorithms [8, 34, 46, 58, 81, 97, 108, 130, 131]. Partial dependence plots have been researched by [130] and applied to decision trees. Partial dependence plots are also applicable to other algorithms but they only allow for a small number of features. This limits their feasibility for complex algorithms. Deep Taylor decomposition is limited to neural networks and saliency maps are limited to image processing algorithms [55, 56].

3.4.4 Metrics

Considering the metrics-based mechanisms, it is apparent that distance, proximity and activation are limited to a range of algorithms [74]. Confidence, algorithm statistics, permutation importance, odds ratio and normalcy exemplar are easy to implement for all algorithms [5, 29, 43, 74, 108].

3.4.5 Model-specific mechanisms

The model-specific mechanisms are both researched within their own domain [18, 23, 43, 63]. Characteristics of clusters have been studied by [71].

3.4.6 Model-agnostic mechanisms

The model-agnostic mechanisms show much promise. Meta model approximation, LIME and SHAP are available to all models and only need little alterations [42, 43, 102, 113, 128]. Localization is not available for trees but can otherwise be implemented for the other categories, ranging in difficulty [58, 108]. The feasibility of Parzen is also dependent on the algorithm [102].

The model-agnostic mechanisms are most promising for the different algorithms. Within this category, Parzen and localization are limited to algorithms with certain features. However, ACE, LIME and SHAP are often feasible with little effort and can provide detailed explanations. These mechanisms were specifically developed to explain algorithms and due to that reason have an advantage over the other mechanisms.

3.5 Determining factors

Within this section, we highlight some of the determining factors that seem to determine the feasibility of certain combinations, as seen in Table 6. There are a few explanation mechanisms that are considerably blue for all algorithms and some are mainly red. These observations can be explained through two causes, 1) the structure of the data and 2) the structure of the algorithms.

The first being the structure of the input data. Sometimes, the structure of the data can make an explanation mechanisms incompatible. Saliency maps are strictly compatible with image processing. However, decision trees are not fit for processing images, making it impossible to combine these two. Another example is the semantic meaning in relationships. To be able to get a semantic meaning, the data set needs to have semantic meaning itself. This is true for graphs databases, making it eligible for only a very small number of algorithms. Some explanation mechanisms are not dependent on the structure of the data and are thus available for most algorithms. Natural language expressions can include different types of data. The same applies to graphs.

The second determining factor is the structure of the outlier detection algorithm. The activation mechanism is limited to neural networks. The same reasoning applies to proximity and reasoning. These two mechanisms are only applicable to algorithms that place the data point on a plane. This factor also enables certain explanation mechanisms to be applicable to all algorithms. Algorithm statistics are available for all algorithms. Furthermore, most of the model-agnostic mechanisms are available for all algorithms because the mechanisms do not depend on the structure of the algorithm.

Both these factors limit the possible combinations but also enable opportunities for future research. These promising areas will be explained below.

3.6 Promising areas

Next to the described combinations, Table 6 contains possibilities for other combinations that have not yet been researched within the included literature. We have used a scale of 1 to 5 to indicate whether the combination is feasible or it will take a lot of effort.

The *feature-based explanation mechanisms* are usually easy to implement for all algorithms and require little effort. These mechanisms are already used with several algorithms but are also a possibility for the others. As the importance of a feature is usually a good and interpretive explanation for decisions, this could be a promising area.

The *semantic explanation mechanisms* show mixed result. IF THEN rules are usually possible but require some effort to acquire. Furthermore, it is questionable whether these rules and their volume are easy to interpret for the human user and whether it is worth the effort. Natural language expressions can increase the interpretability for human users significantly in combination with the feature importance. The expressions are feasible to implement for most algorithms. Semantic meanings of relationships is only feasible for data graphs. In graphs, data points have a relationship which has a semantic meaning. However, as graph mining is only one algorithm in Table 6, this mechanism is too specific.

Visual tools are valuable for explaining algorithms and their decisions to users. Two mechanisms which have mixed opportunities are deep Taylor decomposition and saliency maps. Both cater to a limited audience: respectively neural networks and image processing. Graphs are commonly used and can be applied to almost all algorithms. These in turn can be displayed in interactive explanation facilities. This can be harder depending on the algorithm as some have more information to visualise than others. Partial dependency plots can show relevant explanations. There is just the drawback that only a small number of features can be plotted. As most data sets have a significant number of features, this could increase the number of partial dependence plots and decrease the interpretability.

Metrics-based explanation mechanisms have mixed results. Proximity, distance and activation are dependent on the kind of algorithm. Confidence and algorithm statistics are easy to calculate or already available. These are a good addition and can increase the trust in the algorithm and its decisions. Gini index, permutation importance, odds ratio and normalcy exemplar are usually feasible and can increase the interpretability of a decision. They do require some effort to calculate and it should thus be review whether the effort is worth it.

The *model-specific mechanisms* are naturally limited to the two categories they belong to. However, for these categories it is a readily available or little effort mechanism for explaining the algorithm. Due to their low effort and availability, these are promising for their respective categories of algorithms.

The *model-agnostic mechanisms* are most promising for the algorithms. Parzen and localisation are sometimes limited to algorithms with certain features. However, ACE, LIME and SHAP are often feasible with little effort and can provide detailed explanations.

3.7 Limitations

This research tried to include all the relevant literature within the specified scope. However, it is possible that some studies were not included in our search queries and are thus not accounted for. This limit is always present with literature reviews. By using a proper methodology, we tried to mitigate this situation but it cannot be ruled out that we missed relevant

studies. Furthermore, the indication of feasibility for the different combinations is based on the selected literature. However, there are a lot of factors influencing the feasibility. Future research can gather more literature to confirm the indications and include more factors.

3.8 Conclusions

This section has shown the current body of research on unsupervised outlier detection algorithms in the financial sector and explanation mechanisms. Table 6 has indicated combinations that are fruitful and not yet researched. This matrix has led to the decision to include the following three algorithms in this research: IF, LOF, and OCSVM. This section has shown that these three have shown a good performance in previous research and are all relatively easy to implement. Furthermore, the matrix is used to investigate which explanation mechanisms are possible in combination with the three selected algorithms.

		Feature-based				Semantic explanation			Visualisation				Metrics-based								Model-specific		Model-agnostic methods				
	Name	Score/weights	Inclusion	Relationship	Differences	IF THEN rules	Natural language expressions	Semantic meaning in relationships	Saliency maps	Interactive Explanation interface	Partial dependence plots	Graphs	Deep Taylor decomposition	Confidence	Distance	Proximity	Odds ratio	Activation	Algorithm Statistics	Permutation Importance	Normalcy Exemplar	Tree explainability	Characteristics of cluster	Meta model approximation (ACE)	LIME	SHAP	Localization	Parzen
Trees	Decision tree	1	[47]	2	3	1	2	5	5	2	2	2	5	1	5	5	2	5	1	2	2	[23]	5	2	[102]	2	5	[102]
	Random forest	[43, 78]	[78]	2	3	1	2	5	5	[130]	[130]	2	5	1	5	5	2	5	1	2	2	[43]	5	2	[102]	[43]	5	[102]
	Isolation forest	[3, 110]	[110]	2	3	1	2	5	5	[130]	[130]	[46]	5	1	5	5	2	5	1	2	2	[37, 18]	5	2	2	2	5	3
	Balanced random forest	1	1	2	3	1	2	5	5	3	4	2	5	1	5	5	2	5	1	2	2	1	5	2	2	2	5	3
	Minimum spanning tree	1	1	2	3	1	2	5	5	3	2	2	5	1	5	5	2	5	1	2	2	1	5	2	2	2	5	3
	Gradient boosted tree	1	1	2	3	1	2	5	5	3	2	2	5	1	5	5	2	5	1	2	2	1	5	2	2	2	5	3
	Frequent pattern tree	1	1	2	3	1	2	5	5	3	2	2	5	1	5	5	2	5	1	2	2	1	5	2	2	2	5	3
	C4.5	1	1	2	3	1	2	5	5	2	2	2	5	1	5	5	2	5	1	2	2	1	5	2	2	2	5	3
Support vector machine	Regular SVM	1	[47]	4	2	1	2	5	2	[34]	1	1	2	2	3	3	2	5	1	2	3	5	5	2	[102]	2	3	[102]
	One class SVM	[3]	1	4	2	1	2	5	2	1	1	[81]	[55]	2	3	3	2	5	1	2	3	5	5	2	2	2	3	3
Local outlier factor	Regular LOF	[131]	1	2	2	3	3	5	5	2	2	[8, 131, 46]	5	1	1	1	2	5	1	2	1	5	5	2	2	2	5	3
	Clustering-based LOF	1	1	2	2	3	3	5	5	2	2	1	5	1	1	1	2	5	1	2	1	5	5	2	2	2	5	3
Neural network	Probabilistic NN	1	1	3	3	3	2	5	2	3	3	2	2	2	4	4	2	3	1	2	2	5	4	2	2	2	3	2
	(Deep) NN	[5, 12, 2, 43]	3	3	[5, 74, 2]	5	2	3	3	2	2	[5]	[74]	4	2	3	[29]	2	2	5	4	2	[102]	[43]	3	[102]	2	2
	Recurrent NN	1	1	3	3	3	2	5	2	3	3	2	2	2	4	4	2	3	1	2	2	5	4	2	2	2	3	2
	Autoencoder	[110, 131]	[110]	3	3	3	2	5	[21]	3	3	[81, 58, 97, 131]	2	2	4	4	2	3	1	2	2	5	4	[128]	2	[42]	[58]	2
	Variorational autoencoder	[81]	1	3	3	3	2	5	2	3	3	2	2	2	4	4	2	3	1	2	2	5	4	2	2	2	3	2
	Competitive learning network	1	1	3	3	3	2	5	2	3	3	2	2	2	4	4	2	3	1	2	2	5	4	2	2	2	3	2
	Multi-layer feed forward NN	1	[6]	3	3	3	2	5	2	3	3	2	2	2	4	4	2	3	1	2	2	5	4	2	2	2	3	2
	RBM	1	1	3	3	3	2	5	2	3	3	2	2	2	4	4	2	3	1	2	2	5	4	2	2	2	3	2
	Long short term memory	[20, 110, 88]	[110]	3	3	3	2	5	2	[34]	3	2	2	2	4	4	2	3	1	2	2	5	4	2	2	2	3	2
	CNN	[80]	[80]	3	3	3	[80]	5	2	3	3	2	2	2	4	4	2	3	1	2	2	5	4	2	2	2	3	2
	Self-organizing map	1	1	3	3	3	[32]	5	2	2	3	2	2	2	4	4	2	3	1	2	2	5	4	2	2	2	3	2
	GAN	1	1	3	3	3	2	5	2	[108]	3	2	2	2	4	4	2	3	1	2	[108]	5	5	2	2	2	[108]	3
	Hierarchical cluster-based deep NN	1	1	3	3	3	2	5	2	3	3	2	2	2	4	4	2	3	1	2	2	5	4	2	2	2	3	2
Clustering	Cluster analysis	2	2	3	2	3	2	5	3	2	3	1	4	2	1	1	2	5	1	2	1	5	1	2	2	2	2	3
	Spectral clustering	2	2	3	2	3	2	5	2	2	3	1	4	2	1	1	2	5	1	2	1	5	1	2	2	2	2	3
	K-means	2	2	3	2	3	[32]	5	3	2	3	1	4	2	1	1	2	5	1	2	1	5	1	2	2	2	2	3
	Euclidian distance	2	2	3	2	3	2	5	3	2	3	1	4	2	1	1	2	5	1	2	1	5	1	2	2	2	2	3
	Expectation-maximization	2	2	3	2	3	2	5	3	2	3	1	4	2	1	1	2	5	1	2	1	5	1	2	2	2	2	3
	Gaussian mixture modelling	2	2	3	2	3	2	5	3	2	3	1	4	2	1	1	2	5	1	2	1	5	[71]	2	2	2	2	3
	DBSCAN	2	2	3	2	3	2	5	3	2	3	1	4	2	1	1	2	5	1	2	1	5	1	2	2	2	2	3
Bayesian networks	Naïve Bayes	3	2	2	3	3	2	5	2	3	3	2	5	1	5	5	2	5	1	2	3	5	5	2	2	2	3	3
	Bayesian model	3	[47]	[72]	3	3	2	5	2	3	3	2	5	1	5	5	2	5	1	2	3	5	5	2	2	2	3	3
Other	Principal component analysis	2	2	4	2	3	2	5	2	3	3	[8]	5	2	5	5	2	5	1	2	3	5	5	2	2	2	2	3
	(Multi-nomial) Logistic regression	2	[47]	3	2	3	2	5	3	2	3	2	5	2	5	5	[43]	5	1	2	3	5	5	2	[102]	[43]	2	[102]
	Discriminant analysis	2	2	4	2	3	2	5	3	2	3	2	5	2	5	5	2	5	1	2	3	5	5	2	2	2	2	3
	OneR	1	1	1	2	1	1	5	4	4	3	4	5	2	5	5	2	5	1	2	2	5	5	4	4	4	4	4
	Hidden markov model	2	2	3	2	3	2	5	3	2	3	2	5	2	5	5	2	5	1	2	2	5	5	2	2	2	2	2
	Graph mining	2	2	2	2	2	2	[56]	3	2	3	2	5	2	2	2	2	5	1	2	2	5	5	2	2	2	2	2
	Link analysis	2	2	2	2	2	2	3	4	3	3	2	5	2	5	5	2	5	1	2	4	5	5	4	4	4	4	4
	Peer group analysis	2	2	3	2	3	2	5	3	2	3	2	5	2	5	5	2	5	1	2	1	5	5	2	2	[113]	2	3
	Angle based outlier detection	2	2	3	2	3	2	5	4	2	3	2	5	2	3	3	2	5	1	2	2	5	5	2	2	2	2	3
	Salient Object detection	4	4	4	4	4	4	5	1	2	3	2	5	2	5	5	2	5	1	2	5	5	5	4	3	3	2	3
	Kernel density estimation	2	2	3	2	3	2	5	4	3	3	2	5	2	3	3	2	5	1	2	3	5	5	2	2	2	2	1

Table 6: Assessment matrix between algorithms and explanation mechanisms

4 Research approach

This section describes the approach of this research. It describes all the steps that were taken within this research and a short motivation why certain steps were taken. This project is executed for the Dutch Government at the Dutch Central Government Audit Service (DCGAS) within the data analytics team.

This research consists of two different parts. Figure 7 describes the different steps of both approaches and how they connect. The first part concerns the use of data and unsupervised outlier detection algorithms to detect outlying transactions. For this part, the CRISP-DM methodology is used and this part is referred to as phase 1. This phase answers the first research question and its sub questions. The second part concerns the design and prototyping of an explanation facility for financial auditors. The prepared data and model from phase three and four of CRISP-DM is used for the treatment design. The design of the facility is used to validate the outcomes of the first phase. The second phase answers the second research question. These two phases concern different types of research, namely a data mining problem and a design science problem. Therefore, it is decided to use two different approaches and evaluate both in the final step. Both phases and its evaluation answer the third research question.

1. To what extent can unsupervised outlier detection algorithms help to identify potential financial fraud in invoices of the public sector?
    1.1. What features are important for financial fraud detection (identified by domain experts)?
    1.2. What unsupervised outlier detection algorithm delivers the most promising results on the invoice data of the public sector?
2. How can an explanation facility, aimed at explaining the algorithm and its outcomes to a financial auditor, be structured?
    2.1. What is the purpose of the explanation facility for the financial auditor?
    2.2. What explanation mechanisms should be included in this facility?
3. How can the models and explanations contribute to the assessment of the reliability of the financial statement?

Section 4.1 specifies the scope of this research to clarify what is included and excluded. Section 4.2 describes the first four steps of the CRISP-DM approach. Section 4.3 describes the two iterations of the design science methodology. Section 4.4 evaluates the results of both approaches as the fifth step.

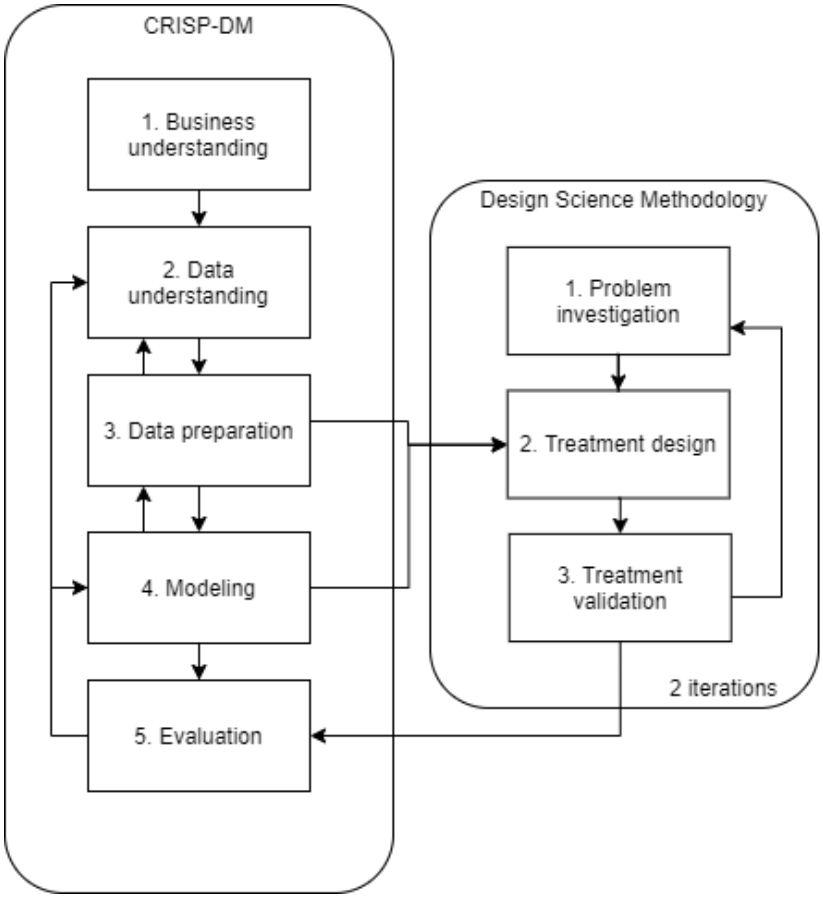


Figure 7: Overview of the approach of this research

4.1 Scope

The scope of this research is defined in this section to clarify why certain topics are not included in this research. The duration of this study was 6 months. Due to the resources, it was decided to include three promising outlier detection algorithms. Furthermore, it was decided to focus on the explanation mechanisms in this research as this is a novel research area. Due to the lack of research on explanation mechanisms in the financial sector, as shown in Section 3, this research more to this research area. Opposed to the area of fraud detection by using outlier detection methods, which have been researched more frequently.

The data that is used for this study is the available data from the DCGAS. This is an elaborate data set which provides a good basis. The participants that took part in this research are financial auditors from the DCGAS. Due to the scale of this research, it is advised to focus on one subgroup. The origin of the data and the participants means that the scope of this research is focused on the public sector which can influence the results.

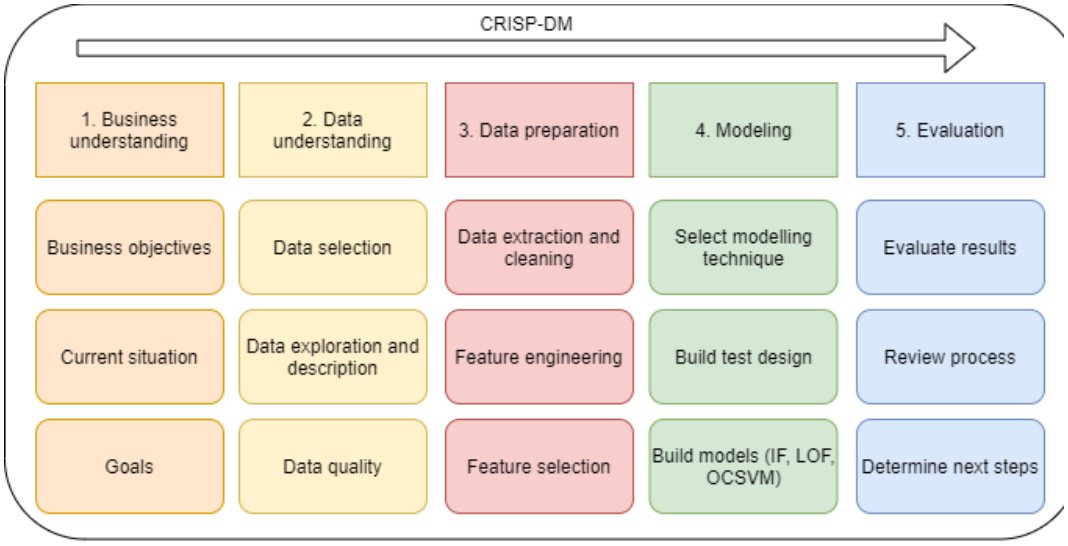


Figure 8: The applied stages of CRISP-DM in phase 1

4.2 Phase 1: Detecting outlying transactions

The methodology that is applied in this phase of the research is CRISP-DM. This is a methodology developed for experiments in data mining [126]. It consists of six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. This research focuses on the first five phases, as shown in Figure 8. The sixth phase is deployment and recommendations are made for this in Section 7. This methodology is chosen for this phase as it is a leading process model for data scientists. Its focus on understanding the project objectives and requirements from a business perspective and translating this into a data mining problem make it fit for the aim of this research phase. Furthermore, previous research has indicated that the methodology is complete, cost-effective and focuses on best practices [9, 77]. Figure 8 describes the different steps of this phase. Some steps of CRISP-DM have been put together into one to keep the overview compact. Below are elaborations on the different steps of this phase.

4.2.1 Business understanding

4.2.1.1 Business objectives

The DCGAS reviews the financial and IT systems of the Dutch government. The goal is to ensure that everything is performed according to the accounting principles. These principles differ per country and the government has their own manual on this subject. The financial auditors are responsible for examining the reliability of the financial statements [124]. One of the methods for this is reviewing a sample of invoices on their correctness as it is not feasible to review all invoices.
When it is not possible to review all invoices, one could look for the transactions with a

higher risk of being incorrect and potentially fraudulent. Transactions that differ from the rest are usually worth to manually review. The task of selecting 'risky' transactions can be automated by using machine learning algorithms to find outliers in the transactions. Outliers are data points that differ significantly from the others and are thus more likely to be incorrect.

The auditors are not responsible for detecting fraud specifically. However, when they encounter fraud, it is considered a risk and reported. This means that when the selection of the sample is automated and outlying samples will be reviewed, the likelihood to detect potential fraud is higher. This could help discover high profile (fraud) cases such as at the Rijkswaterstaat [82]. Rijkswaterstaat found out in 2019 that an employee had submitted false invoices, worth 2.3 million euros. One of the factors that enabled the employee to commit fraud for such a period of time, was that the employee filed small invoices. These are usually not reviewed but automatically processed and verified on a number of rules. Small invoices usually have a lighter review process.

The objective is to automate the selection of invoices that are included in the sample and thereby review the most outlying invoices. This will contribute to the reliability of the financial statement and provide a better and more objective overview of the invoices. This can contribute to the correctness of the financial year statement.

4.2.1.2 Current Situation

Currently, the auditors use statistical methods to determine the sample of invoices they will review and manually select these. They review the invoice according to the accounting principles. It is not possible to review all invoices and thus a sample suffices. However, as only a small sample is reviewed, there is a higher chance that incorrect or fraudulent invoices will pass.

The financial statement is accepted under the agreement that 95% of the money is accounted for. Thus, there remains a chance that fraud is not detected. This can still be a huge amount of money and makes it worth to research methods of helping the auditors.

There are a few requirements and constraints to this case that are described below.

- The ratio of outlying transactions to normal transactions is very imbalanced.
- There is no manually labelled training set yet.
- It is important to prevent false negatives as the data is sensitive.
- The results and model should be explainable to the financial auditors. Without background knowledge, the auditor needs to understand what the algorithm does and trust the outcomes.

4.2.1.3 Goals

To achieve the business objectives, certain data mining goals have been set. It is important to discover what algorithm can find outlying transactions most accurately. Next, the algorithm should be able to provide explanations to the financial auditors. Some mechanisms have been equipped to achieve this. This means that this research will consists of experiments on data sets to review the most accurate unsupervised outlier detection algorithms.

These goals have been achieved by applying several algorithms to the data set provided by the DCGAS and validating the outcomes with financial auditors.

4.2.2 Data understanding

In this phase, the aim is to understand the data better to discover what preparation is needed. Different types of analysis have been performed and are described below.

4.2.2.1 Data selection

The available data consists of the transactions made by the public sector and are recorded in ERP systems of the government. There is access to the transactions of seven departments over 2019 which contain around 800.000 transactions. These transactions are automatically collected from the ERP(SAP) systems that the departments use. It is available through a structured database (MS SQL server).

4.2.2.2 Data exploration and description

The first description of the data can be found in appendix 9.1. It describes the two databases that are used including their attributes, format and descriptions.

The supplier master data (SMD) database contains 38 columns, the accounts payable (APA) database contains 91 columns. Very little numerical data is available, either discrete or continuous. Most of the columns are either nominal or binary. Ordinal occurs mostly due to different times and dates.

Appendix 9.2 contains Table 18 and Table 19 which describe the initial analysis of the data. The number of data points, empty values and unique values is described. It is shown that certain features have a very high percentage of empty values. The values either incline to 100 percent or are very low. Furthermore, it is shown that most nominal columns have limited categories.

4.2.2.3 Data quality

DAMA UK is a community of data professionals that have developed several, widely accepted data management strategies. They have also created a list of six dimensions to assess the quality of data [25]. These are completeness, uniqueness, timeliness, validity, accuracy and consistency. Completeness has been evaluated in Table 20 and Table 21 by evaluating

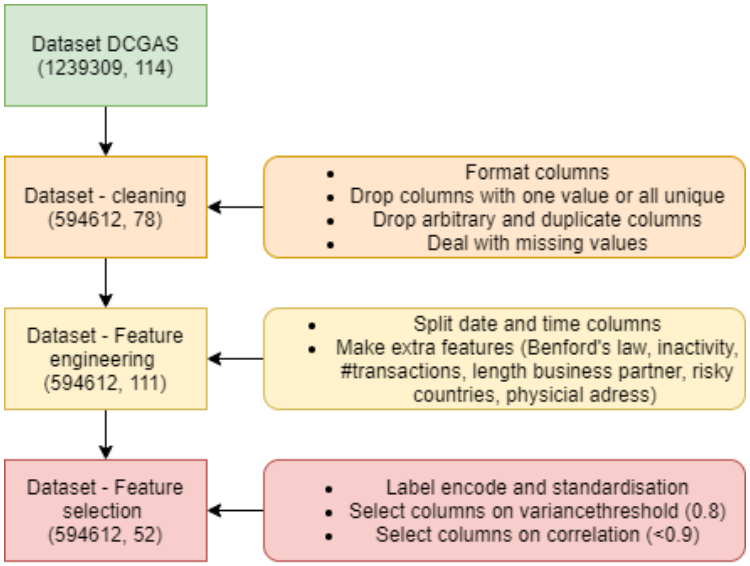


Figure 9: The phases of data preparation

the percentage of empty values for each attribute. Next, the data was also reviewed to see the different values that appeared in each attribute. This was done to see if an empty value meant that the data was not available or it had meaning. Uniqueness has been identified through the count of unique values. This is also supported in the feature selection by variance in the feature selection phase in Section 4.2.3.3. Timeliness constitutes the degree to which data represent reality from the required point in time. The dataset contains all the invoices of the year 2019, confirming its timeliness and its recentness. Data is valid if it conforms to the syntax (format, type, range) of its definition. Most columns were in the right format and could be reformatted where needed. Consistency means data across all systems reflects the same information and are in sync with each other. As only the data from one ERP system is used, this dimension does not apply.

4.2.3 Data preparation

In this phase, the data was prepared to be used as input for the algorithms. The different phases are described below. Figure 9 gives an overview of the different steps and the resulting data set.

4.2.3.1 Data extraction and cleaning

Both databases are used for this research project. The APA database with the transactions is used as a base and the information on the business partners is added to each transaction. The data is extracted from the MS SQL server and loaded into a Pandas dataframe. From this extract, the transformations are applied.

In this phase, the data was reviewed and further cleaned for processing. The following actions have been applied to the data to increase its quality.

1. Entire empty columns are dropped.
2. Some columns are arbitrary and are dropped. These columns can be derived from others and do not gain any extra information.
3. Columns with either one value or only unique values are dropped.
4. Empty values are correctly formatted. E.g. instead of an 'X', a NaN is used.
5. All columns are formatted to the correct type such as integer, float, datetime etc.
6. If columns have more than 50% empty values, they are dropped.
7. Records that are duplicates are removed from the data.

The indicators that are not numerical are transformed to numerical features by using a label encoder. This preprocessing step ensures that all categorical data is transformed to an integer and can thus be applied to all algorithms.

Another crucial activity is to standardise the features. Especially the distance-based algorithms are sensitive to features of different scales. Standardisation helps to get all features on the same scale and make fair comparisons. Furthermore, standardisation is not sensitive to outliers and does not require the assumption of a normal distribution, as opposed to normalisation [28].

4.2.3.2 Feature Engineering

Additional features are constructed for the data. Section 2 describes different features of fraud that are found in literature. As there are only limited methods of selecting useful features, it can be fruitful to consult domain experts. The authors of [106] have established that the use of features indicated by domain experts increased the accuracy of the algorithms. In Section 2, literature was consulted on potential red flags in transactions. The results from this study are combined with the features that are recorded in the data from the DCGAS. This resulted in a list of potential indicators. These were incorporated in a survey that asks domain experts to judge whether these would indicate if a transaction is outlying and whether it could indicate fraud. By making use of a survey, auditors outside of the DCGAS could also be included and much more data could be collected compared to interviews.

This survey is created by using Qualtrics software and is distributed among financial auditors and serves as basis for the feature selection. The content of the survey, including the informed consent can be found in appendix 9.5. It was decided to use Likert scales to collect ordinal data on the effectiveness of the features [11]. These scales have proven to be useful

for comparing statements and should be analysed using the mean and the median or mode. All three are calculated to properly compare the results of the individual features.

The first part gathered general information about the participant such as sector, years of experience, experience with fraud etc. The second, third and fourth part asked about their opinion on different indicators and whether these could predict outliers and fraud. The participants have the opportunity to elaborate on their choice. The second part concerned indicators about transactions details, the third about time and date, and the fourth part concerned indicators about the business partner. The fifth part asked about the fraud triangle and gave the opportunity to deliver their own indicators.

Based on the results of the survey, the features that are ranked as the most effective are constructed for the data set provided by the DCGAS, when not already available. The first constructed feature is Benford's law applied on the amount. This is a probability distribution that states that the first number appears according to a uniform distribution. The distribution of the amounts of all transactions is calculated and the transactions with amounts that do not adhere to the uniform distribution are marked. Previous research has indicated that this delivers good results for detecting fraud in transactions [7, 98].

Furthermore, the period in which the Government has worked with a supplier and thus received invoices of a business partner is recorded. The date when a business partner is registered for the first time is used and the difference with the current time is calculated. Another feature that is constructed is the number of transactions that a business partner has performed within a year. Whether the business partner is from a risky country, according to the European Commission and whether a physical address is recorded are also constructed as features. Finally, the length of inactivity of each business partner is recorded. This is the difference between the selected transaction and the most recent, previous transaction.

The code that was created to construct these features can be found in Appendix 9.3.

4.2.3.3 Feature selection

Not all features will be used as some are less useful. First, a feature selection method will be selected to discover what the most attributing features are. The authors of [109] have reviewed different unsupervised feature selection methods. They concluded that the selection method will increase the accuracy in unsupervised methods but is heavily dependent on the algorithm. As unsupervised learning limits the feature selection methods significantly, only a few simple methods can be used to limit the amount of features. A variance threshold is applied to lower the number of features. Columns are dropped when more than 80% of their values is the same value. Furthermore, correlation was used to find columns that correlated for more than 90%. Highly correlated columns provide information that is very similar, therefore one can be dropped. Both these feature selection methods are aimed at reducing the number of features to increase the results and make the algorithms run smoother.

4.2.4 Modelling

In this phase, the models were selected, described and its parameters are determined.

4.2.4.1 Select modelling technique

There are a lot of invoices available and the features are in different data types. Most features are categorical and few are numerical. The literature review in Section 2 resulted in a list of 45 algorithms that are used in the financial domain for unsupervised outlier detection. The four most popular algorithms from varying categories are Isolation Forest, Local outlier factor, one-class SVM and an autoencoder.

These models are all based on different assumptions. Isolation forest utilises the fact that anomalies are few and differ significantly from the inliers [130]. Local outlier factor is based on the assumption that outliers have a substantially lower density than their neighbours [39]. OCSVM assume that the training data only has inliers and are considered normal [50]. Low frequency of outliers are allowed in the training data. Autoencoders denoise the input again and thus require the input to be robust and stable. Due to time and resource constraints and the complexity of the algorithms, it is decided that the first three are applied.

4.2.4.2 Build test design

The first test design concerns the algorithms. An experimental setup with data and experiments is sufficient. However, as the data set is not labelled, the model cannot be tested by comparing the predictions to the values of a test set. Therefore, a different approach was used. The DCGAS provided a list of outlying invoices based on certain features. With these invoices, the outcomes of the three algorithms were reviewed.

Box plots of all three algorithms are created to see how outlying the actual outliers were scored. The best performing algorithm is selected and a selection of predictions is manually labelled with financial auditors. The top 100 most outlying invoices were presented to the auditors. The auditors indicated if they would look into these invoices.

4.2.4.3 Build models

Table 7 describes the different parameters and settings from the different models. These are based on the default settings of the chosen implementation. The motivation for this is that these settings are proven to work best in general situations [90]. As this is an unsupervised problem, it is not possible to refine the settings based on the performance. Therefore, the default and general settings are chosen based on the implementation.

The implementation of the algorithms was done by using the implementation of scikit-learn [90]. Their algorithms have proven to have a good performance and are widely used and recognised for this.

Algorithm/mechanism	Implementation	Parameters	Settings
Isolation Forest	sklearn.ensemble	n_estimators	default = 100
		max_samples	auto = min(256, n_samples)
		contamination	auto
		max_features	max_features
		bootstrap	False (sampling without replacement)
		n_jobs	None
		random_state	np.random
		verbose	[1:3]
		warm_start	False
Local Outlier Factor	sklearn.neighbors	n_neighbors	default = 20
		algorithm	[ball_tree, kd_tree, brute]
		leaf_size	default = 30
		metric	[minkowski]
		p	default = 2
		metric_params	None
		contamination	auto
		novelty	False
		n_jobs	None
OCSVM	sklearn.svm	kernel	[rbf]
		degree	default = 3
		gamma	scale (1/(n_features*X.var())
		coef0	default = 0
		tol	default = 1e-3
		nu	default = 0.5
		shrinking	True
		cache_size	default = 200
		verbose	False
		max_iter	default = -1

Table 7: Parameters and implementation of the algorithms and mechanisms

4.3 Phase 2: Design an explanation facility

The research approach for this phase is based on the design science methodology (DSM) developed by the authors of [125]. This methodology is aimed at solution-oriented research. It studies the artefact in context. In this part of the research, the explanation mechanisms in the context of accounting are researched. An artefact is designed, specific for financial auditors. This artefact is validated and refined based on the feedback. Due to this being a design problem, this specific methodology is fitting for this phase. Furthermore, the design science methodology is iterative by nature. Figure 10 elaborates all the different phases of this methodology. The design cycle is part of the DSM and is used to answer knowledge questions about the artefact in the context. Below, each section represents one of the phases and will describe the proposed research approach.

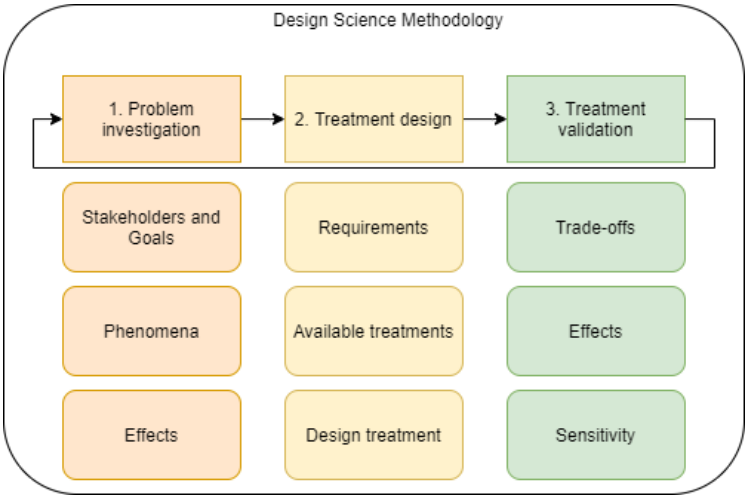


Figure 10: The applied stages of DSM in phase 2

4.3.1 Problem investigation: Iteration 1

The following findings and assumptions were found regarding the interaction between financial auditors and algorithms in general. Three meetings were held with financial auditors that have been employed for the DCGAS over ten years. This was an unstructured interview where they were asked to share their experience with financial fraud within the DCGAS. The responses were recorded and analysed to find the requirements.

- Algorithms show promise for use in auditing as they are able to analyse all invoices in a very short time.
- Accountants need to justify their actions.
- When auditors use algorithms, the algorithms need to justify their decisions with explanations.

- Financial auditors are not familiar with machine learning algorithms and need explanation on the origin of their predictions and outcomes.

From these observations, it can be derived that there is a need for explanations of algorithms and their decisions in a language that is adjusted to the level of knowledge on IT for auditors. Once this is adequate, auditors can make use of algorithms.

- **Stakeholders**: Financial auditors, Government
- **Goals**: Enabling auditors to use algorithms in a transparent and justified manner. The authors of [41, 115, 114] have described the criteria of explainability in 7 factors. These include: transparency, scrutability, trust, effectiveness, persuasiveness, efficiency and satisfaction. For this research, the goals are providing transparency, increasing effectiveness and increasing trust.
- **Phenomena**: The algorithm does not provide any explanation when determining if a transaction is outlying.
- **Effects**: The financial auditor cannot understand the algorithm and process and hence will not trust the outcomes. This will lead to the auditor not using any algorithm.

4.3.2 Treatment design: Iteration 1

4.3.2.1 Requirements

First, we get a grasp of what auditors require of an explanation facility and what they wish to see. Therefore, in the first round, we have interviewed two financial auditors and two data analytics employees to discover their requirements.The interviews were held through an online, synchronous video-communication system with financial auditors. The authors of [107] have proven that synchronous, online interviews are as fruitful as live interviews or user trials. It is fruitful to get qualitative data on the needs of the accountants before designing the explanation facility and therefore interviews are chosen. This method is called upstream engagement and has proven to be useful [24, 116]. First, it is discovered how a financial auditor uses data analytics in their process of reviewing transactions. Next, it is discovered with what goal they will use the facility and what is absolutely required in the facility. Finally, different types of explanations are shown to the participants and they are asked to comment on what they would prefer. The design of the interview questions and the different explanation mechanisms shown can be found in appendix 9.6 and 9.7.

The lowest level of Figure 5 contains 7 categories of explanations and accompanying explanation mechanisms. These are shown to the participants in the last part of iteration 1. Each category and its mechanisms are shown on one page with an financial example, as shown in appendix 9.7. Their answers and feedback is collected and summarised.

4.3.2.2 Available treatments

Literature describes the different treatments that are already available. These are discussed in Section 3 and Section 2. However, one of these available treatments cannot fulfil all the requirements found in Section 6. It is therefore decided to use multiple of the existing treatments and combine these into one, new treatment that satisfies the requirements.

4.3.2.3 Design treatment

Based on the requirements found in Section 6, a selection of appropriate explanation mechanisms was combined into a new prototype. First of all, it was decided that the medium of the explanation facility is a web application. This medium provides the ability to have interactive interfaces and is easily accessible for participants that will test it. Furthermore, any data source can be easily connected which creates abilities to work with the real life data set. Streamlit was used to quickly build a web application. This is a package in Python that enables users to quickly develop interactive web applications. The following design decisions were made for the prototype.

- The explanations will be provided per step of the process. This means that there will be an overview, input, model and output page that provide explanations.
- The needs of auditors vary and thus there will be a choice between visual and textual explanations.
- The layout needs to be clean and easy.
- There should be a focus on the financial impact of invoices. Expensive invoices have a higher priority.
- The facility should provide explanations about general mechanisms behind the algorithms applied in the system, as well as explanations about how an algorithm reached to a fraud indication level assignment for each individual invoice.

Interactive visualisations, algorithm statistics and textual explanations are self-explanatory in their workings. However the SHAP values will be explained below.

The goal of SHAP is to explain a prediction of an algorithm by computing the contribution of each feature to the prediction [113]. The theory of SHAP is based on game theory where the feature values are players in a coalition. The prediction is fairly distributed among the players. When computing the Shapley values, the idea is that some players are playing and some are absent. This is then used to describe the importance of each feature. The explanation is specified as:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

where g is the explanation model, $z' \in 0,1^M$ are the simplified features, M is the maximum coalition size and $\phi'_j \in R$ is the feature attribution for a feature j, the Shapley values. Once the values of each feature is calculated, these can be communicated back, on individual transactions levels or on a higher model level. The implementation of SHAP is provided by the authors of [69]. This is the most extensive implementation of SHAP for Python yet.

There can be a variety of complexity in the explanations. This is tested in the treatment validation using A/B testing. Therefore, multiple interfaces are created to use in the usability test. Every web page has a version A and B. Version A is the simple interface with limited options. Version B is more complex and offers multiple options to the user. A description of the different options is found in Table 8.

	Version	
Interface	A	B
Overview	Overview of all invoices	Make a selection of the invoices based on outlier score
Input	Only raw data and textual explanation	Include visualisations and explanations of each indicator
Model	List of sufficient and insufficient indicators	Include graph and confusion matrix
Input	List top 5 important features	Include the feature scores and extra visualisations

Table 8: Summary of the different interfaces and their options

4.3.3 Treatment validation: Iteration 1

A usability test was designed according to the data that needs to collected according to Table 9 and performed with six auditors to receive the data to validate the treatment. The requirements state that there are four goals: transparency, trust, efficiency and satisfaction. The usability test measures the influence of the treatment on these four goals. The goals are translated to indicators and variables that are measurable. These are connected to different collection components. An overview of this is given in Table 9.Table 10 describes the tasks that were given to the participants. The participants were asked to think aloud and the time to complete each task was recorded [19]. The think aloud method has proven to uncover additional data on the opinion of participants on the prototype. Second, the different interfaces were shown to the participants and they were asked for their preference and the reasoning. Finally, they were asked several questions on the goals that could not be measured through the tasks. The questions are found in appendix 9.8.

Goals	**Indicator**	**Variable**	**Collection**
Transparency	Actual understanding of how the system works	Knowledge on presented information	Tasks (1,3,4)
	Perceived understanding of how the system works	Intuition	Tasks (1,3,4), Survey
Trust	Dependable	Represent actual real world information	Survey
	Reliable	Presented information remains the same and understandable	Tasks (4,5,6,7)
	Thrustworthy	Rely on information	Tasks (4,5,6,7), Survey
Effectiveness	Make better decisions	Correct choices	Tasks (5,6,7)
Satisfaction	Enjoyable to use	Likeability	Survey
		Comments	Ratio of positive to negative comments

Table 9: Captured data in validation

Task	Goal	Summary
1	Transparency	Find the used data as input
2	Satisfaction	Explore the indicators of the input
3	Transparency	Explore the general model
4	Trust, transparency	Explore the specific model
5	Effectiveness, trust	Explore the general output
6	Effectiveness, trust	Pick an individual invoice
7	Effectiveness, trust	Explore the individual output

Table 10: Overview of the tasks for the usability test

4.3.4 Problem investigation: Iteration 2

The second iteration is based on the results from the first iteration. This iteration is more compact as only feedback is incorporated. The following problems were identified in the results of the treatment validation.

- The explanation per web page is not clear enough.
- The algorithm metrics are too complicated.
- The connection to their own risk control is not clear enough.

4.3.5 Treatment design: Iteration 2

The problems that are identified are translated into changes in the design of the prototype. The following changes were made to the prototype.

- The confusion matrix is dropped from the model page.
- The explanation of the input and output page are improved.
- The overview page indicates how much the financial impact is of the selection.
- The input data set explanation is set to default expanded.

Furthermore, the results of the preferences on version A and B, as shown in Section 6, is implemented in the web app.

4.3.6 Treatment validation: Iteration 2

The results of the second iteration are validated by using real-world transactions as provided by the DCGAS. The outcomes of the most promising outlier detection algorithm are loaded into the prototype. Three financial auditors were asked to participate in a manual labelling session where they go through the predictions and confirm whether these outliers are correct. These sessions validate if the prototype is complete and can be implemented. The comments given by the auditors are recorded as final feedback.

4.4 Evaluation

This is the final stage of this research and consists of three sub stages.

4.4.0.1 Evaluating results

Within this section, the results were reviewed according to the business goals that are defined in the first step. The implications of the results are discussed. The results of this stage are described in the discussion of this thesis in Section 7. This step also answered the third research question based on the gathered results. All the information that is gathered during the previous interviews is reviewed again and all comments on how the outlier detection algorithms can be used are summarised. Furthermore, in the final treatment validation iteration where the prototype is validated by using real-world transactions, the auditors are explicitly asked how they would use the algorithm in their daily work.

4.4.0.2 Reviewing the process

In this section, the entire process of this research was reviewed. Limitations that were met are described and causes are discussed. Possible recommendations are made to further improve in the future.

4.4.0.3 Determining the next steps

This section includes the recommendations for future work. This depends on the output of the models and their business value. A list of possible actions is made to continue this line of research. This is found in Section 7.

5 Design of the Artefact

This section describes the final design of the artefact as designed according to Section 4. The design was created in two iterations and this section describes the final version. The changes made in the second iteration were not drastic, therefore it was chosen to exclude screenshots of the first version. Table 8 describes the interfaces of the first version in text.

5.1 Overview page

Figure 11 shows the homepage of the prototype. There is a banner on the left side that is visible on each page. From there, one can navigate the four different pages. Furthermore, some interesting transactions can be selected as a reminder and there is a short welcome message with some explanation. There is a more elaborate explanation on the overview page itself. Below that is an overview of all the transactions and their outlier score. The slider lets you select a percentage of the most outlying transactions. Some statistics show how many transactions are selected and how much the amount is on the total amount of money.

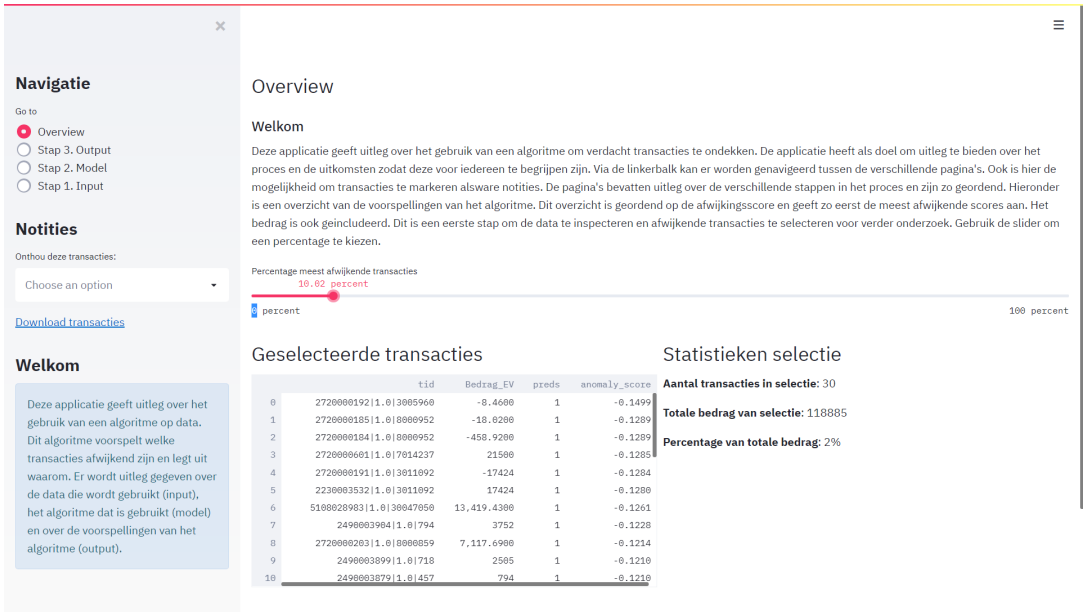


Figure 11: The overview page of the explanation facility

5.2 Input page

Figure 12 shows the top half of the input page. There are several options that you can select. By default, all options are off. You can select to see the textual explanation of the input data and the raw data itself in a table.

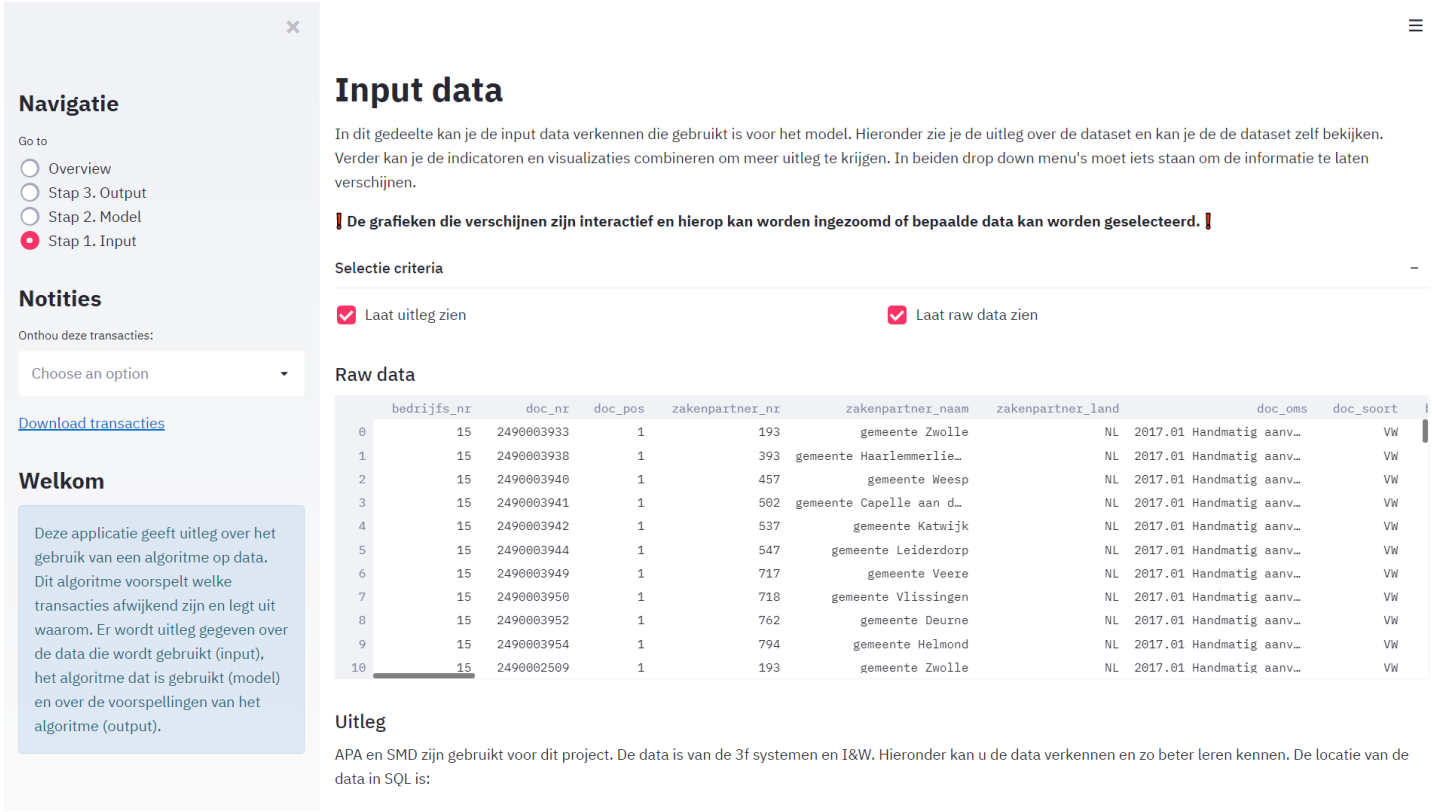


Figure 12: The input page of the explanation facility with the explanation and raw data

Figure 13 shows the bottom half of the input page. In the selection criteria on the top half, you can select any feature that you want to see some explanation about. Furthermore, you can choose from three types of explanations: textual explanation, statistics and a graph of the distribution. All three types of explanation are displayed in Figure 13. All graphs are interactive and anything can be selected and zoomed in upon.

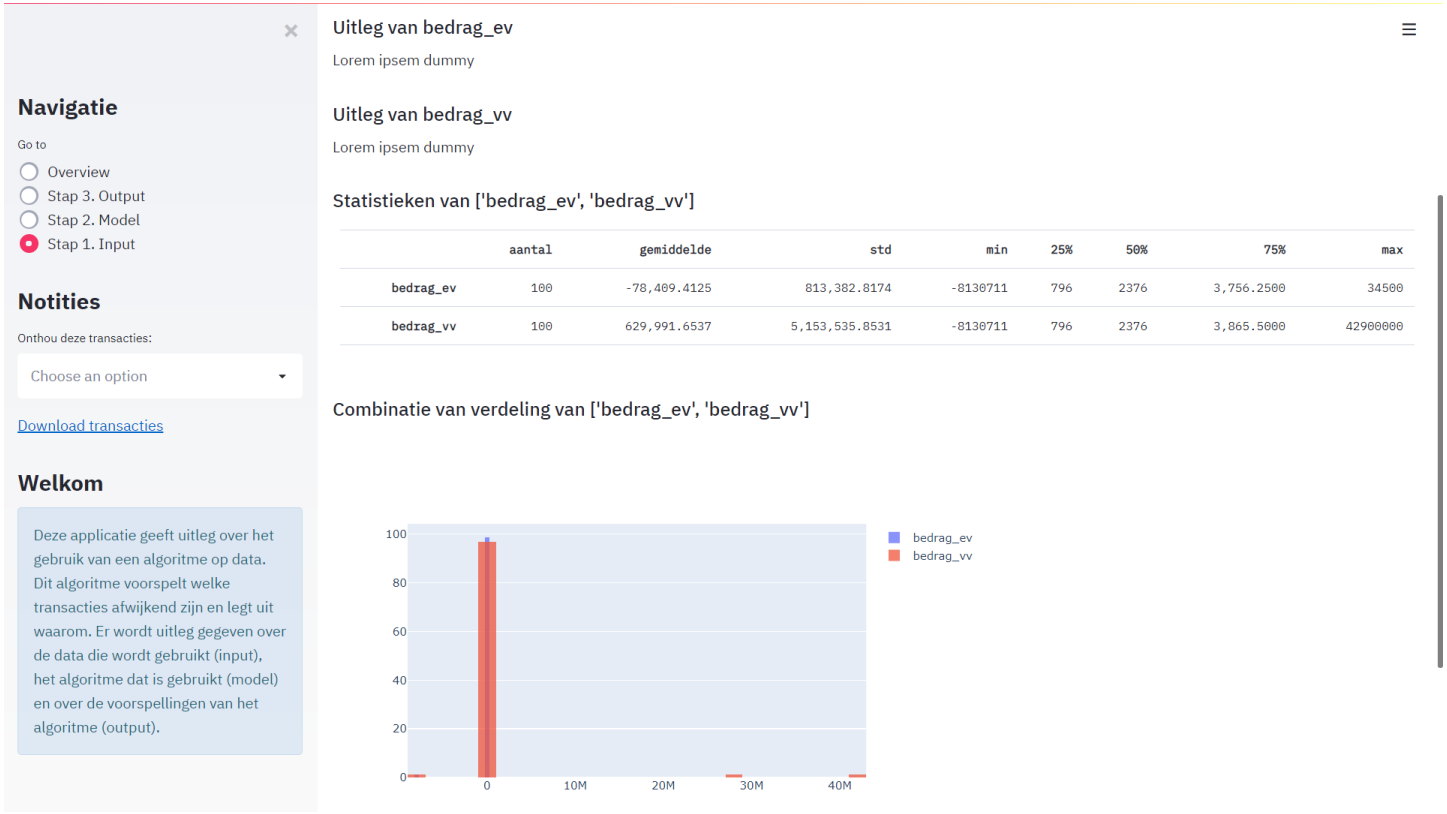


Figure 13: The input page of the explanation facility with the different explanations per indicator

5.3 Model page

Figure 14 shows the top half of the model page. There are two options for explanations: general and project specific. For the general explanation, you can select a textual explanation and a video. For the project specific explanation, you can choose to read the motivation for choosing this model and to see some metrics that describe the reliability of this model. The metrics are shown in a graph with some thresholds to help auditors determine what the value means. Figure 15 shows that each metric is also separately mentioned and when clicked upon shows additional explanations.

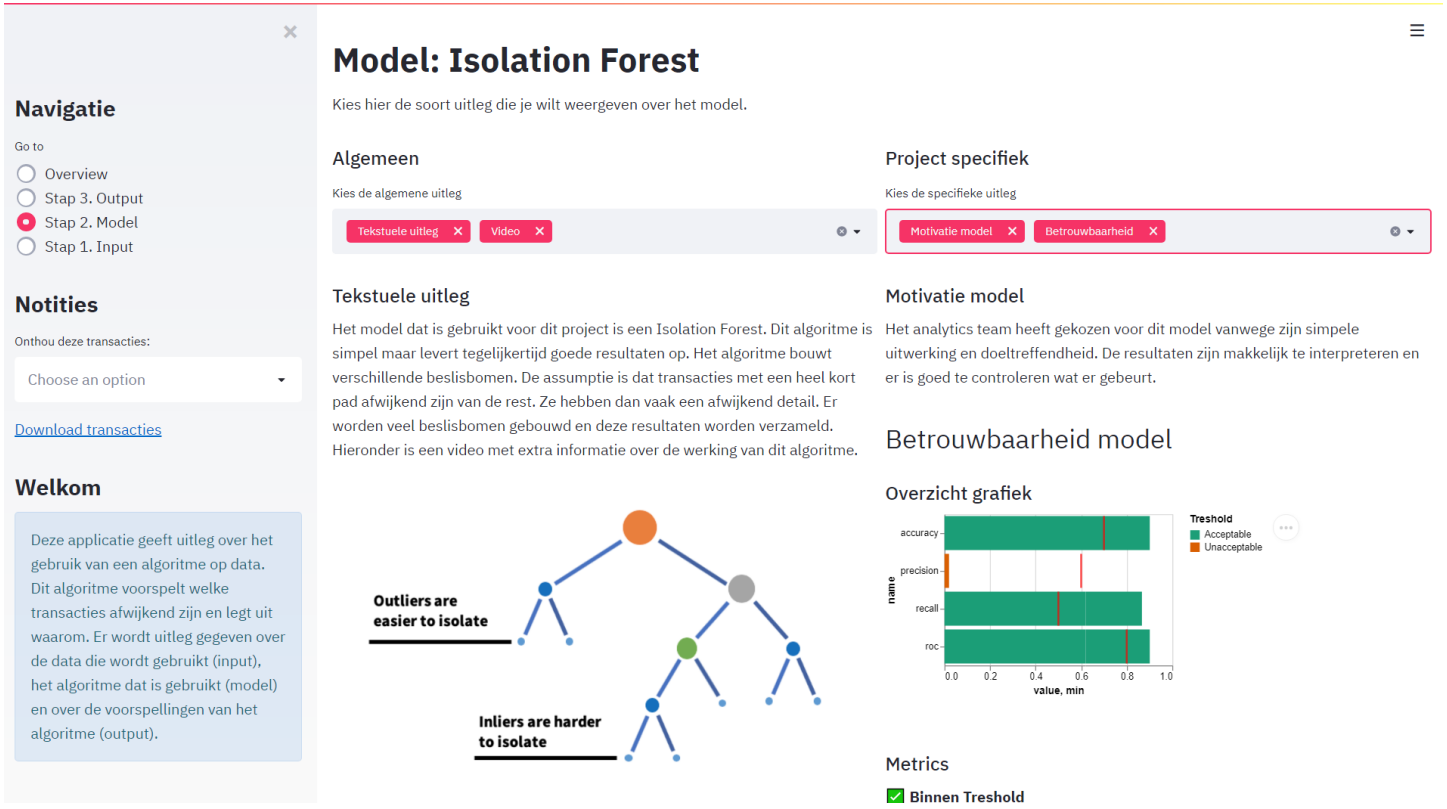


Figure 14: The model page of the explanation facility with the different explanations on the model and motivation

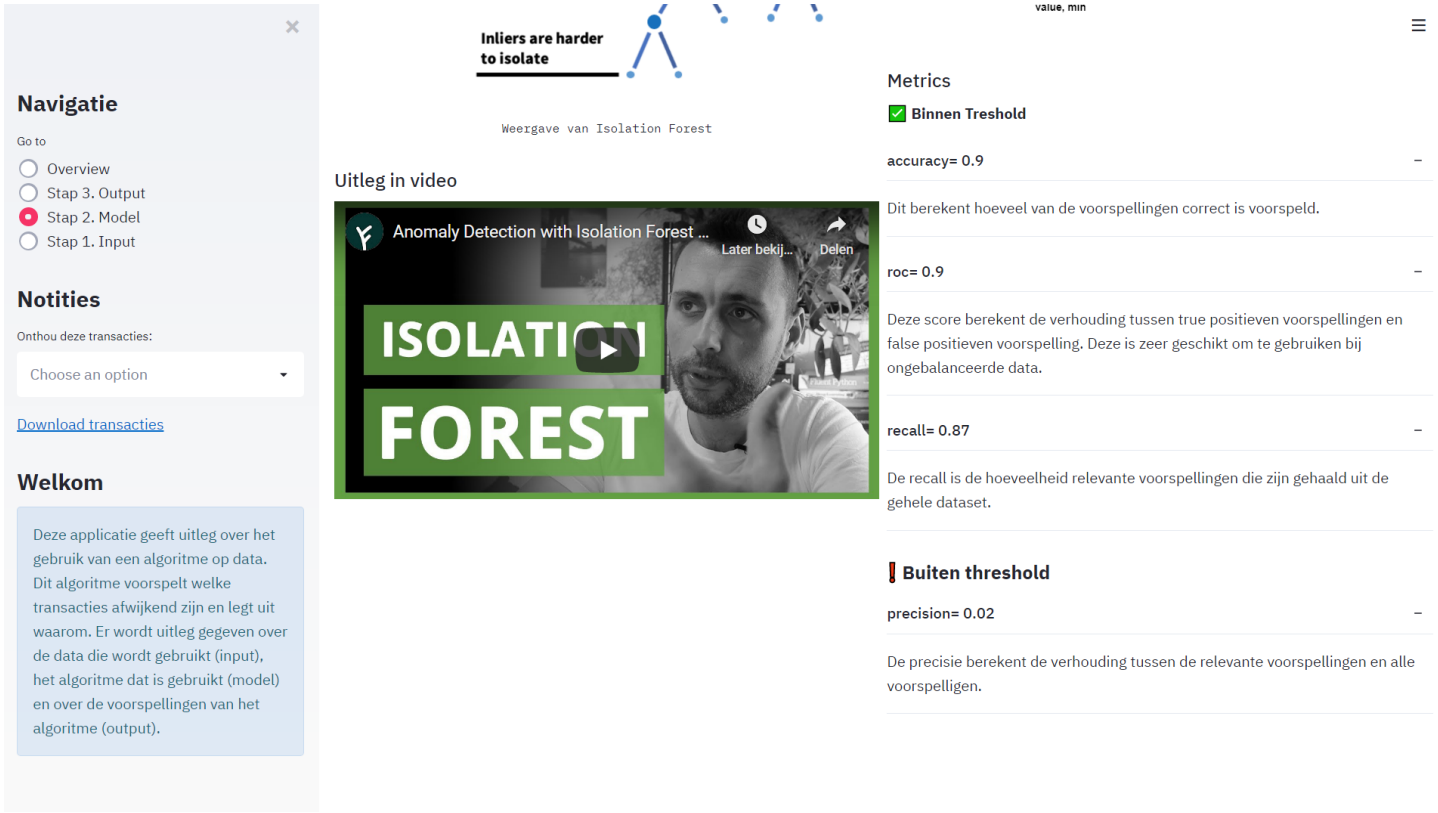


Figure 15: The model page of the explanation facility with the different explanations on the model and reliability of the model

5.4 Output page

Figure 16 shows the output page of the prototype. The user can choose between explanations on the individual transaction level or on the entire model level. Furthermore, they can choose the type of explanation, either the feature relevance or visualisations. Figure 16 shows the feature relevance of the entire model in a graph.

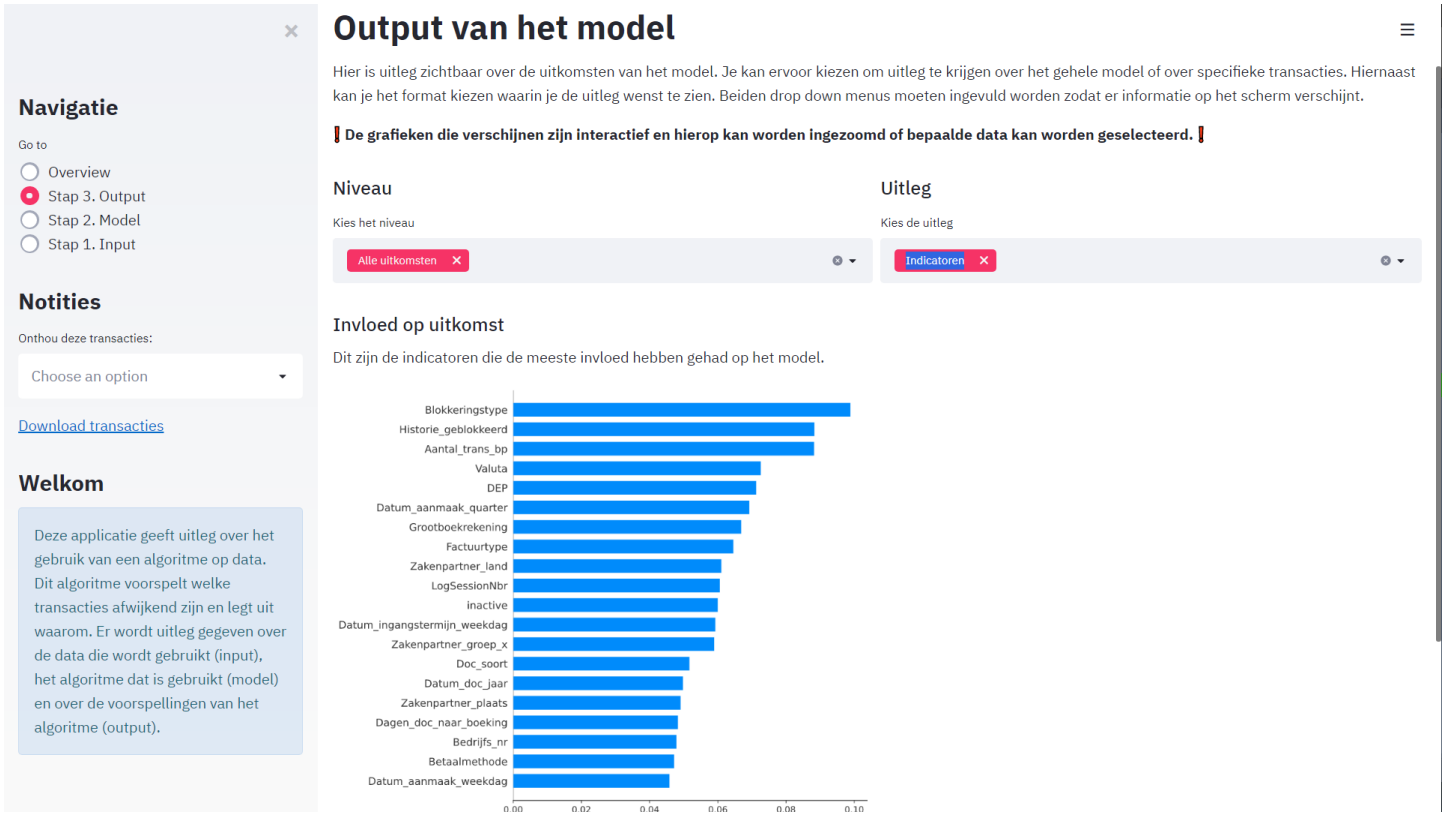


Figure 16: The output page of the explanation facility with the indicator explanations on the entire model

Figure 17 shows the visualisations of the feature "Bedrag_EV" of the entire model. The graph on the left shows the box plot of the outliers and normal transactions. The graph on the right show the different values of the feature.

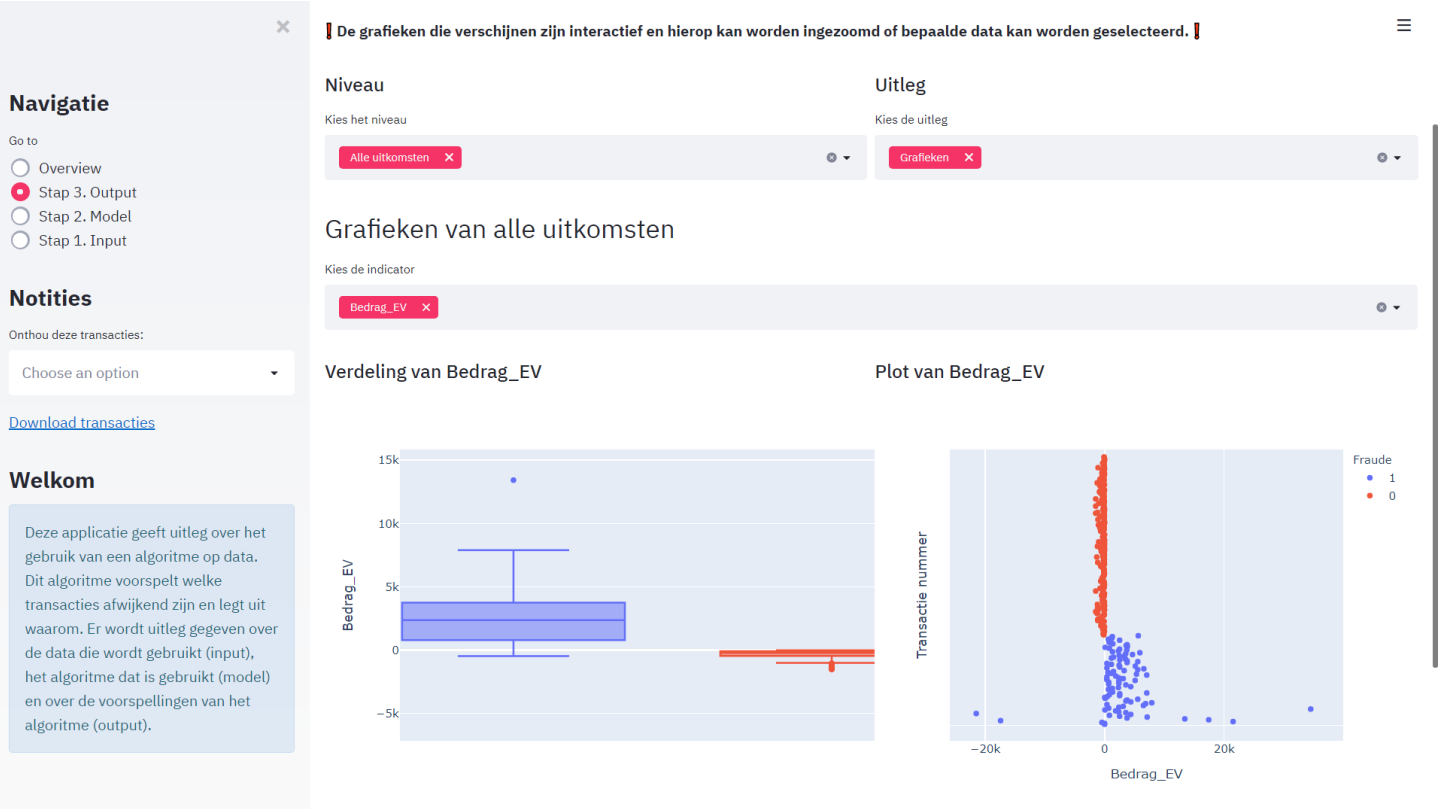


Figure 17: The output page of the explanation facility with the visual explanations on the entire model

Figure 18 shows the feature relevance of a single transaction. The user can select the transaction they want to investigate. The prediction is shown next to the selection box. The user can also review all the features and their values of this transaction. Beneath, the five features are shown that either indicate if the transaction would be normal or outlying.

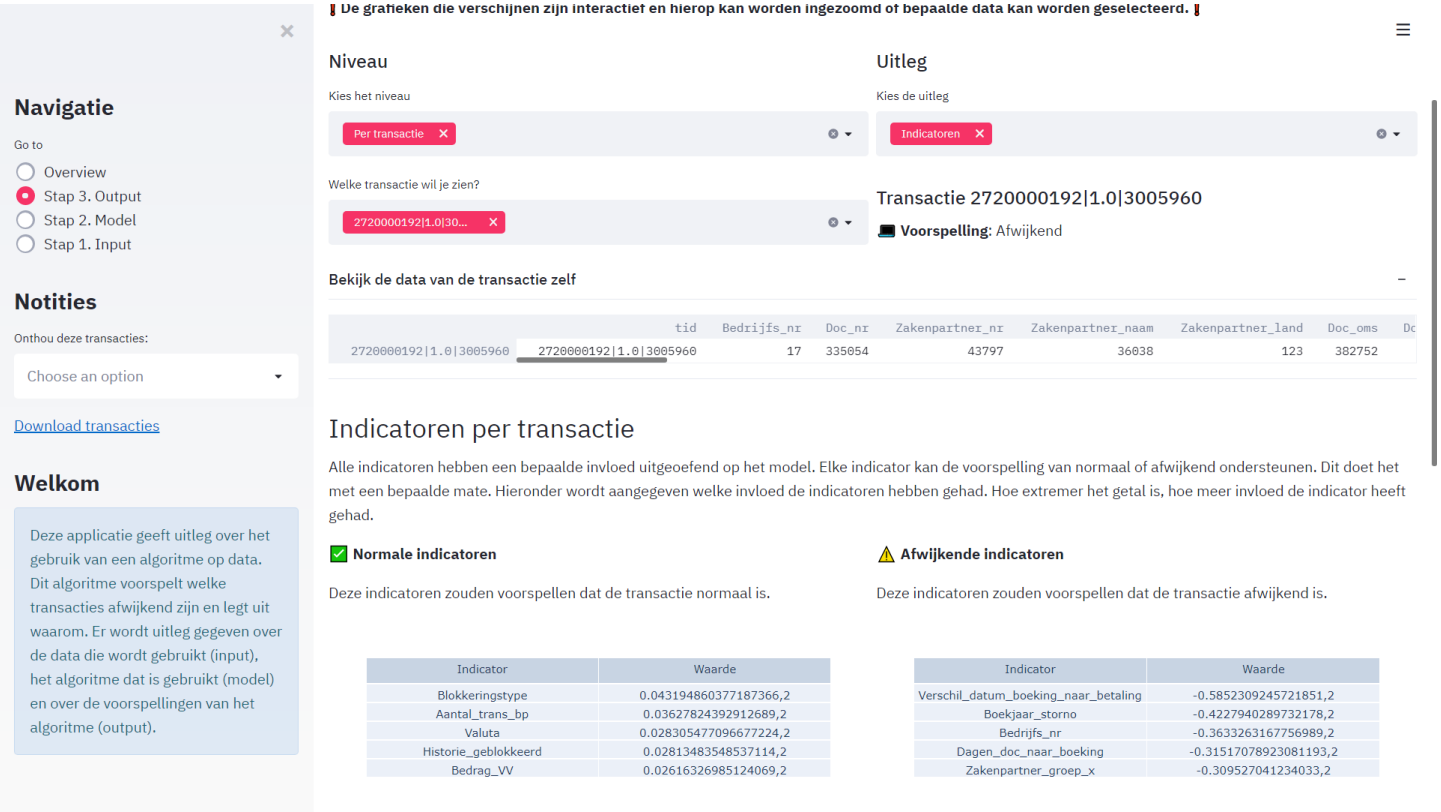


Figure 18: The output page of the explanation facility with the indicator explanations on an individual transaction

Figure 19 shows the visualisations of the feature "Bedrag_EV" of the selected, individual transaction. The graphs are the same as in Figure 17 but the individual transaction is highlighted in green.

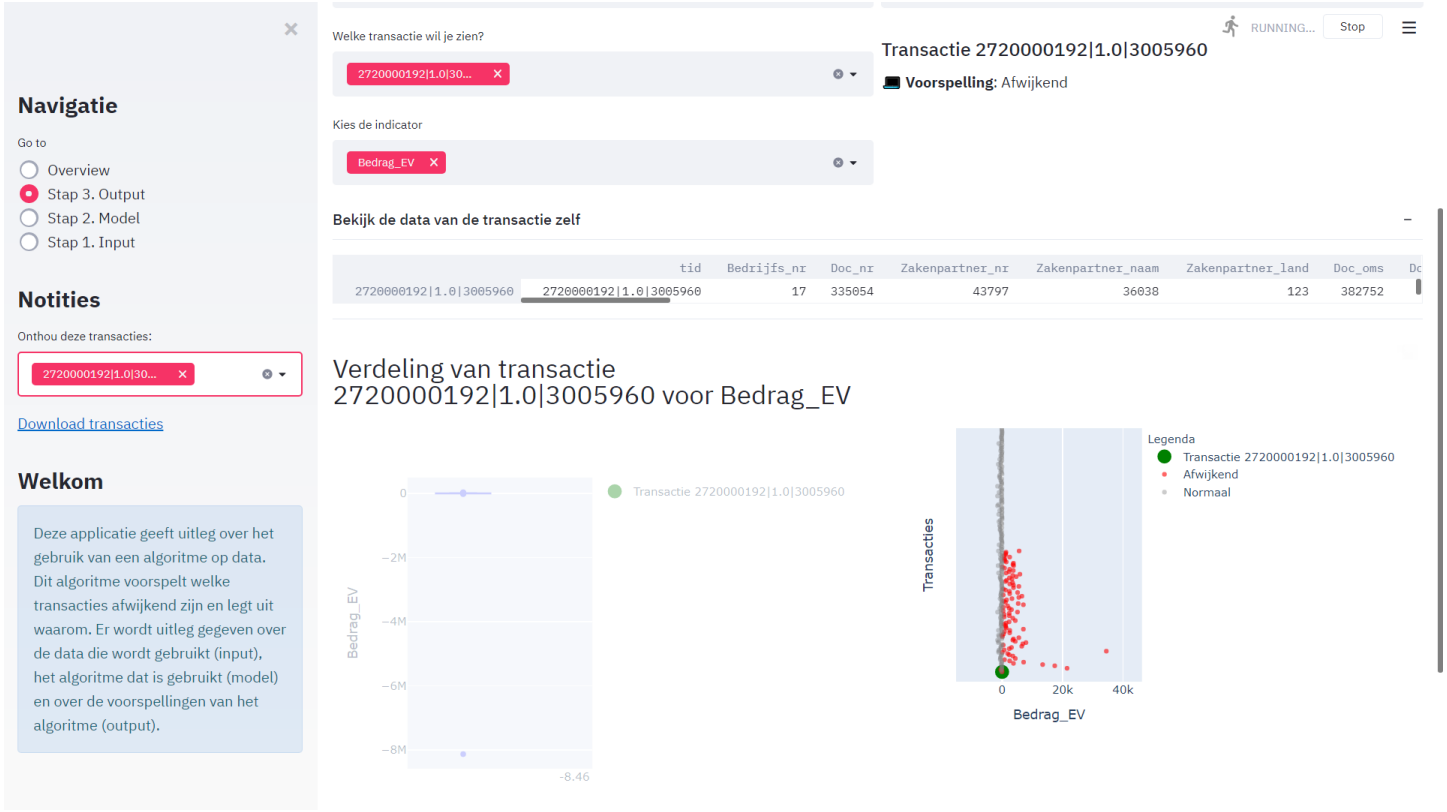


Figure 19: The output page of the explanation facility with the visual explanations on an individual transaction

6 Results

In this section, the results of the research approach is presented. Section 6.1 describes the results from the survey on fraud indicators. Section 6.2 describes the results from the outlier detection on the data of the DCGAS. Section 6.3 describes the results from the design cycle of the explanation facility.

6.1 Fraud indicators

6.1.1 Survey

The survey was distributed among accountants and was divided into five parts. The results are discussed per part in subsections below. The entire survey and informed consent form are shown in Appendix 9.5.

6.1.1.1 General

The number of participants is 28 (n=28). Three of the participants have not completed the survey in the given time of three weeks. These responses will not be included as they are not complete, meaning that the final number of responses included is 25. 18 respondents work in the public sector, 5 respondents work in the private sector. 88% are working as a financial auditor, and 73% are officially registered as accountant ("RA"). The respondents have on average 18 years of experience with accounting and are give their familiarity with fraud in invoices on average a four out of five.

As the number of participants is not very high and the participants are not divided equally over the factors mentioned above, the statistics will be of descriptive nature. No significant changes are measured between different groups of participants.

The overall results of the three categories are described in Table 11 for the three main categories in which the features were divided. Appendix 9.4 shows all the results per indicator. Figure 20 shows all the individual features and their scores. The colour indicates if the features was seen as useful for fraud detection (green), or only as error detection (red). The scale is from 1 to 5, with 1 being least effective and 5 being most effective. Two indications are given for each feature. First, the error indication indicates whether the feature would be useful for detecting errors in the invoice. Second, the fraud indication indicates whether the feature could also indicate fraud, next to the error indication. For the fraud indication, 1 is applicable for fraud, 2 is not applicable for fraud. One of the first observations is that the average effectiveness for the categories differs quite much. Business partner indicators are deemed most effective to detect errors/outliers. After that, invoice details are second and date and time is last. The scores of whether an indicator can be used as fraud indicator are closer to each other. The subsections go deeper into the answers.

Figure 20: The results for each feature from the survey

	Error			Fraud		
Category	Mean	Mode	Median	Mean	Mode	Median
Invoice details	3.16	4	3	1.50	1	1
Date and time	2.57	2	2	1.67	2	2
Business partner	3.53	4	4	1.39	1	1

Table 11: General results on the three main categories

6.1.1.2 Invoice details

The results vary within this category.The most popular quantitative indicators are double payments, large amounts of employee expenses, and large and complicated transactions at the end of the year. In the comments, it is explained that the last indicator is already manually checked. Double payments are in itself wrong, but usually an error. Large amounts of employee expenses can be a factor, but should be checked according to the allowed expenses of an employee. Invoices paid through an offshore company and anonymous deposits are also seen as effective but many responses debate whether this is even possible in the systems.

From the comments, it is described that type of entry and currency can be useful when this is different from the usual ones. Furthermore, the amount can be interesting when it is just under the "light boundary" which is the boundary of amount that needs to be manually checked.

6.1.1.3 Date and Time

Date and time indicators are overall seen as less effective for detecting errors and fraud. The most useful indicators are time and date of processing invoices and whether this is during office hours. The others are only useful in extreme cases. One useful remark is that in international organisations, the auditors work in all time zones and thus becomes time useless.

6.1.1.4 Business Partner

In this category, the results vary from medium effective to very effective. The most effective indicators include non-existing business partner, risky origin country and different activities than registered. Apart from the country is it hard to check the other two features. You would need additional data sources, especially if the invoices come from abroad. More suitable popular indicators are many journal entries and corrections from the same business partner, no physical address and the history of blocking of a business partner.

The comments show that there are a few very useful indicators in this category, but that the feasibility is also harder to achieve. It would require additional data sources. Furthermore,

it is mentioned that some indicators are already prevented by the use of business rules. E.g. missing bank details would automatically prevent a transaction from happening.

6.1.1.5 Fraud triangle

Most respondents indicated that most fraud can be assigned to the opportunity side of the triangle, as shown in Figure 2. This means that the focus should be on the prevention of opportunities to commit fraud.

There was also the possibility to mention some other indicators that were missed in the survey. The following indicators were mentioned and can also be applicable to data:

- Qualitative aspect of invoices such as the description
- "Light procedure", meaning invoices that have an amount low enough to be automatically reviewed.
- The user that makes the entries
- Change of bank details

The light procedure is something that can be used in the data. The rest of the answers focuses on business rules and checks that can be implemented in the systems but are not fit as features. This is one of the challenges to this problem: the auditor has a different mindset and does not fully understand the machine learning approach. The answers from the survey must be "translated" to features fit for the algorithm. The top 10 performing indicators are translated into features and are included in the process.

6.2 Outlier detection Algorithm

6.2.1 Comparison of three algorithms

The three algorithms are applied on the test set provided by the DCGAS. The comparison is made between the ranking of the outliers according to their outlier score and if they are confirmed to be an outlier. Figure 21 shows the results. The box plots are used to see how high the outlier score was given to the test set. The Y axis is the index of all transactions in the test set, ordered by their outlier score. The transactions with the lowest ID have the highest outlier score. The box plots show the distributions of the confirmed outlying transactions. The closer the confirmed outliers are to zero, the better performing the algorithm is. The results show that Isolation forest performs the best on the test set. This algorithm is thus used for the manual labelling sessions.

6.2.2 Manual labelling

Table 12 shows the results from the manual labelling sessions with the three financial auditors. The goal of the manual labelling is to validate the predictions of the IF. The predictions of

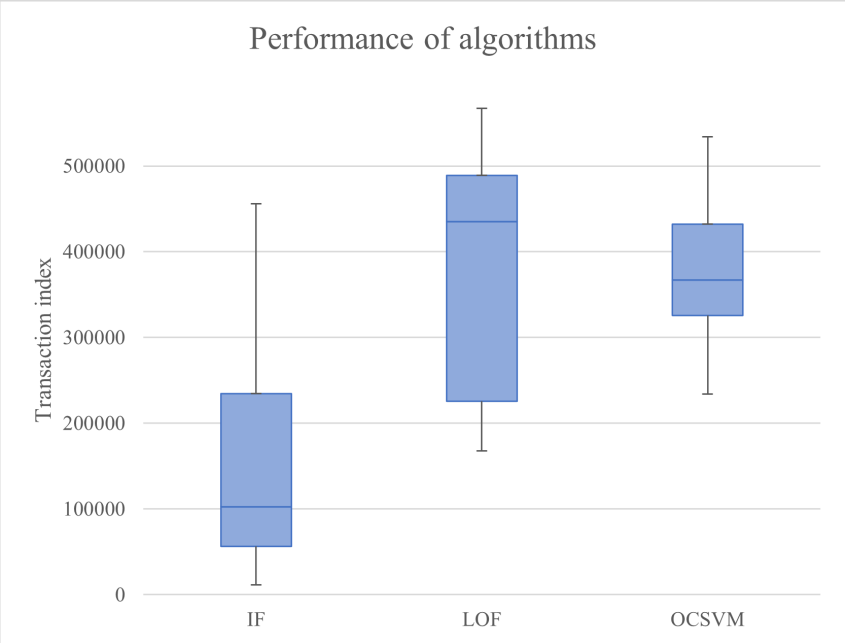


Figure 21: The distribution of the outliers per algorithm

the algorithm were presented to the auditors. They were asked to indicate if the prediction of the algorithm was correct. All auditors were presented with a mix of outliers. Table 12 present the number of correct and incorrect predictions by the algorithm. The percentage represents the number of correct predictions made by the algorithm in the sample reviewed by the auditors. E.g. auditor 1 has labelled 73 invoices, of which 28 were predicted correctly by the algorithm and 45 were incorrectly labelled according to his expertise. This results in a percentage of 38.4% of correctly identified outliers by the Isolation Forest.

There was a selection of overlapping invoices. The results from the manual labelling can be found in Table 13 to compare the behaviour of the three auditors.

	Auditor		
	1	2	3
Labelled	73	13	13
Correct	28	11	12
Incorrect	45	2	1
Percentage	38.4	84.6	92.3

Table 12: The results of the labelling per auditor

	Auditor			
Transaction	1	2	3	Agreement
192	1	0		No
185	1	1		Yes
184	1	1		Yes
604	1	1		Yes
191	0	1		No
532	0	1		No
983	1	1		Yes
904	1	1		Yes
203	1	1		Yes
899	1	1		Yes
879	0	0		Yes
654	1	1	1	Yes
631	0	1		No
4224	0		1	No
2641	1		1	Yes
3908	0		0	Yes
3974	1		1	Yes
631	1		1	Yes
750	0		1	No
2640	1		1	Yes
2650	1		1	Yes
3869	0		1	No
190	0		1	No
3892	0		1	No
1458	0		1	No
Total reviewed as outlier	14	11	12	15

Table 13: Overlapping invoices between the auditors and their labels

6.3 Explanation facility for auditors

6.3.1 Iteration one: treatment design

6.3.1.1 Interview setup

The first round of interviews was designed to receive information about the auditor and the explanation facility. The interview is divided into four parts. First, it is some general information on what the participant does and how they work with accountants. Second, some information is shared on their general take on working with accountants and what important is for them. Third, some questions are asked about the goals and content of an explanation facility for auditors. Finally, some existing examples of explanation mechanisms are shown and their opinions on these was asked. The results will be discussed below per part. Furthermore, from the descriptions given by the participants, different requirements are deducted and written down.

Four persons were selected with varying roles within the Dutch Governmental Audit Services. Two participants are working in the data analytics team that create data analyses for the auditors. One participant is the manager of this team and also educated as an auditor. One participant acts as a financial auditor within a department and manages a team on certain projects. This mix of participants was chosen as it is important to hear from both perspectives. The more technically inclined participants can comment on what accountants need and what would contribute to their work. The participants that are more business inclined can comment on the feasibility for accountants and what they want and understand. The technical participants have less years of experience than the business participants.

6.3.1.2 Interview results

All participants describe working with auditors as a thorough process. The auditors know what they are doing and their work is very reliable and thorough. There is a great diversity between the auditors, some are looking for confirmation that everything is going according to the protocols, others are looking for the errors and mistakes. An important factor to consider when working with auditors is that they act in accordance to the law and standards considering accounting. Their work needs to be documented in an audit trail and everything needs to be able to be tracked. A certain risk needs to be covered and for this certain standards and guidelines need to be followed. It is important to ask the auditor for the specific guidelines, standards and laws that are applicable before you start to work with them on a project. When working together with auditors, the data analysts are usually involved from the beginning. Once a certain project is started, they look at the question and see if a data analysis can support the process. If so, the analysts look at what is possible and what is needed. When more details are gathered from the auditors, they can make an analysis. They present the analysis to the auditors once done, and help them to understand it and guide them on how

to use the analyses. Some teams are multidisciplinary and the entire process can happen within the team. The acceptance of the technology by auditors differs greatly. Some teams are very enthusiastic and some are more conservative. Overall, it is seen as an extra tool, not a fundamental process. The technology should always add something and should not be there for the sake of it.

An explanation facility is crucial for an accountant because they need to be able to see the entire process and know what certain outcomes are based on. They need to be able to base their statement on some rules or certainties. Currently, there are rules concerning the random sampling of invoices. However, there are no rules yet for the algorithm. The auditor does not need to be able to reproduce the process but they need to be able to trust it, guarantee that all activities have been performed and sell it to others. Next, they need to produce an audit trail of what has been done. Explanations can provide the reasoning for this trail. Finally if the algorithm does not explain why an invoice is an outlier, it would take considerate time to discover the reasons. When talking about necessary components in an explanation facility for auditors, several things are mentioned. It should include information about the selection of the algorithm, why it is allowed and they need to be able to judge if they can get certainty from this algorithm. It should include criteria on which the auditor can decide which invoices to further review. E.g. if you include the amount, the auditor can decide to review the outlying invoices with the highest amount. The most described goals for an explanation facility for auditors are: transparency, trust, effectiveness and satisfaction. Others are also mentioned but not by all participants. The first ideas for an explanation facility by the participants include the explanation of the input, process and output. It should show which features are important and could be connected to their "vaktechnische bijsluiter". The facility should include a timestamp and should be able to be exported as to provide reliable documentation.

The different categories according to Table 5 and the according mechanisms are presented to the participants through a case. Transaction 1007 is shown as an outlier and different types of explanations are provided. The following feedback was provided per category.

	Sector Person	IT 1	IT 2	Business 3	Business 4
Process-based	Model	Works good when concrete. Only possible with decision tree.	Clear, will explain a lot.	Supportive, case-based works for auditors.	Gives overview to review process.
	Explanation model	Auditors will not read this.	Always possible, needs to be there.	Text can support the explanation.	Less interested in as auditor.
Outcome-based | By example	Proximity	This does not provide why it is an outlier. Interesting to use for identifying clusters.	This can create infinite loops, where to stop?	Start with the entire sample of invoices too see the outlying total.	Only interesting if you can see the entire set of invoices. Now you cannot see what makes it an outlier.
Outcome-based | Visual	Graphs	Useful, when plotted against the outliers.	Very useful, a lot of information in one overview. Show outlier information in graphs.	Again, start with entire sample but interesting overview.	Always useful.
	Partial dependence plot	This can show relationships but should be supported by textual explanations.	Useful, but too difficult for auditors.	Very useful but needs more explanation.	Need more explanation, too difficult.
	Interactive visualisation	Good idea, start with highest amount of money.	Use this for all mechanisms, neighbours etc.	More useful than static graphs.	Seems very useful.
Outcome-based | Local	LIME	Very useful to keep the explanation simple.	Support & contradict are too difficult. Group it into two columns and list the features on importance.	This can give indication for further research. You can add the values of the selected invoice.	Per transaction it too detailed, will not use this. Also needs more explanation before it can be used.
Outcome-based | Natural language	IF THEN rules	Can be used but natural language is better.	Very informative.	Helps to understand the algorithm.	Basic, but easy to understand.
	Natural language expressions	Simple language and links to more information would be useful.	This is better, documents automatically the process.	Understandable text, the interpretation is nice to have.	Gives a better overview.
Outcome-based | Simplification	Algorithm statistics	You need to document this but does not give any explanation to auditor.	Useful for IT but auditors will be confused.	Very good to have but needs explanation and thresholds.	No clue what the statistics mean, explanation and thresholds can be useful.
Outcome-based | Feature relevance	Inclusion	Ordered by importance is clear and useful.	Add scores behind each feature and make two columns with support and contradict.	Too simple.	Interesting, can be useful.
	Scores	Too complex for the auditors.	Useful, but make it easier.	Prefer scores to enable the accountant to judge themselves.	The scores are fruitful to know where to research further.
	SHAP	Visual representation is easier to understand.	The best visual, however too much information and will confuse auditors too much.	Very interesting, also analyses the input.	Too complicated to understand as auditor.

Table 14: Feedback on the shown examples

6.3.1.3 Goals

All participants have confirmed that the following goals are important for the explanation facility, as stated in previous requirements:

- to provide transparency around the process.
- to increase the users confidence.
- to help the user make better decisions.
- to be user-friendly.

6.3.1.4 Requirements

To achieve these goals, several requirements are made that the explanation facility needs to adhere to. These requirements are based on the input from the auditors. Per goal, these are mentioned below.

Transparency

- The facility should show information on the input.
- The facility should show information about the model and throughput.
- The facility should show information about the output.
- The facility should be able to convey decisions made by the data scientist.

Trust

- The facility should disclose information to the end user about the performance of the algorithm and should enable the user to judge if it can be trusted.

Effectiveness

- The facility should be able to disclose the information that the end user needs to make a decision.
- The facility should be able to direct the user to relevant information.
- The facility should help the user making the decision.

Satisfaction

- The facility should be visually attractive.
- The facility needs to be intuitive.

6.3.2 Iteration one: treatment validation

6.3.2.1 Tasks

Table 15 shows the times in seconds that each participant used per task. The average and standard deviation per task is also calculated. The task time cannot be bench marked to any other situation, as this is a new situation. However, the standard deviation can be compared to see individual differences per task. It can be assumed that when the standard deviation is lower, the task execution and thus the design is clear for all participants.

	Participant							
Task	1	2	3	4	5	6	Average	Standard Deviation
1: Explanation on input data	69	152	34	95	127	8	140.2	132.0
2: Explore indicators of input data	520	55	118	150	82	81	115.9	140.3
3: Explore general explanation model	54	90	16	18	56	15	72.2	44.7
4: Explore specific model explanation	91	63	170	45	119	80	81.7	48.1
5: Explore general output	70	57	22	28	152	9	70.4	39.8
6: Pick individual invoice	46	51	131	63	84	32	93.3	38.2
7: Explore individual output	157	113	47	115	126	54	102	39.2
Total	1007	581	538	514	746	279	610.83	183.7

Table 15: Time in seconds per task per participant

6.3.2.2 Interfaces

Table 16 shows the results of the A-B testing with different interfaces. The screenshots presented in Chapter 5 represent the extended version B of the interfaces. Table 16 describes the preferences of the five participants, the mode which is the most frequently chosen interface and the comments that were given during the A-B testing. The different interfaces are described in Table 8. Version A of the interfaces is designed for simplicity while version B provides more options.

	Participant							
Interface	1	2	3	4	5	6	Mode	Comments
Home	B	A	B	B	B	A	B	Including the slider gives the opportunity to focus on the financial impact.
Input	A	B	B	B	A	B	B	Visualisations can be more appealing to accountants.
Model	B	B	A	A	B	B	B	The graph is easier to interpret than the list of metrics.
Output	B	B	B	B	A	A	B	Even though the numbers can be confusing, it can show the amount of influence.

Table 16: The results of the A-B testing

6.3.2.3 Interview

Table 17 shows the answers that were given to the different questions asked at the end of the usability test. The questions were aimed at four subjects of the applications, namely general, goals, content, and design. The answers are summarised into these four categories. The questions can be found in Appendix 9.8.

Subject	Feedback
General	Good first impression. It seems that this application can provide indications of outlying transactions.
Goals | Transparency	The transparency has increased due to the explanations on all levels. The explanations of the model can be extended as they are limited and hard to comprehend.
Goals | Trust	The trust has increased for most. However, as some indicators are without context, this was not true for all.
Goals | Satisfaction	All agreed that the application was easy to use. With more time, people could get further used to the application.
Content	There was sufficient content to determine whether a transactions had to be researched. Some did want to see more connection to the financial statement or the financial impact of a transaction.
Design	All indicated that they would use this application again.

Table 17: The results of the interview part of the usability test

6.3.3 Iteration two: treatment design

6.3.3.1 Prototype design

The result from this phase is the final design of the explanation facility. Several components have been improved according to previous results and the screenshots of the final design can be found in Section 5.

6.3.4 Iteration two: treatment validation

6.3.4.1 Real case testing

The results of this phase include the outcomes of the manual labelling, as described in Section 6.2. However, during the manual labelling sessions, feedback and ideas were given by the auditors that will be summarised below. These are specifically targeted on the prototype and not on the outcomes of the algorithm.

The three auditors all agreed that the explanation facility was very useful for reviewing outliers and finding the causes. The output interface was mainly used to perform the manual labelling. The focus was on using the SHAP values per invoice to see which indicators contributed to the prediction. The ability to view all data from the individual transactions proved to be fruitful as this was used to confirm the predictions. The visualisations were used less often, only to see how the individual transaction related to the mass.

All three auditors suggested the same improvement for the explanation facility. As some outliers had the same characteristics, it would be preferred if these could be grouped together. The auditor does not need to review all the same outliers then, but could only review a few from this subgroup. This would make the explanation facility more efficient.

Finally, the auditors indicated how this explanation facility could contribute to the review of the financial statement. It was suggested that the facility could be used to select a sample of invoices that will be reviewed. This sample represents the entire mass of invoices that are included in the financial statement. The facility and algorithm could create a sample of outlying transactions that need to be reviewed and are more interesting than randomly selected invoices. Auditors one and three also indicated that the facility could be used after the financial statement review to detect any other irregularities. This would not only be incorrect invoices but also on the processes of the organisation. Auditor one recognised a pattern in the invoices that many were paid too late. Once discovered, this can be communicated to the

financial management of the organisation.

7 Discussion

7.1 Interpretation of results

This section discusses the meaning of the results and how these are used to answer the research questions. The sections are divided into three, each discussing and answering one research question. Section 7.1.1 reflects on the outcomes of the outlier detection algorithms and the features of fraud. Section 7.1.2 discusses the results of the design of the explanation facility. Section 7.1.3 evaluates the contribution of the results to the review of the financial statement.

7.1.1 Outlier detection

7.1.1.1 Fraud features

The results from the survey have indicated which indicators can predict errors and potential fraud most correctly, as shown in Figure 20. When looking at the three categories, invoice details, date and time, and business partner, it is clear that indicators on the business partner are ranked highest in their ability to predict fraud. This is thus focused on external fraud. Especially indicators such as non-existing business partner, other activities than registered and risky countries of origin scored high in the survey. These indicators in itself are abnormal in comparison to e.g. type of entry. That could be a potential reason for their high score.

The invoice details are also fruitful to consider as useful indicators. However, it is visible that some of the highest ranking indicators in this category are also abnormal in itself such as double payments and anonymous deposits. Large and complicated transactions at the end of the year and large amounts of employee expenses are not in itself abnormal. There is some balance between these two types of indicators in this category.

Date and time as a category is considered useless according to the survey responses. Some have indicated that there are usually normal reasons for invoices not paid in time. Furthermore, participants have indicated that they work for foreign customers and that time is relative to their country of origin. In the current situation, many auditors work from home and are therefore more likely to work in the weekends or on odd hours. These reasons validate the outcome that the focus should not be on on any datetime features.

Sixteen of the thirty-two features have a score above the average of 3 and can thus be deemed as useful.

7.1.1.2 Comparison of algorithms

The comparison between the IF, LOF and OCSVM is made according to a set of outlying transactions as identified by the DCGAS. The outlier detection algorithms give each invoice an outlier score. This was used to see how the confirmed outliers were distributed according to the algorithms. A box plot was created to show this distribution, as seen in Figure 21. The results show that the Isolation Forest seems to provide the best results. The mean, minimum and maximum values are closest to zero and thus closest to the highest outlying score.

The test set of outlying transactions has been selected based on a certain business rules, determined by the DCGAS. The exact rules are not known but it was indicated that indicators were selected that matter to the auditors and extreme values are included. This should be considered when interpreting the results of the algorithms. Isolation Forests create many decision trees and use the average path lengths to calculate the outlier score [130]. Decision trees could potentially recognise these business rules easier than the other two as decision trees are used to structure the presentation of a series of closely related business rules. This could thus explain their improved performance.

Previous research has already indicated that Isolation Forests perform well for fraud detection, as stated in section 3. These results confirm this finding. The outcomes are positive as the Isolation Forest is a simple algorithm to understand and it is quite fast in comparison. These characteristics make it a good fit for the explanation facility.

7.1.1.3 Outcomes of Isolation Forest

The Isolation Forest had the best results in the previous phase and was therefore used in the manual labelling sessions. Three auditors have indicated for a sample of outliers if the algorithm was correct and the invoices required further investigation. The results show that the results varied per auditor. The first auditor has labelled the most invoices and indicated that only 38.4% of the invoices was truly an outlier. Auditor two and three have both labelled thirteen invoices and have respectively shown that 85% and 92% of the invoices was truly an outlier. The average of all three auditors is 71.8%.

Because the results differed significantly between the first and the other auditors, their behaviour was compared on overlapping invoices. Table 13 shows that there is an overlap of 60% in their behaviour. However, the first auditor has a much lower percentage of positive indications of outliers. This could be a reason of the differing results of the manual labelling sessions per auditor. Another cause could be that the second and third auditors have labelled a smaller portion of the invoices. As they started at the most outlying invoices, it is logical that their rate of positive indications is much higher as the invoices are more likely to be outlying.

The results can be interpreted as an indication that the Isolation Forest can be used to detect outlying invoices. The auditors indicated that the predictions do not point toward fraud, but do indicate invoices that are worth to further research. This will ultimately lead to better work if the Isolation Forest can help detect potential errors. The Isolation Forest is a good start for using algorithms to detect outliers in invoices. Once there are more labelled examples available, the algorithm can be refined and improved upon.

The results confirm in line with other literature that the Isolation Forest is a good start for detecting outliers in the invoices of the public sector. The results have indicated that a certain amount of the detected outliers is indeed worth to review.

7.1.2 Explanation facility

7.1.2.1 Iteration one

The results of the first iteration include the duration for each task that it took the participants, as shown in Table 15. All tasks were completed by all participants, making the completion ratio 100%. As there is not a base scenario to which the times can be bench marked, the standard deviation was calculated per task. Tasks three to six have a relative lower standard deviation compared to task one, two and seven. This indicates that the differences in time per participant for task 3, 4, 5, and 6 are smaller. This could mean that the task can easily be completed by all participants and the design allows for an easy completion.

The interfaces indicate that overall the auditors prefer to have version B as interface, as shown in Table 8. Version B is the more complex interface that offers various options. The participants all belong to a subgroup, but the preferences among auditors vary. Therefore, by having more options, a larger target group can be provided for. These results confirm the findings from the requirement interviews that auditors prefer to have options in explanations, both visually and textual.

The interview questions were included to receive feedback on the goals for the explanation facility. All auditors agreed that the explanation facility provided a good basis for presenting explanations, as presented in Table 17. There are a few comments on small alterations that need to be made. Considering the goals, the auditors indicated that the transparency was increased by the explanation facility. The division into input, model and output turned out to be fruitful for the transparency of the process. Overall, the trust in the algorithm was also increased. However, as the facility provided explanations on which features were relevant, the trust was also decreased. Some auditors noted that some features and their importance did not make sense. All auditors agreed that the explanation facility was easy to use because of its clean and simple interface. The auditors indicated that most relevant information was present in the facility to make a correct decision. Effectiveness was thus also achieved. These results mean that the basis of the explanation facility has achieved the

goals that this research focused on. Some small alterations are needed to further improve the facility.

7.1.2.2 Iteration two

The results from the second iteration include the manual labelling sessions and the feedback that was given in these rounds. The fact that the facility was able to facilitate the manual labelling session indicates that it is ready to be used by auditors for real cases. The facility was able to provide all the necessary information for the auditors to determine the correctness of the outliers.

The auditors did indicate that an additional feature would be preferred where you could detect patterns in the outliers. This is logical as auditors are looking for outlying invoices and patterns in this. This would make the process even more efficient as they would only have to check a few outliers from the same pattern.

7.1.3 Contribution to financial statement review

The results in the last phase give an indication of the contribution to the financial statement review. Within all phases of this research, auditors were involved. Through interviews and the usability testing, a sense of relevance of this tool was developed. Three main contribution are identified for the algorithms and explanation facility.

7.1.3.1 Finding certain risks

At the beginning of a year, the auditors make a planning of which objects will be reviewed and which risks will be identified. They set certain indicators on which they focus during the review in the coming year. The participating auditors have mentioned that they can use the explanations of the relevant indicators on the entire model in this process. This would mean that they can use the outlying indicators from the past year as risks for the next year.

7.1.3.2 During the review of invoices

Participating auditors have indicated that they could use the algorithm to make a sample of invoices that are particularly outlying on certain indicators. This could replace the random sample selection technique and increase the quality of the review.

7.1.3.3 After the review of the financial statement

The algorithm and its outcomes can also be used after the financial statement review to find errors and correct these. As one auditor mentioned that late payments came up a lot and it would mean malpractices at the financial management. This implies that some indicators can also review the quality of the processes and reflect on these.

7.2 Contribution

The contribution of this research is divided into two subsections: the theoretical contribution and the practical contribution.

7.2.1 Theoretical contribution

This research has contributed to the theoretical body of research in different manners. First, the research to state of the art outlier detection methods has provided an overview of the most recent studies on unsupervised outlier detection algorithms. Ample research has been performed on using unsupervised outlier detection algorithms to detect errors and fraud in financial transactions. However, research in this area innovates quickly, and overviews become outdated quickly.
Second, this research provides an overview of explanation mechanisms that have been developed for outlier detection methods and to what extent these have been applied in the financial world. To our knowledge, there was no overview of explanation mechanisms described in the financial domain. Very little research can be found on any explanation mechanisms applied in the financial domain. Therefore, an overview of explanation mechanisms in all domains was created.
Third, the previous overviews were combined in an assessment matrix that reviews each combination of outlier detection algorithm and explanation mechanism on their feasibility. This assessment led to the discovery of gaps and opportunities in this field of research. The assessment of these combinations was not made before and therefore directly contributes to the theoretical body of research. It provides researchers with gaps in the existing research and suggests opportunities.
Fourth, this research provides a list of indicators that could predict errors and potential fraud in invoices, according to auditors. The survey on fraud indicators has resulted in a list of indicators that can be considered as red flags in invoices. Several sources have identified cases and red flags in auditing practices. However, these were not connected to any data indicators. The list from this research are indicators that can be identified in data sets, thus making the indicators suitable for any machine learning approach.
Fifth, this research compares three algorithms and their performance with real life data provided by the DCGAS. Isolation Forest is confirmed to outperform the other algorithms. This provides more evidence on the performance of the different outlier detection algorithms.
Finally, this research contributes by designing an explanation facility for auditors. In this process, it was discovered what requirements financial auditors have, what explanations they prefer and what information is needed to achieve the goals. To our knowledge, there has not been any research performed to explanation mechanisms for financial auditors.

7.2.2 Practical contribution

This research has also provided some practical contributions. The research was performed for the DCGAS and the practical contributions will mainly benefit their organisation. First, the application of the Isolation Forest can be used as a basis for outlier detection within their invoices. The results showed promise and alterations can easily be implemented.
Second, the design of the explanation facility contributes as it can be used to further test the algorithm and its outcomes. The facility is designed to be generic and any type of data set and algorithm can be connected to it. The DCGAS can use the facility to test any algorithm on any data set and provide explanations to the auditors.
Finally, context was provided on how to use the explanation facility and outlier detection algorithm in their daily practice. The DCGAS can use these recommendations to apply the research in certain parts of their processes and research its potential.

7.3 Limitations

There are certain limitations to this research that need to be discussed. These have influenced the results and validity of this research. However, by discussing these limitations future research can take these into account.

First, the feature selection and feature engineering stage can be improved upon. Due to time constraints, no text mining techniques were applied to textual features. As auditors rely on the descriptions of invoices, this could be a rich source of information. Furthermore, the data set can be combined with other data sources to get more information. An example would be to use information from the chamber of commerce (KvK) to validate business partners. This limitation contributed to the implementation of the outlier detection algorithms being very standard. The scope of this research was mainly focused on the design and implementation of the explanation facility and was therefore not able to go deeper into the implementation of the outlier detection algorithms.
Second, the comparison of the three outlier detection algorithms was based on the results of the outlier test set provided by the DCGAS. These outliers were selected by certain business rules and based on the values of certain indicators. This leads to a certain kind of outlier being included in the test set. Having a mixed set of outliers would improve the quality and validity of the comparison of algorithms.
Third, the outcomes of the Isolation Forest were manually labelled by three financial auditors. However, all three auditors have different backgrounds and experiences with reviewing invoices. Their estimate of the invoices being outlying is to a certain extent subjective. Their differences in labelling became evident in the results. Overlapping transactions can provide insight in the differences in their labelling behaviour. However, the resources of this research allowed for only a small sample of overlapping invoices.
Fourth, the explanation facility has been tested in two iterations with respectively six and

three participants. These participants only represent a very small part of the entire group of financial auditors. While research indicates that most problems are found in usability testing with five participants, the participant number should be increased to give the results more validity.
Finally, this research was executed with the assumption that an outlying invoice means that there is an error or potential fraud. However, the financial auditors indicated that the outliers that were detected, were not automatically errors. The information could be correct and just different from the mass. Therefore, it is advised to frame this research as an outlier detection case instead of potential fraud detection.

7.4 Recommendations for future work

This section will describe the recommendations that are made for future research. There are three main recommendations.

First, it is recommended to improve the feature selection and engineering phase. As described in the limitation section, this phase was done in a very general manner. If the resources allow it, multiple feature sets can be made and their performance could be compared to each other. It is recommended to apply text mining techniques to certain features to gain more information. The option to use data from an external source can also be explored.

Second, the financial auditors had indicated that the explanation facility needs to be extended to include an option to work with subsets of data. These could be subsets of the same types of outliers and would enable auditors to only review a few of the subset instead of all individual invoices. This would improve the efficiency of the processes and convince the auditors even more to use the explanation facility.

Third, it is recommended to test the explanation facility and algorithms on a larger scale. The number of participants in the testing was limited in this research (n=9) due to the resources. The number of participants was sufficient for the design of the prototype. However, before implementing the explanation facility, testing at a larger scale needs to be performed. It is also recommended to test the explanation facility in the different stages of the auditing process. As described, there are three stages where this explanation facility and algorithms can be used. By testing them in these three stages, it can be discovered where they will be of most relevance.

Finally, it would be interesting to investigate if the results of this research are only applicable to the public sector or if it is also useful within the private sector. The authors of [14] have identified systemic differences between these two sectors. They both face similar managerial-level IT issues and challenges but the outcome suggests that there is not a one size fits all approach. This research is based on data of the public sector and the participants that took part in this research were all employed for the DCGAS. This means that this research is

completely based upon the public sector and some further research could compare the results to the private sector to see its usefulness.

8 Conclusion

This section provides a conclusion to this research by providing answers to the research questions.

RQ1 To what extent can unsupervised outlier detection algorithms help to identify potential financial fraud in invoices of the public sector?
The results of the manual labelling show that there is some potential for identifying outlying invoices. The approach to start with the most outlying transactions was found to be fruitful. Overall, 72% of the labelled invoices were found to be classified correctly.
However, it should be noted that the auditors indicated whether these invoices would be of interest to further research. The auditors found that many outliers were probably not an error or potential fraud. This result points out that the outlier detection algorithms are not necessarily useful to detect potential financial fraud, but to detect suspicious, outlying invoices.
The algorithms were perceived very useful to the auditors to detect anomalies in the invoices and to help them select invoices to review. This is a promising basis and the algorithms can be refined and researched more to enable them to detect potential financial fraud.

RQ1.1 What features are important for potential financial fraud detection (identified by domain experts)?
The results have indicated that 16 indicators of the 32 proposed are seen as useful for identifying potential fraud by auditors, as seen in Table 1. The results have shown that most highly ranked features belong to the category of business partner. This would imply that the focus of the participants was on external fraud or that it is easiest to recognise this. The following features are listed as important for both error detection and potential fraud detection: "History blocks business partner", "No physical address", "Activity after inactivity", "Many journal entries and corrections", "Risky country of origin", "Other activities than registered", "Non-existing business partner", "Date of entry invoice", "Large and complicated transactions at the end of the year", "Large number of purchases from new business partner", "Anonymous deposits", "Large number of expensive transactions of employees", "Large amounts of travel and phone costs", "Payment through offshore company", and "Double payments".

RQ1.2 What unsupervised outlier detection algorithms delivers the most promising results on the invoice data of the public sector?
From the comparison of the three algorithms on the test set, it was found that the Isolation Forest delivered the best quantitative results. It labelled the outliers with the highest outlier scores as compared to the Local Outlier Factor and the One Class Support Vector Machine. Furthermore, the Isolation Forest is very fast and easy to interpret compared to the other algorithms. It provides a good basis for outlier detection and can be further refined according

to the recommendations to increase its performance.

RQ2 How can an explanation facility, aimed to explain the algorithm and its outcome to a financial auditor, be structured?
Some of the key design decisions that were validated are the following. First, the use of a web application proved to be fruitful as it offers an interactive platform that is easily accessible for the auditors. The auditors have indicated that would need explanations about the entire process to trust the outcomes. Therefore, it was decided to provide explanations on the input, model and output and to keep this structure evident in the web application. Third, the auditors indicated that preferences on the format of explanations differ in the target group. This led to the inclusion of different options of explanations such as visualisations but also tables and text.

RQ2.1 What is the purpose of the explanation facility for the financial auditor?
It was found that the explanation facility had four goals for the financial auditors. These include effectiveness, trust, transparency and satisfaction. The explanation facility prototype did provide for these four goals and improved on all of them.

It was identified that the explanation facility was needed to provide an auditor with all information to review the decisions made by an algorithm.

RQ2.2 What explanation mechanisms should be included in this facility?
The results indicated that the auditors prefer to have information on the model, interactive visualisations, algorithm statistics, local and global feature importance and textual information. All these explanation mechanisms are included in the first prototype of the explanation facility. During usability testing and the real case testing phase, it became apparent that auditors also prefer to have an explanation mechanism that indicates the proximity of outliers to others. This could be used to discover patterns and make the reviewing process more efficient.

RQ3 How can the models and explanations contribute to the assessment of the reliability of the financial statement?
The results indicated that there are three possible options to use this facility. First, it can be used at the beginning of the year to identify the outlying indicators of the previous year and decide if these are risks specifically watched in the coming years. Second, it can be used as a sample selection technique for the invoice sample that represents all the invoices for the financial statement. Normally, this sample is taken randomly but with the use of the algorithm, outlying transactions can automatically be selected. Finally, the algorithm and facility can be used after the financial statement has been approved. For the approval, not all invoices are reviewed. However, the algorithm can still be used to analyse all the invoices and check if there are any errors. These could be corrected after identification, meaning that

the quality of the bookkeeping will increase. Further research will indicate for which of these three options the algorithm and explanation facility can offer the most benefits.

References

[1] Cos240. *NBA* (2020).

[2] ACHITUVE, I., KRAUS, S., AND GOLDBERGER, J. Interpretable online banking fraud detection based on hierarchical attention mechanism. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)* (2019), IEEE, pp. 1–6.

[3] ALDAIRI, M., KARIMI, L., AND JOSHI, J. A trust aware unsupervised learning approach for insider threat detection. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)* (2019), IEEE, pp. 89–98.

[4] ALMELO, L. V. Rode vlaggen; frauderisico's ontdekken en melden. *Accountant (NBA)* (2020).

[5] AMARASINGHE, K., KENNEY, K., AND MANIC, M. Toward explainable deep neural network based anomaly detection. In *2018 11th International Conference on Human System Interaction (HSI)* (2018), IEEE, pp. 311–317.

[6] AMARASINGHE, K., AND MANIC, M. Improving user trust on deep neural networks based intrusion detection systems. In *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society* (2018), IEEE, pp. 3262–3268.

[7] AMARASINGHE, T., APONSO, A., AND KRISHNARAJAH, N. Critical analysis of machine learning based approaches for fraud detection in financial transactions. In *Proceedings of the 2018 International Conference on Machine Learning Technologies - ICMLT '18* (2018), ACM Press, pp. 12–17.

[8] ARNALDO, I., VEERAMACHANENI, K., AND LAM, M. ex2: a framework for interactive anomaly detection. In *IUI Workshops* (2019).

[9] AZEVEDO, A. I. R. L., AND SANTOS, M. F. Kdd, semma and crisp-dm: a parallel overview. *IADS-DM* (2008).

[10] BAEK, H., OH, J., KIM, C. Y., AND LEE, K. A model for detecting cryptocurrency transactions with discernible purpose. In *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)* (2019), IEEE, pp. 713–717.

[11] BOONE, H. N., AND BOONE, D. A. Analyzing likert data. *Journal of extension 50*, 2 (2012), 1–5.

[12] BROWN, A., TUOR, A., HUTCHINSON, B., AND NICHOLS, N. Recurrent neural network attention mechanisms for interpretable system log anomaly detection. In *Proceedings of the First Workshop on Machine Learning for Computing Systems - MLCS'18* (2018), ACM Press, pp. 1–8.

[13] CABANES, G., BENNANI, Y., AND GROZAVU, N. Unsupervised learning for analyzing the dynamic behavior of online banking fraud. In *2013 IEEE 13th International Conference on Data Mining Workshops* (2013), IEEE, pp. 513–520.

[14] CAMPBELL, J., MCDONALD, C., AND SETHIBE, T. Public and private sector it governance: Identifying contextual differences. *Australasian Journal of Information Systems 16*, 2 (Mar. 2010).

[15] CANILLAS, R., HASAN, O., SARRAT, L., AND BRUNIE, L. Supplier impersonation fraud detection using bayesian inference. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)* (2020), IEEE, pp. 330–337.

[16] CARCILLO, F., LE BORGNE, Y.-A., CAELEN, O., KESSACI, Y., OBLÉ, F., AND BONTEMPI, G. Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences* (2019).

[17] CARDOSO, B., SEDRAKYAN, G., GUTIÉRREZ, F., PARRA, D., BRUSILOVSKY, P., AND VERBERT, K. Intersectionexplorer, a multi-perspective approach for exploring recommendations. *International Journal of Human-Computer Studies 121* (2019), 73 – 92. Advances in Computer-Human Interaction for Recommender Systems.

[18] CARLETTI, M., MASIERO, C., BEGHI, A., AND SUSTO, G. A. Explainable machine learning in industry 4.0: Evaluating feature importance in anomaly detection to enable root cause analysis. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (2019), IEEE, pp. 21–26.

[19] CHARTERS, E. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education: A Journal of Educational Research and Practice 12*, 2 (2003).

[20] CHATTERJEE, J., AND DETHLEFS, N. Deep learning with knowledge transfer for explainable anomaly prediction in wind turbines. *Wind Energy 23*, 8 (2020), 1693–1710.

[21] CHEN, V., YOON, M.-K., AND SHAO, Z. Novelty detection via network saliency in visual-based deep learning. In *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)* (2019), IEEE, pp. 52–57.

[22] CHOI, D., LEE, K., AND YOU, I. An artificial intelligence approach to financial fraud detection under iot environment: A survey and implementation. *Sec. and Commun. Netw. 2018* (Jan. 2018).

[23] CHUNG, Y., KRASKA, T., POLYZOTIS, N., TAE, K. H., AND WHANG, S. E. Slice finder: Automated data slicing for model validation. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)* (2019), IEEE, pp. 1550–1553.

[24] CORNER, A., PIDGEON, N., AND PARKHILL, K. Perceptions of geoengineering: public attitudes, stakeholder perspectives, and the challenge of 'upstream' engagement. *Wiley Interdisciplinary Reviews: Climate Change 3*, 5 (June 2012), 451–466.

[25] DAMA, U. The six primary dimensions for data quality assessment. *DAMA UK, October* (2013).

[26] DANESHPAZHOUH, A., AND SAMI, A. Semi-supervised outlier detection with only positive and unlabeled data based on fuzzy clustering. In *The 5th Conference on Information and Knowledge Technology* (2013), IEEE, pp. 344–348.

[27] DE ROUX, D., PEREZ, B., MORENO, A., VILLAMIL, M. D. P., AND FIGUEROA, C. Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2018), ACM, pp. 215–222.

[28] DEFILIPPI, R. R. Standardize or normalize? - examples in python, Apr 2018.

[29] DETHISE, A., CANINI, M., AND KANDULA, S. Cracking open the black box: What observations can tell us about reinforcement learning agents. In *Proceedings of the 2019 Workshop on Network Meets AI & ML - NetAI'19* (2019), ACM Press, pp. 29–36.

[30] DINAPOLI, T. P., AND HANCOX, S. J. Red flags for fraud. *Office of the State Comptroller*.

[31] DOMINGUES, R., FILIPPONE, M., MICHIARDI, P., AND ZOUAOUI, J. A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition 74* (2018), 406 – 421.

[32] DROBICS, M., BODENHOFER, U., AND WINIWARTER, W. Mining clusters and corresponding interpretable descriptions – a three–stage approach. *Expert Systems 19*, 4 (2002), 224–234.

[33] DURTSCHI, C., HILLISON, W., AND PACINI, C. The effective use of benford's law to assist in detecting fraud in accounting data. *Journal of forensic accounting 5*, 1 (2004), 17–34.

[34] EICHMANN, P., SOLLEZA, F., TAN, J., TATBUL, N., AND ZDONIK, S. Metro-viz: Black-box analysis of time series anomaly detectors. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), ACM, pp. 1–6.

[35] ESHGHI, A., AND KARGARI, M. Introducing a method for combining supervised and semi-supervised methods in fraud detection. In *2019 15th Iran International Industrial Engineering Conference (IIIEC)* (2019), pp. 23–30.

[36] ESHGHI, A., AND KARGARI, M. Introducing a new method for the fusion of fraud evidence in banking transactions with regards to uncertainty. *Expert Systems with Applications 121* (2019), 382 – 392.

[37] FAN, S., LIU, G., AND CHEN, Z. Anomaly detection methods for bankruptcy prediction. In *2017 4th International Conference on Systems and Informatics (ICSAI)* (2017), IEEE, pp. 1456–1460.

[38] FEMI, P. S., AND GANESH VAIDYANATHAN, S. Comparative study of outlier detection approaches. In *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)* (2018), IEEE, pp. 366–371.

[39] GAO, Z. Application of cluster-based local outlier factor algorithm in anti-money laundering. In *2009 International Conference on Management and Service Science* (2009), IEEE, pp. 1–4.

[40] GARCHERY, M., AND GRANITZER, M. Identifying and clustering users for unsupervised intrusion detection in corporate audit sessions. In *2019 IEEE International Conference on Cognitive Computing (ICCC)* (2019), IEEE, pp. 19–27.

[41] GEDIKLI, F., JANNACH, D., AND GE, M. How should i explain? a comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies 72*, 4 (2014), 367 – 382.

[42] GIURGIU, I., AND SCHUMANN, A. Additive explanations for anomalies detected from multivariate temporal data. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (2019), ACM, pp. 2245–2248.

[43] GOLBIN, I., LIM, K. K., AND GALLA, D. Curating explanations of machine learning models for business stakeholders. In *2019 Second International Conference on Artificial Intelligence for Industries (AI4I)* (2019), IEEE, pp. 44–49.

[44] GOLDSTEIN, M., AND UCHIDA, S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLOS ONE 11*, 4 (04 2016), 1–31.

[45] GRETARSSON, B., ODONOVAN, J., BOSTANDJIEV, S., HALL, C., AND HÖLLERER, T. SmallWorlds: Visualizing social recommendations. *Computer Graphics Forum 29*, 3 (Aug. 2010), 833–842.

[46] GUPTA, N., ESWARAN, D., SHAH, N., AKOGLU, L., AND FALOUTSOS, C. Beyond outlier detection: Lookout for pictorial explanation. In *ECML/PKDD* (2018).

[47] HAJEK, P., AND HENRIQUES, R. Mining corporate annual reports for intelligent detection of financial statement fraud – a comparative study of machine learning methods. *Knowledge-Based Systems 128* (2017), 139 – 152.

[48] HOFFMAN, R. R., MUELLER, S. T., KLEIN, G., AND LITMAN, J. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[49] ICO. Explaining decisions made with ai, May 2020.

[50] JERAGH, M., AND ALSULAIMI, M. Combining auto encoders and one class support vectors machine for fraudulant credit card transactions detection. In *2018 Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)* (2018), IEEE, pp. 178–184.

[51] JIANYUN XU, SUNG, A., AND QINGZHONG LIU. Tree based behavior monitoring for adaptive fraud detection. In *18th International Conference on Pattern Recognition (ICPR'06)* (2006), IEEE, pp. 1208–1211.

[52] JUNG, A., AND NARDELLI, P. H. J. An information-theoretic approach to personalized explainable machine learning. *IEEE Signal Processing Letters 27* (2020), 825–829.

[53] JUSZCZAK, P., ADAMS, N. M., HAND, D. J., WHITROW, C., AND WESTON, D. J. Off-the-peg and bespoke classifiers for fraud detection. *Computational Statistics Data Analysis 52*, 9 (2008), 4521 – 4532.

[54] KASSEM, R., AND HEGAZY, M. Financial reporting fraud: Do red flags really help? *Journal of Economics and Engineering* (June 2010).

[55] KAUFFMANN, J., MÜLLER, K.-R., AND MONTAVON, G. Towards explaining anomalies: A deep taylor decomposition of one-class models. *Pattern Recognition 101* (2020), 107198.

[56] KHAZANE, A., RIDER, J., SERPE, M., GOGOGLOU, A., HINES, K., BRUSS, C. B., AND SERPE, R. DeepTrax: Embedding graphs of financial transactions. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)* (2019), IEEE, pp. 126–133.

[57] KIM, J., KIM, H.-J., AND KIM, H. Fraud detection for job placement using hierarchical clusters-based deep neural networks. *Applied Intelligence* (02 2019).

[58] KITAMURA, S., AND NONAKA, Y. *Explainable Anomaly Detection via Feature-Based Localization*. Springer International Publishing, Cham, 2019, pp. 408–419.

[59] KITCHENHAM, B., AND CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering.

[60] KULTUR, Y., AND CAGLAYAN, M. U. A novel cardholder behavior model for detecting credit card fraud. In *2015 9th International Conference on Application of Information and Communication Technologies (AICT)* (2015), IEEE, pp. 148–152.

[61] LANE, J.-E. *The public sector: concepts, models and approaches*. Sage, 2000.

[62] LEI, J. Z., AND GHORBANI, A. A. Improved competitive learning neural networks for network intrusion and fraud detection. *Neurocomputing 75*, 1 (2012), 135 – 145. Brazilian Symposium on Neural Networks (SBRN 2010) International Conference on Hybrid Artificial Intelligence Systems (HAIS 2010).

[63] LI, Z., LIU, G., WANG, S., XUAN, S., AND JIANG, C. Credit card fraud detection via kernel-based supervised hashing. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (2018), IEEE, pp. 1249–1254.

[64] LIPTON, Z. C. The mythos of model interpretability. *Queue 16*, 3 (June 2018), 31–57.

[65] LIU, F. T., TING, K. M., AND ZHOU, Z.-H. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining* (2008), IEEE, pp. 413–422.

[66] LIU, J., AND WANG, G. Outlier detection based on local minima density. In *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference* (2016), IEEE, pp. 718–723.

[67] LIU, Z., AND LU, A. Explainable visualization for interactive exploration of CNN on wikipedia vandal detection. In *2019 IEEE International Conference on Big Data (Big Data)* (2019), IEEE, pp. 2354–2363.

[68] LOU, Y.-I., AND WANG, M.-L. Fraud risk factor of the fraud triangle assessing the likelihood of fraudulent financial reporting. *Journal of Business & Economics Research (JBER) 7*, 2 (Feb. 2011).

[69] LUNDBERG, S. M., AND LEE, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[70] MA, T., QIAN, S., CAO, J., XUE, G., YU, J., ZHU, Y., AND LI, M. An unsupervised incremental virtual learning method for financial fraud detection. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)* (2019), IEEE, pp. 1–6.

[71] MARINO, D. L., WICKRAMASINGHE, C. S., RIEGER, C., AND MANIC, M. Data-driven stochastic anomaly detection on smart-grid communications using mixture poisson distributions. In *IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society* (2019), IEEE, pp. 5855–5861.

[72] MATSUBARA, T. Bayesian deep learning: A model-based interpretable approach. *Nonlinear Theory and Its Applications, IEICE 11* (01 2020), 16–35.

[73] MCGOVERN, A., LAGERQUIST, R., JOHN GAGNE, D., JERGENSEN, G. E., ELMORE, K. L., HOMEYER, C. R., AND SMITH, T. Making the black box more transparent: Understanding the physical implications of machine learning. 2175–2199.

[74] MEJIA-LAVALLE, M. Outlier detection with innovative explanation facility over a very large financial database. In *2010 IEEE Electronics, Robotics and Automotive Mechanics Conference* (2010), IEEE, pp. 23–27.

[75] MITTAL, S., AND TYAGI, S. Performance evaluation of machine learning algorithms for credit card fraud detection. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (2019), IEEE, pp. 320–324.

[76] MONAMO, P., MARIVATE, V., AND TWALA, B. Unsupervised learning for robust bitcoin fraud detection. In *2016 Information Security for South Africa (ISSA)* (2016), IEEE, pp. 129–134.

[77] MORO, S., LAUREANO, R., AND CORTEZ, P. Using data mining for bank direct marketing: An application of the crisp-dm methodology.

[78] MORRIS, B. Explainable anomaly and intrusion detection intelligence for platform information technology using dimensionality reduction and ensemble learning. In *2019 IEEE AUTOTESTCON* (2019), IEEE, pp. 1–5.

[79] MOYES, G., YOUNG, R., AND DIN, H. Malaysian internal and external auditor perceptions of the effectiveness of red flags for detecting fraud. *Int. J. of Auditing Technology 1* (01 2013), 91 – 106.

[80] MUNIR, M., SIDDIQUI, S., KÜSTERS, F., MERCIER, D., DENGEL, A., AND AHMED, S. Tsxplain: Demystification of dnn decisions for time-series using natural language and statistical features. pp. 426–439.

[81] NGUYEN, Q. P., LIM, K. W., DIVAKARAN, D. M., LOW, K. H., AND CHAN, M. C. GEE: A gradient-based explainable variational autoencoder for network anomaly detection. In *2019 IEEE Conference on Communications and Network Security (CNS)* (2019), IEEE, pp. 91–99.

[82] NU.NL. Fraude door medewerker rijkswaterstaat blijkt groter dan gedacht, Jul 2019.

[83] OMIDI, M., MIN, Q., MORADINAFTCHALI, V., AND PIRI, M. The efficacy of predictive methods in financial statement fraud. *Discrete Dynamics in Nature and Society 2019* (05 2019), 1–12.

[84] OUNACER, S., AIT EL BOUR, H., OUBRAHIM, Y., GHOUMARI, M., AND AZZOUAZI, M. Using isolation forest in anomaly detection: the case of credit card transactions. *Periodicals of Engineering and Natural Sciences (PEN) 6* (11 2018), 394.

[85] PARRENO-CENTENO, M., ALI, M., GUAN, Y., AND VAN MOORSEL, A. *Unsupervised Machine Learning for Card Payment Fraud Detection*. 02 2020, pp. 247–262.

[86] PARRENO-CENTENO, M., ALI, M. A., GUAN, Y., AND MOORSEL, A. V. Unsupervised machine learning for card payment fraud detection. In *Risks and Security of Internet and Systems*, S. Kallel, F. Cuppens, N. Cuppens-Boulahia, and A. Hadj Kacem, Eds., vol. 12026. Springer International Publishing, 2020, pp. 247–262. Series Title: Lecture Notes in Computer Science.

[87] PASINI, A., AND BARALIS, E. Detecting anomalies in image classification by means of semantic relationships. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)* (2019), IEEE, pp. 231–238.

[88] PATIL, A., WADEKAR, A., GUPTA, T., VIJAN, R., AND KAZI, F. Explainable LSTM model for anomaly detection in HDFS log file using layerwise relevance propagation. In *2019 IEEE Bombay Section Signature Conference (IBSSC)* (2019-07), IEEE, pp. 1–6.

[89] PAULA, E. L., LADEIRA, M., CARVALHO, R. N., AND MARZAGAO, T. Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (2016), IEEE, pp. 954–960.

[90] 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., AND DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[91] PHUA, C., LEE, V. C.-S., SMITH-MILES, K., AND GAYLER, R. W. A comprehensive survey of data mining-based fraud detection research. *ArXiv abs/1009.6119* (2007).

[92] PICKERING, C., AND BYRNE, J. The benefits of publishing systematic quantitative literature reviews for phd candidates and other early-career researchers. *Higher Education Research Development 33* (12 2013), 534–548.

[93] PUMSIRIRAT, A., AND YAN, L. Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. *International Journal of Advanced Computer Science and Applications 9* (01 2018).

[94] PYLE, D. An executive's guide to machine learning. *McKinsey Company* (Jun 2015).

[95] QUAH, J. T., AND SRIGANESH, M. Real-time credit card fraud detection using computational intelligence. *Expert Systems with Applications 35*, 4 (2008), 1721–1732.

[96] RAHUL, K., SETH, N., AND DINESH KUMAR, U. Spotting earnings manipulation: Using machine learning for financial fraud detection. In *Artificial Intelligence XXXV*, M. Bramer and M. Petridis, Eds., vol. 11311. Springer International Publishing, 2018, pp. 343–356. Series Title: Lecture Notes in Computer Science.

[97] RAJENDRAN, S., MEERT, W., LENDERS, V., AND POLLIN, S. Unsupervised wireless spectrum anomaly detection with interpretable features. *IEEE Transactions on Cognitive Communications and Networking 5*, 3 (2019), 637–647.

[98] RAVISANKAR, P., RAVI, V., RAGHAVA RAO, G., AND BOSE, I. Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems 50*, 2 (2011), 491–500.

[99] RAZA, M., AND QAYYUM, U. Classical and deep learning classifiers for anomaly detection. In *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)* (2019), IEEE, pp. 614–618.

[100] RECHTBANK DEN HAAG. Syri wetgeving in strijd met europees verdrag voor de rechten van de mens, Feb 2020.

[101] REZAPOUR, M. Anomaly detection using unsupervised methods: Credit card fraud case study. *International Journal of Advanced Computer Science and Applications 10*, 11 (2019).

[102] RIBEIRO, M. T., SINGH, S., AND GUESTRIN, C. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), ACM, pp. 1135–1144.

[103] RIECK, K., AND LASKOV, P. Visualization and explanation of payload-based anomaly detection. In *2009 European Conference on Computer Network Defense* (2009), IEEE, pp. 29–36.

[104] ROBOTICS, AND AI. Ethics guidelines for trustworthy ai. *Shaping Europe's digital future - European Commission* (Nov 2019).

[105] RUFF, L., ZEMLYANSKIY, Y., VANDERMEULEN, R., SCHNAKE, T., AND KLOFT, M. Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In *Proceedings of the 57th Annual Meeting of the Association*

for Computational Linguistics (2019), Association for Computational Linguistics, pp. 4061–4071.

[106] RUSHIN, G., STANCIL, C., SUN, M., ADAMS, S., AND BELING, P. Horse race analysis in credit card fraud—deep learning, logistic regression, and gradient boosted tree. In *2017 Systems and Information Engineering Design Symposium (SIEDS)* (2017), IEEE, pp. 117–121.

[107] SEIDMAN, I. *Interviewing as qualitative research: A guide for researchers in education and the social sciences*. Teachers college press, 2006.

[108] SMITH-RENNER, A., RUA, R., AND COLONY, M. Towards an explainable threat detection tool. In *IUI Workshops* (2019).

[109] SOLORIO-FERNÁNDEZ, S., CARRASCO-OCHOA, J. A., AND MARTÍNEZ-TRINIDAD, J. F. A review of unsupervised feature selection methods. *Artificial Intelligence Review 53*, 2 (Jan. 2019), 907–948.

[110] SONG, F., DIAO, Y., READ, J., STIEGLER, A., AND BIFET, A. EXAD: A system for explainable anomaly detection on big data traces. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)* (2018), IEEE, pp. 1435–1440.

[111] SUN, W., CHEN, M., YE, J.-X., ZHANG, Y., XU, C.-Z., ZHANG, Y., WANG, Y., WU, W., ZHANG, P., AND QU, F. Semi-supervised anti-fraud models for cash pre-loan in internet consumer finance. In *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)* (2019), IEEE, pp. 635–640.

[112] SUSTO, G. A., TERZI, M., MASIERO, C., PAMPURI, S., AND SCHIRRU, A. A fraud detection decision support system via human on-line behavior characterization and machine learning. In *2018 First International Conference on Artificial Intelligence for Industries (AI4I)* (2018), IEEE, pp. 9–14.

[113] TAKEISHI, N. Shapley values of reconstruction errors of PCA for explaining anomaly detection. In *2019 International Conference on Data Mining Workshops (ICDMW)* (2019), IEEE, pp. 793–798.

[114] TINTAREV, N., AND MASTHOFF, J. A survey of explanations in recommender systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop* (2007), pp. 801–810.

[115] TINTAREV, N., AND MASTHOFF, J. *Designing and Evaluating Explanations for Recommender Systems*. Springer US, Boston, MA, 2011, pp. 479–510.

[116] TONEKABONI, S., JOSHI, S., MCCRADDEN, M. D., AND GOLDENBERG, A. What clinicians want: contextualizing explainable machine learning for clinical end use. *arXiv preprint arXiv:1905.05134* (2019).

[117] VAN DEN BERG, M., AND KUIPER, O. Xai in the financial sector.

[118] VANHOEYVELD, J., MARTENS, D., AND PEETERS, B. Value-added tax fraud detection with scalable anomaly detection techniques. *Applied Soft Computing 86* (2020), 105895.

[119] VOJÍŘ, S., ZEMAN, V., KUCHAŘ, J., AND KLIEGR, T. EasyMiner.eu: Web framework for interpretable machine learning based on rules and frequent itemsets. *Knowledge-Based Systems 150* (2018), 111–115.

[120] WALIGORA, R. Alleged fraud for 2019 has reached over £1 billion. *KPMG* (Jan 2020).

[121] WANG, D., YANG, Q., ABDUL, A., AND LIM, B. Y. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), CHI '19, Association for Computing Machinery, p. 1–15.

[122] WEBSTER, J., AND WATSON, R. T. Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly* (2002), xiii–xxiii.

[123] WESTON, D. J., HAND, D. J., ADAMS, N. M., WHITROW, C., AND JUSZCZAK, P. Plastic card fraud detection using peer group analysis. *Advances in Data Analysis and Classification 2*, 1 (2008), 45–62.

[124] WHITE, G. I., SONDHI, A. C., AND FRIED, D. *The analysis and use of financial statements*. John Wiley & Sons, 2002.

[125] WIERINGA, R. J. *Design science methodology for information systems and software engineering*. Springer, 2014.

[126] WIRTH, R., AND HIPP, J. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (2000), Springer-Verlag London, UK, pp. 29–39.

[127] ZAMINI, M., AND MONTAZER, G. Credit card fraud detection using autoencoder based clustering. In *2018 9th International Symposium on Telecommunications (IST)* (2018), IEEE, pp. 486–491.

[128] ZHANG, X., MARWAH, M., LEE, I.-T., ARLITT, M., AND GOLDWASSER, D. ACE – an anomaly contribution explainer for cyber-security applications. In *2019 IEEE International Conference on Big Data (Big Data)* (2019), IEEE, pp. 1991–2000.

[129] ZHANG, Y., YOU, F., AND LIU, H. Behavior-based credit card fraud detecting model. In *2009 Fifth International Joint Conference on INC, IMS and IDC* (2009), IEEE, pp. 855–858.

[130] ZHAO, X., WU, Y., LEE, D. L., AND CUI, W. iForest: Interpreting random forests via visual analytics. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2019), 407–416.

[131] ZHENG, F., AND LIU, Q. Anomalous telecom customer behavior detection and clustering analysis based on ISP's operating data. *IEEE Access 8* (2020), 42734–42748.

[132] ZHUANG, Y., SMALL, D. L., SHU, X., YU, K., ISLAM, S., AND DING, W. Galaxy: Towards scalable and interpretable explanation on high-dimensional and spatio-temporal correlated climate data. In *2018 IEEE International Conference on Big Knowledge (ICBK)* (2018), IEEE, pp. 146–153.

9 Appendix

9.1 Description of the data

Column name	Format	Description
Mandant	Nominal	Internal ID
Bedrijfs_nr	Nominal	Number of company
Bedrijfs_nr_oms	Nominal	Description of number of company
Zakenpartner_nr	Nominal	Number of business partner
Zakenpartner_naam	Nominal	Name of business partner
Zakenpartner_land	Nominal	Country of business partner
Zakenpartner_plaats	Nominal	City of business partner
Zakenpartner_adres	Nominal	Address of business partner
Zakenpartner_sorteerveld	Nominal	Short name of business partner
Zakenpartner_postcode	Nominal	Zipcode of business partner
Zakenpartner_postbus	Nominal	POB business partner
Zakenpartner_postbus_postcode	Nominal	POB zip code business partner
Zakenpartner_postbus_plaats	Nominal	POB city business partner
Zakenpartner_groep	Nominal	Billing group
Zakenpartner_groep_oms	Nominal	Billing group
Meeboekrekening	Nominal	Reconcilliation account of business partner
Meeboekrekening_oms	Nominal	Description of reconcilliation account
Zakenpartner_btw_nr	Nominal	VAT number
Datum_aanmaak	Ordinal	Date of creation
Jaar_maand_aanmaak	Ordinal	Year and month of creation
Gebr_naam_aanmaak	Nominal	User that created business partner
Betaalmethode	Nominal	Method of payment
Betaaltermijn	Discrete	Payment term
Betaaltermijn_oms	Nominal	Description payment term
Is_eenmalige_zakenpartner	Binary	Indicator if it is a one time business partner
Heeft_check_dubbele_facturen	Binary	Indicator if the partner has a check for double invoices
Zakenpartner_bedrijfs_nr	Nominal	Company number business partner
Is_zakenpartner_verwijderd	Binary	Indicator if the business partner is removed
Is_zakenpartner_geblokkeerd	Binary	Indicator if the business partner is blocked
Is_zakenpartner_verwijderd_ of_geblokkeerd	Binary	Indicator if the business partner is blocked or removed
Is_zakenpartner_actief	Binary	Indicator if the business partner is active
Gegevens_incompleet	Binary	Indicator if the details are incomplete
Bankgegevens_incompleet	Binary	Indicator if the bank details are incomplete
Valuta_compleet	Binary	Indicator if the currency details are complete
Incoterms_compleet	Binary	Indicator if the international commercial terms are complete
Heef_landspecifiek_btw_nr	Binary	Indicator if business partner has a country specific VAT code
DEP	Nominal	Department
LogSessionNbr	Nominal	Number of the logging session

Table 18: Description of database SMD

Column name	Format	Description
Mandant	Nominal	Internal ID
Bedrijfs_nr	Nominal	Number of company
Bedrijfs_nr_oms	Nominal	Description of company number
Boekjaar	Ordinal	Financial year
Jaar_maand	Ordinal	Combination of financial year and month
Doc_nr	Ordinal	Document number
Doc_pos	Ordinal	Document position
Zakenpartner_nr	Nominal	Number of business partner
Zakenpartner_naam	Nominal	Name of business partner
Zakenpartner_land	Nominal	Country of business partner
Doc_oms	Nominal	Document description
Doc_soort	Nominal	Type of document
Doc_soort_oms	Nominal	Description of document type
Boekingssleutel	Nominal	Entry key
Boekingssleutel_oms	Nominal	Description of key
Debet_credit	Binary	If entry is debit or credit
Valuta	Nominal	Currency
Bedrag_VV	Continuous	Amount in foreign currency
Bedrag_EV	Continuous	Amount in own currency
Functioneel_gebied	Nominal	Functional area
Gebr_naam	Nominal	Username
Gebr_type	Nominal	User type
Gebr_type_oms	Nominal	Description of user type
Transactiecode	Nominal	Code of transaction
Doc_nr_vereffening	Ordinal	Liquidation of document number
Status	Binary	Status of the entry
Business Transaction	Nominal	Business transaction
Referentie	Nominal	Reference
Doc_status	Nominal	Status of entry
Grootboekrekening	Nominal	Ledger account
Profit_center	Nominal	Profit center
Operatie_ref_HD	Nominal	Reference to table
Aanvullende_betaalwijze	Nominal	Extra method of payment
MM_gerelateerd	Binary	MM related
Factuur_nr_MM	Nominal	Invoice number MM
Boekjaar_MM	Ordinal	Financial year MM
Betaaltermijn	Nominal	Payment term
Datum_doc	Ordinal	Date of document

Column name	Format	Description
Datum_invoer	Ordinal	Date of creation
Tijd_invoer	Continuous	Time of creation
Datum_boeking	Ordinal	Date of entry
Datum_vereffening	Ordinal	Date of liquidation
Datum_vrijgave_FI	Ordinal	Date of release in FI
Datum_vrijgave_MM	Ordinal	Date of release in MM
Datum_ingangstermijn	Ordinal	Entry period date
Dagen_verstreken_1	Discrete	Days expired 1
Dagen_verstreken_2	Discrete	Days expired 2
Dagen_verstreken_3	Discrete	Days expired 3
Kortingspercentage_1	Continuous	Discount percentage 1
Kortingspercentage_2	Continuous	Discount percentage 2
Korting	Continuous	Discount
Betaalmethode	Nominal	Payment method
Aantal_dagen	Discrete	Number of days between creation
and liquidation		
Aantal_dagen_groep	Ordinal	Group of number of days
Datum_verval	Ordinal	Date of expiration
Aantal_dagen_laat	Discrete	Number of days late
Dagen_doc_naar_boeking	Discrete	Invoice date vs. creation date
Dagen_boeking_naar_betaling	Discrete	Creation date vs. payment date
Verschil_datum_boeking_naar_betaling	Discrete	
Gebr_naam_aanmaak	Nominal	Username that created entry
Doc_nr_storno	Nominal	Document number
Boekjaar_storno	Discrete	Financial year cancellation
Boeking_negatief	Binary	Indicator if it is a negative entry
Betaaltermijn_SMD	Discrete	Payment term from business partner
Betaalmethode_SMD	Nominal	Payment method from business partner
Is_eenmalige_zakenpartner	Binary	Indicator if business partner is one time
Heeft_check_dubbele_facturen	Binary	Indicator if business partner has
check on double invoices		
Zakenpartner_groep	Nominal	Group of business partner
Zakenpartner_groep_oms	Nominal	Description of business partner group
Betaaltermijn_MM	Nominal	Payment term of MM
Storno_FI	Nominal	Cancellation of FI
Storno_MM	Nominal	Cancellation of MM
Ind_GO	Binary	Indicator if goods are received
Bedrag_EV_GO	Continuous	Amount of received goods

Column name	Format	Description
Ind_IO	Binary	Indicator if it is a purchasing order
Geblokkeerd	Binary	Indicator if it is blocked
Historie_blokkade_in_MM	Binary	History of blocking in MM
Historie_geblokkeerd_FI	Binary	History of blocking in FI
Historie_geblokkeerd	Binary	History of blocking
Blokkeringstype	Nominal	Type of blocking
Historie_geparkeerd_FI	Binary	History of parking in FI
Historie_registratie_in_MM	Binary	History of registration in MM
Betaaltermijn_berekend	Continuous	Calculated payment term
Groep_boekingsgang	Nominal	Group of book entry process
Groep_AP	Nominal	Group of the type of booking
Groep_facturen	Nominal	Group of invoices
Factuurtype	Nominal	Type of invoice
Type_crediteur_boeking	Nominal	Entry of creditor type
Datum_scan	Ordinal	Date of scan invoice
DEP	Nominal	Department
LogSessionNbr	Nominal	Log session

Table 19: Description of database APA

9.2 Analysis of the data

Column name	Total values	Empty percentage	Unique count
Mandant	398560.0	0.0	1.0
Bedrijfs_nr	398258.0	0.1	20.0
Bedrijfs_nr_oms	398258.0	0.1	20.0
Zakenpartner_nr	398560.0	0.0	246663.0
Zakenpartner_naam	398557.0	0.0	206807.0
Zakenpartner_land	398560.0	0.0	211.0
Zakenpartner_plaats	398543.0	0.0	19299.0
Zakenpartner_adres	387957.0	2.7	197568.0
Zakenpartner_sorteerveld	159692.0	59.9	62626.0
Zakenpartner_postcode	389208.0	2.3	112183.0
Zakenpartner_postbus	87663.0	78.0	9550.0
Zakenpartner_postbus_postcode	86950.0	78.2	8393.0
Zakenpartner_postbus_plaats	6294.0	98.4	823.0
Zakenpartner_groep	398560.0	0.0	12.0
Zakenpartner_groep_oms	398560.0	0.0	12.0
Meeboekrekening	398256.0	0.1	6.0
Meeboekrekening_oms	398256.0	0.1	6.0
Zakenpartner_btw_nr	89337.0	77.6	35340.0
Datum_aanmaak	398560.0	0.0	3866.0
Jaar_maand_aanmaak	398560.0	0.0	184.0
Gebr_naam_aanmaak	398560.0	0.0	182.0
Betaalmethode	394755.0	1.0	33.0
Betaaltermijn	398218.0	0.1	5.0
Betaaltermijn_oms	299525.0	24.8	3.0
Is_eenmalige_zakenpartner	5.0	100.0	1.0
Heeft_check_dubbele_facturen	398258.0	0.1	1.0
Zakenpartner_bedrijfs_nr	0.0	100.0	0.0
Is_zakenpartner_verwijderd	398560.0	0.0	2.0
Is_zakenpartner_geblokkeerd	398560.0	0.0	2.0
Is_zakenpartner_verwijderd_of_geblokkeerd	398560.0	0.0	2.0
Is_zakenpartner_actief	398560.0	0.0	3.0
Gegevens_incompleet	398560.0	0.0	2.0
Bankgegevens_incompleet	398560.0	0.0	2.0
Valuta_compleet	398560.0	0.0	2.0
Incoterms_compleet	398560.0	0.0	1.0
Heef_landspecifiek_btw_nr	398560.0	0.0	2.0
DEP	398560.0	0.0	8.0
LogSessionNbr	398560.0	0.0	2.0

Table 20: Analysis of the database SMD

Column name	Total values	Empty percentage	Unique count
Mandant	1239309.0	0.0	1.0
Bedrijfs_nr	1239309.0	0.0	20.0
Bedrijfs_nr_oms	1239309.0	0.0	20.0
Boekjaar	1239309.0	0.0	12.0
Jaar_maand	1239309.0	0.0	84.0
Doc_nr	1239309.0	0.0	1173364.0
Doc_pos	1239309.0	0.0	40.0
Zakenpartner_nr	1239309.0	0.0	72546.0
Zakenpartner_naam	1239309.0	0.0	68609.0
Zakenpartner_land	1239309.0	0.0	195.0
Doc_oms	711875.0	42.6	492989.0
Doc_soort	1239309.0	0.0	39.0
Doc_soort_oms	1239309.0	0.0	42.0
Boekingssleutel	1239309.0	0.0	16.0
Boekingssleutel_oms	1239309.0	0.0	15.0
Debet_credit	1239309.0	0.0	2.0
Valuta	1239309.0	0.0	104.0
Bedrag_VV	1239309.0	0.0	428760.0
Bedrag_EV	1239309.0	0.0	405898.0
Functioneel_gebied	0.0	100.0	0.0
Gebr_naam	1238664.0	0.1	274.0
Gebr_type	1238664.0	0.1	2.0
Gebr_type_oms	1238664.0	0.1	2.0
Transactiecode	1206789.0	2.6	24.0
Doc_nr_vereffening	1232607.0	0.5	491289.0
Status	1239309.0	0.0	2.0
Business Transaction	1238640.0	0.1	3.0
Referentie	716383.0	42.2	495735.0
Doc_status	50.0	100.0	1.0
Grootboekrekening	1239309.0	0.0	7.0
Profit_center	307.0	100.0	1.0
Operatie_ref_HD	1238664.0	0.1	4.0
Aanvullende_betaalwijze	0.0	100.0	0.0
MM_gerelateerd	1239309.0	0.0	2.0
Factuur_nr_MM	1239309.0	0.0	267113.0
Boekjaar_MM	1239309.0	0.0	4.0
Betaaltermijn	736141.0	40.6	6.0
Datum_doc	1239309.0	0.0	1459.0

Datum_invoer	1239309.0	0.0	757.0
Tijd_invoer	1238664.0	0.1	69073.0
Datum_boeking	1239309.0	0.0	760.0
Datum_vereffening	1239309.0	0.0	387.0
Datum_vrijgave_FI	75331.0	93.9	435.0
Datum_vrijgave_MM	86123.0	93.1	466.0
Datum_ingangstermijn	1239309.0	0.0	845.0
Dagen_verstreken_1	1239309.0	0.0	5.0
Dagen_verstreken_2	1239309.0	0.0	1.0
Dagen_verstreken_3	1239309.0	0.0	1.0
Kortingspercentage_1	1239309.0	0.0	1.0
Kortingspercentage_2	1239309.0	0.0	1.0
Korting	1239309.0	0.0	1.0
Betaalmethode	878452.0	29.1	21.0
Aantal_dagen	1239309.0	0.0	843.0
Aantal_dagen_groep	1239309.0	0.0	4.0
Datum_verval	1239309.0	0.0	862.0
Aantal_dagen_laat	1232607.0	0.5	527.0
Dagen_doc_naar_boeking	1239309.0	0.0	1314.0
Dagen_boeking_naar_betaling	1232607.0	0.5	457.0
Verschil_datum_boeking_naar_betaling	1239309.0	0.0	564.0
Gebr_naam_aanmaak	123932.0	90.0	171.0
Doc_nr_storno	31991.0	97.4	31862.0
Boekjaar_storno	1238664.0	0.1	5.0
Boeking_negatief	27943.0	97.7	1.0
Betaaltermijn_SMD	1239309.0	0.0	4.0
Betaalmethode_SMD	1170321.0	5.6	23.0
Is_eenmalige_zakenpartner	32.0	100.0	1.0
Heeft_check_dubbele_facturen	1239309.0	0.0	1.0
Zakenpartner_groep	1239309.0	0.0	10.0
Zakenpartner_groep_oms	1239309.0	0.0	10.0
Betaaltermijn_MM	0.0	100.0	0.0
Storno_FI	1239309.0	0.0	2.0
Storno_MM	1239309.0	0.0	2.0
Ind_GO	267112.0	78.4	2.0
Bedrag_EV_GO	267084.0	78.4	36625.0
Ind_IO	267112.0	78.4	2.0
Geblokkeerd	1239309.0	0.0	2.0
Historie_blokkade_in_MM	267112.0	78.4	2.0

Historie_geblokkeerd_FI	1239309.0	0.0	2.0
Historie_geblokkeerd	1239309.0	0.0	2.0
Blokkeringstype	1239309.0	0.0	3.0
Historie_geparkeerd_FI	1239309.0	0.0	1.0
Historie_registratie_in_MM	267112.0	78.4	2.0
Betaaltermijn_berekend	1239309.0	0.0	7.0
Groep_boekingsgang	1239309.0	0.0	5.0
Groep_AP	1239309.0	0.0	5.0
Groep_facturen	1239309.0	0.0	2.0
Factuurtype	1239309.0	0.0	8.0
Type_crediteur_boeking	1239309.0	0.0	5.0
Datum_scan	438076.0	64.7	519.0
DEP	1239309.0	0.0	7.0
LogSessionNbr	1239309.0	0.0	2.0

Table 21: Data analysis of the database APA

9.3 Feature engineering

The following Python code is used to create the different features described in section 4.2.3.2

```
#%% - Feature engineering
#-----------------------------------------------
#Split data values
def split_date(dataframe):
    for col in dataframe.columns:
        if dataframe[col].dtypes == 'datetime64[ns]':
            dataframe[col + '_jaar'] = pd.to_datetime(
                dataframe[col]).dt.to_period('Y').astype(np.
                int64)
            dataframe[col + '_maand'] = pd.to_datetime(
                dataframe[col]).dt.to_period('M').astype(np.
                int64)
            dataframe[col + '_dag'] = pd.to_datetime(
                dataframe[col]).dt.to_period('D').astype(np.
                int64)
            dataframe[col + '_weekdag'] = pd.to_datetime(
                dataframe[col]).dt.weekday.astype(np.int64)
            dataframe[col + '_quarter'] = pd.to_datetime(
                dataframe[col]).dt.quarter.astype(np.int64)
            dataframe = dataframe.drop(col, axis=1)
    return dataframe



def split_time(dataframe):
    for col in dataframe.columns:
        if col == 'Datum_invoer':
            dataframe[col + '_hour'] = dataframe[col].dt.
                hour.astype(np.int64)
            dataframe[col + '_minute'] = dataframe[col].dt.
                minute.astype(np.int64)
            dataframe[col + '_second'] = dataframe[col].dt.
                second.astype(np.int64)
            dataframe[col + '_businesshour'] = 0
            for i in range(len(dataframe["Datum_invoer"])):
```

```
                if ((dataframe[col + '_hour'].iloc[i] > 8)
                    and ((dataframe[col + '_hour'].iloc[i] <
                    18))):
                     dataframe[col + '_businesshour'].iloc[i]
                         = 1
            dataframe = dataframe.drop(col, axis=1)
    return dataframe

#Benford's law - make distribution
def count_dist(dataframe):
    dist = pd.DataFrame({'numbers': [1,2,3,4,5,6,7,8,9], '
       blaw': [30.1,17.6,12.5,9.7,7.9,6.7,5.8,5.1,4.6]})
    for col in dataframe.columns:
        if ("Bedrag" or "bedrag") in col:
            dist['count' + col] = 0
            dist['percent' + col] = 0
            for i in range(len(dataframe[col])):
                first = str(dataframe[col].iloc[i])[0]
                for j in range(len(dist)):
                    if first == str(dist['numbers'].iloc[j])
                       :
                        dist['count' + col].iloc[j] = dist['
                           count' + col].iloc[j] + 1
                        dist['percent' + col].iloc[j] = ((
                           dist['count' + col].iloc[j] + 1)/
                           len(dataframe[col])*100)
    return dist
#Benfordslaw - make feature
def blaw(dataframe, dist):
    for col in dataframe.columns:
        if ("Bedrag" or "bedrag") in col:
            dataframe['blaw'+ col] = 0
            for z in range(len(dataframe)):
                first = str(dataframe[col].iloc[z])[0]
                for x in range(len(dist)):
                    if first == str(dist['numbers'].iloc[x])
                       :
                        if ((dist['percent' + col].iloc[x]))
                            > ((dist['blaw'].iloc[x])+1):
```

```
                                dataframe["blaw" + col].iloc[z]
                                    = 1
    return dataframe


#Length of business partner
def length_bp(dataframe):
    if 'Datum_aanmaak' in dataframe.columns:
        dataframe['Lengte_SMD'] = (pd.to_datetime('today') -
            df['Datum_aanmaak']).astype(np.int64)
    return dataframe


#Number of transactions per BP
def count_trans(dataframe):
    freq = dataframe['Zakenpartner_nr'].value_counts()
    dataframe['Aantal_trans_bp'] = 0
    for i in range(len(dataframe)):
        part_nr = dataframe['Zakenpartner_nr'].iloc[i]
        num = freq.loc[part_nr]
        dataframe['Aantal_trans_bp'].iloc[i] = num
    return dataframe

#Business partner physical adress
def physical_adress(dataframe):
    dataframe['has_adress'] = 0
    for i in range(len(dataframe["Zakenpartner_adres"])):
        if not pd.isnull(dataframe["Zakenpartner_adres"].
            iloc[i]):
             dataframe['has_adress'].iloc[i] = 1
    dataframe.drop(['Zakenpartner_adres', '
       Zakenpartner_postcode'],inplace=True,axis=1)
    return dataframe

#Non-complying countries
def risky_countries(dataframe):
    countries = ['AF' , 'BA', 'GY', 'IQ', 'LA', 'SY', 'UG',
       'VU', 'YE', 'ET', 'LK', 'TT', 'TN', 'PK', 'KP']
    dataframe['risky_country'] = 0
    for i in range(len(dataframe['Zakenpartner_land'])):
```

```
        if dataframe['Zakenpartner_land'].iloc[i] in
           countries:
            dataframe['risky_country'].iloc[i] = 1
    return dataframe

#Activity after inactivity
def inactive(dataframe):
    dataframe['inactive'] = 0
    for i in range(len(dataframe)):
        sub = dataframe[dataframe['Zakenpartner_nr'] ==
           dataframe['Zakenpartner_nr'].iloc[i]]
        if len(sub)>1:
            sub = sub.sort_values('Datum_doc', ascending=
               False)
            sub = sub.reset_index()
            r = sub[sub['Doc_nr'] == dataframe['Doc_nr'].
               iloc[i]].index[0]
            if len(sub)>(r+1):
                dataframe['inactive'].iloc[i] = sub['
                   Datum_doc'].iloc[r] - sub['Datum_doc'].
                   iloc[r+1]
    return dataframe
```

9.4 Results of the survey

	Error			Fraud				
Invoice details	**Mean**	**Mode**	**Median**	**Mean**	**Mode**	**Median**	**Text responses sum**	**Summary text**
Double payments	3.88	4	4	1.4	1	1	12	Usually an error, but indication for further research.
Paid through offshore company	3.48	4	4	1.28	1	1	10	Sounds suspicious, but often not applicable.
High travel and phone costs	3	3	3	1.44	1	1	10	Susceptible for fraud, needs further research.
Large amounts of employee expenses	3.92	4	4	1.32	1	1	7	Possible but dependent on the organisation.
Benford's law	3	3	3	1.64	2	2	8	Previous research did not show results.
Anonymous deposits	3.68	4	4	1.28	1	1	9	Very suspicious, but often not possible.
Large numbers of sales from new business partner	3.32	4	3	1.4	1	1	6	Possible, needs further analysis.
Type of entry	3.08	2	3	1.64	2	2	4	Some entries are more susceptible to fraud.
Debit or credit	2.28	2	2	1.76	2	2	4	Not very relevant.
Currency	2.24	2	2	1.68	2	2	5	Only with unexpected currency.
Payment method	2.4	3	3	1.72	2	2	3	Manual entries are more susceptible.
Amount	2.84	2	3	1.68	2	2	4	Invoices just under the light-boundary.
Large and complicated transactions at the end of the year	3.96	4	4	1.24	1	1	5	Susceptible for fraud, are often reviewed.

	Error			Fraud				
Date and time	**Mean**	**Mode**	**Median**	**Mean**	**Mode**	**Median**	**Text responses sum**	**Summary text**
Invoice not paid in time	2.08	2	2	1.96	2	2	3	Only with extreme values.
Date on invoice	2.28	2	2	1.76	2	2	2	In combination with rest of invoice.
Date of processing invoice	2.64	3	3	1.52	2	2	4	On weird times, outside of office hours.
Date of invoice entry	2.68	3	3	1.48	1	1	3	On weird times, outside of office hours.
Date of clearing invoice	2.48	1	2	1.68	2	2	2	Should be investigated when not paid on entry.
Time of processing invoice	2.88	3	3	1.52	2	2	5	Outside office hours can be suspicious.
Invoice processed during working hours	2.68	2	2	1.76	2	2	5	Possibility but some customers work 24/7 or in different time zones.
Time difference between dates mentioned above	2.84	2	2	1.64	2	2	5	Depends on the context.

	Fouten				Fraude			
Business partner	**Mean**	**Mode**	**Median**	**Answers**	**Mean**	**Mode**	**Text responses sum**	**Summary text**
Non-existing business partner	4.36	5	4	1	1.16	1	2	Useful, but not feasible.
Other activities than registered	3.8	4	4	1	1.2	1	3	Larger risk, not always fraudulent.
Risky origin country	3.84	4	4	1	1.24	1	4	Useful if such a list exists.
Many journal entries and corrections	3.72	4	4	1	1.28	1	4	Useful, but can also be a mismatch.
Active after inactivity	3.52	4	4	1	1.36	1	4	Potential for outlier
No physical adress	3.68	4	4	1	1.28	1	5	Very useful, but in some systems not possible.
One time business partner	3	3	3	1	1.64	2	6	Worth investigating.
History blocking	3.52	4	4	1	1.44	1	4	Depending on the reason.
Type of blocking	3.32	3	3	1	1.44	1	3	Depending on the type.
Incomplete bank details	3.24	3	3	1	1.56	2	6	When incomplete, the invoice will not be paid.
Location of business partner	2.8	3	3	1	1.64	2	5	Depending on the expectation of the organisation.

Table 22: Analysis of the results of the survey

9.5 Survey Fraud Indicators

Introduction

Beste lezer,

Allereerst, bedankt voor uw medewerking aan deze enquête. Deze enquête is onderdeel van mijn masteropdracht voor de studie Business&IT en hierin word ik begeleid door de Universiteit Twente. Het onderzoek is in opdracht van de Auditdienst Rijk, onderdeel van het Ministerie van Financiën.

Financiële fraude is iets waar de meeste bedrijven mee te maken krijgen. Deze fraude kan in verschillende vormen plaatsvinden. Het gebruik van algoritmes zou kunnen helpen om deze fraude te detecteren door te zoeken naar afwijkende zaken in de boekhouding. Dit is gebaseerd op het volgende statement uit COS240: “Afwijkingen in financiële overzichten kunnen het gevolg zijn van fraude of fouten. De onderscheidende factor tussen fraude en fouten is het al dan niet opzettelijke karakter van de handeling die aan de afwijking in de financiële overzichten ten grondslag ligt.”

Het onderzoek is erop gericht om algoritmes te gebruiken om (in de eerste plaats) fouten te detecteren in facturen door het zoeken naar afwijkingen. De assumptie wordt hierin gemaakt dat binnen deze afwijkingen ook potentiële fraude opgespoord kan worden. Hiernaast onderzoek ik hoe deze algoritmes hun keuzes kunnen uitleggen en zo het vertrouwen van accountants in algoritmes kan vergroten.

Om de algoritmes te kunnen trainen, is het nodig om aan te kunnen geven welke indicatoren belangrijk zijn voor het detecteren van afwijkingen. Academische- en vakliteratuur is gebruikt om indicatoren te vinden. Hieruit is een selectie van indicatoren gemaakt die te vinden zijn in transactie- en factuurdata. Ik vraag u een inschatting te maken in hoeverre deze indicatoren afwijkingen, fouten en eventueel fraude zouden kunnen voorspellen vanuit uw expertise.

Deze enquête betreft de eerste stap van het onderzoek. De enquête zal ongeveer 15 minuten in beslag nemen. De gegevens en de resultaten van het onderzoek zullen wij anoniem en vertrouwelijk verwerken. De gegevens en resultaten zullen uitsluitend voor analyse gebruikt worden en niet aan derden worden verstrekt. U behoudt zich het recht voor om op elk moment zonder opgaaf van redenen de deelname aan dit onderzoek te beëindigen.

Mocht u verder nog vragen hebben dan kunt u mij te allen tijde bereiken via:
l.h.hamelers@student.utwente.nl of 0651566916.

Ik hoop u hiermee voldoende geïnformeerd te hebben.
Lieke Hamelers

Verklaart u:
- op een duidelijke wijze te zijn ingelicht over de methode en het doel van het onderzoek;
- geheel vrijwillig deel te nemen aan deze enquête;
- volledig in te stemmen met bovenstaande informatie?

- ◯ Ja
- ◯ Nee

Algemene gegevens

In welke sector bent u werkzaam?

- ◯ Private sector
- ◯ Publieke sector
- ◯ Anders, namelijk: [ ]

Wat is de grootte van het bedrijf waar u werkzaam bent?

- ◯ Minder dan 100 werknemers
- ◯ Tussen de 100 en 1000 werknemers
- ◯ Tussen de 1000 en 10.000 werknemers

- ◯ Meer dan 10.000 werknemers

Bent u werkzaam als een financial auditor?

- ◯ Ja
- ◯ Nee
- ◯ Anders, namelijk: _____

Hoe staat u ingeschreven in het register?

- ◯ RA
- ◯ AA
- ◯ Niet ingeschreven
- ◯ Anders, namelijk: _____

Hoeveel jaar ervaring heeft u in uw vak? Vul alleen het aantal jaar in (bv. 5).

_____

In hoeverre bent u bekend met fraude in facturen?

	Niet bekend			Zeer bekend
1	2	3	4	5
Bekend met fraude in facturen				

Details transactie

De vragen in dit blok gaan over details die gekoppeld zijn aan de factuur en/of transactie. Wanneer nodig staat er tussen de haakjes extra uitleg en een eventueel voorbeeld van de indicator.

In welke mate kunnen de volgende indicatoren wijzen op een fout in de transactie? Hiernaast, zou deze indicator ook kunnen wijzen op fraude?

	Hoe effectief is deze indicator voor het vaststellen van afwijkingen?					Detecteert deze indicator ook fraude?		Waarom wel/niet? Voor welke soort afwijkingen/fraude?
	Niet effectief	Licht effectief	Matig effectief	Zeer effectief	Extreem effectief	Ja	Nee	Uitleg (Optioneel)
Dubbele betalingen (Twee keer dezelfde factuur betaald)	○	○	○	○	○	○	○	
Betaald door een offshore bedrijf (Betaling van bedrijf in India terwijl business partner van bedrijf uit Nederland komt)	○	○	○	○	○	○	○	
Hoge reis- en telefoonkosten (Hoger dan andere werknemers)	○	○	○	○	○	○	○	
Grote hoeveelheden dure spullen en stortingen medewerkers (Bepaalde medewerkers die veel en hoge declaraties indienen)	○	○	○	○	○	○	○	
Benford's Law (Een bepaalde, natuurlijk voorkomende verdeling van getallen, de 1 komt vaker voor dan de 2 enz.)	○	○	○	○	○	○	○	
Anonieme stortingen (Geen afzender van de stortingen)	○	○	○	○	○	○	○	

	Hoe effectief is deze indicator voor het vaststellen van afwijkingen?					Detecteert deze indicator ook fraude?		Waarom wel/niet? Voor welke soort afwijkingen/fraude?
	Niet effectief	Licht effectief	Matig effectief	Zeer effectief	Extreem effectief	Ja	Nee	Uitleg (Optioneel)
Grote aantallen aankopen van nieuwe zakenpartner (Veel bestellingen na net aanmelden nieuwe verkoper)	○	○	○	○	○	○	○	
Soort boeking (Factuur, creditnota etc.)	○	○	○	○	○	○	○	
Debet of Credit	○	○	○	○	○	○	○	
Valuta (EUR, GBP, etc.)	○	○	○	○	○	○	○	
Betaalmethode (Automatische incasso, overschrijving etc.)	○	○	○	○	○	○	○	
Bedrag (Het bedrag dat is vermeldt op de factuur)	○	○	○	○	○	○	○	
Grote en ingewikkelde transacties aan het einde van het boekjaar	○	○	○	○	○	○	○	

Tijd en datum

De vragen in dit blok gaan over details die gekoppeld zijn aan de tijd en datum van de factuur en/of transactie. Wanneer nodig staat er tussen de haakjes extra uitleg en een eventueel voorbeeld van de indicator.

In welke mate kunnen de volgende indicatoren wijzen op een fout in de transactie? Hiernaast, zou deze indicator ook kunnen wijzen op fraude?

	Hoe effectief is deze indicator?					Detecteert deze indicator ook fraude?		Waarom wel/niet Voor welke soor afwijkingen/fraud
	Niet effectief	Licht effectief	Matig effectief	Zeer effectief	Extreem effectief	Ja	Nee	Uitleg (optionee
Factuur niet op tijd betaald (Niet betaald voor de uiterste betalingstermijn)	○	○	○	○	○	○	○	
Factuurdatum (Datum op de factuur)	○	○	○	○	○	○	○	
Datum invoer factuur (Datum waarop de factuur in het systeem is gezet)	○	○	○	○	○	○	○	
Datum boeking factuur (Datum waarop de factuur is opgenomen in het boekhoudingssysteem)	○	○	○	○	○	○	○	
Datum vereffening factuur (Datum waarop de factuur wordt betaald)	○	○	○	○	○	○	○	
Tijd invoer (Tijd waarop de factuur in het systeem wordt ingevoerd)	○	○	○	○	○	○	○	
Factuur ingevoerd tijdens werktijd (Indicator of de factuur tussen 6 en 22 uur wordt ingevoerd)	○	○	○	○	○	○	○	

	Hoe effectief is deze indicator?					Detecteert deze indicator ook fraude?		Waarom wel/niet? Voor welke soor afwijkingen/fraud
	Niet effectief	Licht effectief	Matig effectief	Zeer effectief	Extreem effectief	Ja	Nee	Uitleg (optionee
Verschil in dagen tussen bovengenoemde data (De duur tussen invoer van de factuur en boeking, de duur tussen boeking en vereffening etc.)	○	○	○	○	○	○	○	

Zakenpartner

De vragen in dit blok gaan over details die gekoppeld zijn aan de zakenpartner die de goederen of services levert. Wanneer nodig staat er tussen de haakjes extra uitleg en een eventueel voorbeeld van de indicator.

In welke mate kunnen de volgende indicatoren wijzen op een fout in de transactie? Hiernaast, zou deze indicator ook kunnen wijzen op fraude?

	Hoe effectief is deze indicator?					Detecteert deze indicator ook fraude?		Waarom wel/niet? Voor welke soort afwijkingen/fraude?
	Niet effectief	Licht effectief	Matig effectief	Zeer effectief	Extreem effectief	Ja	Nee	Uitleg (optioneel)
Niet bestaande zakenpartner (Niet bestaand bij KvK etc.)	○	○	○	○	○	○	○	

	Hoe effectief is deze indicator?					Detecteert deze indicator ook fraude?		Waarom wel/niet? Voor welke soort afwijkingen/fraude?
	Niet effectief	Licht effectief	Matig effectief	Zeer effectief	Extreem effectief	Ja	Nee	Uitleg (optioneel)
Bedrijf voert andere activiteiten uit dan geregistreerd staat	○	○	○	○	○	○	○	
Risicovol land van afkomst (Land van zakenpartner op lijst met risicovolle landen)	○	○	○	○	○	○	○	
Veel journal entries en correcties van dezelfde zakenpartner (Veel activiteit rondom dezelfde zakenpartner)	○	○	○	○	○	○	○	
Actief na lange periode van inactiviteit (Opeens weer veel facturen na een periode niks)	○	○	○	○	○	○	○	
Geen fysiek adres (De zakenpartner heeft geen fysiek adres opgegeven)	○	○	○	○	○	○	○	
Eenmalige zakenpartner (De zakenpartner heeft maar eenmaal een service geleverd)	○	○	○	○	○	○	○	

	Hoe effectief is deze indicator?					Detecteert deze indicator ook fraude?		Waarom wel/niet? Voor welke soort afwijkingen/fraude?
	Niet effectief	Licht effectief	Matig effectief	Zeer effectief	Extreem effectief	Ja	Nee	Uitleg (optioneel)
Historie blokkering zakenpartner (Of de zakenpartner is geblokkeerd)	○	○	○	○	○	○	○	
Blokkeringstype (Handmatig, Factuur verificatie blokkering, etc.)	○	○	○	○	○	○	○	
Incomplete bankgegevens (Niet alle gegevens zijn volledig)	○	○	○	○	○	○	○	
Locatie van zakenpartner (Land en stad waar de zakenpartner zich bevindt)	○	○	○	○	○	○	○	

Overig

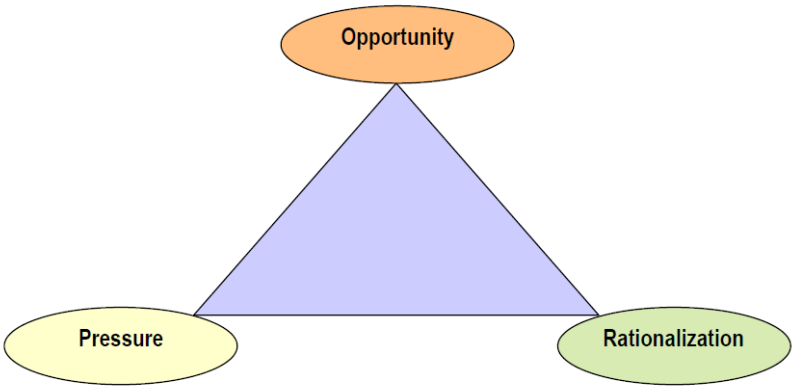
Een van de bekendste modellen dat gebruikt wordt om aanwijzingen voor fraude op te sporen is de fraudedriehoek. Deze wordt hieronder weergegeven. Dit model voert drie redenen aan waarom iemand fraude zou kunnen plegen.

In het gros van de gevallen vormt druk (pressure) de basis voor iemands beweegredenen om fraude te plegen. Deze druk kan zowel intern als extern zijn, en zowel financieel als persoonlijk.

De tweede frauderisicofactor is gelegenheid (opportunity): de kans op fraude wordt groter naarmate het potentiële fraudeurs makkelijker wordt gemaakt. Een onderneming

biedt bijvoorbeeld gelegenheid tot het plegen van fraude wanneer er sprake is van een onduidelijke, chaotische organisatiestructuur en een gebrek aan controle en toezicht.

Het laatste element dat kenmerkend is voor fraudeurs is rationalisatie (rationalization): een fraudeur zal zijn gedrag doorgaans proberen goed te praten. De fraudeur rechtvaardigt zijn normafwijkend gedrag bijvoorbeeld door zichzelf voor te houden dat ‘iedereen het doet’, of door te stellen dat hij het wederrechtelijk verkregen voordeel verdient vanwege zijn harde werk binnen de onderneming.



Ziet u dat een van de hoeken van de fraudedriehoek vaker voorkomt met betrekking tot fraude in facturen? Meerdere antwoorden zijn mogelijk.

- ☐ Opportunity
- ☐ Rationalization
- ☐ Pressure

Zijn er nog indicatoren voor fouten en/of fraude die u mist in deze enquête?

Wilt u de op de hoogte blijven van dit onderzoek? Dan kunt u hieronder uw e-mailadres invullen en zullen de resultaten worden opgestuurd naar u zodra het onderzoek is afgerond.

Powered by Qualtrics

9.6 Survey Round 1 Upward Stream Engagement

Informed consent

Beste lezer,

Allereerst, bedankt voor uw medewerking aan dit interview. Deze enquête is onderdeel van mijn masteropdracht voor de studie Business&IT en hierin word ik begeleid door de Universiteit Twente. Het onderzoek is in opdracht van de Auditdienst Rijk, onderdeel van het Ministerie van Financiën.

Financiële fraude is iets waar de meeste bedrijven mee te maken krijgen. Deze fraude kan in verschillende vormen plaatsvinden. Het gebruik van algoritmes zou kunnen helpen om deze fraude te detecteren door te zoeken naar afwijkende zaken in de boekhouding. Dit is gebaseerd op het volgende statement uit COS240: “Afwijkingen in financiële overzichten kunnen het gevolg zijn van fraude of fouten. De onderscheidende factor tussen fraude en fouten is het al dan niet opzettelijke karakter van de handeling die aan de afwijking in de financiële overzichten ten grondslag ligt.”

Het onderzoek is erop gericht om algoritmes te gebruiken om (in de eerste plaats) fouten te detecteren in facturen door het zoeken naar afwijkingen. De assumptie wordt hierin gemaakt dat binnen deze afwijkingen ook potentiële fraude opgespoord kan worden. Een belangrijke stap in dit onderzoek is om te kijken hoe uitleg bij de beslissingen accountants kan ondersteunen.

Om te bepalen wat een goede uitleg faciliteit is, moeten we er eerst achter komen wat een accountant nodig heeft. Hiernaast is het belangrijk om te bepalen om wat voor soort uitleg een accountant goed reageert. Dit interview heeft als doel om de doelen en requirements voor zo'n faciliteit te ontdekken en te functioneren als basis voor het eerste prototype.

Dit interview zal ongeveer 30-45 minuten in beslag nemen. Tijdens dit interview zullen wij de audio opnemen. De gegevens en de resultaten van het onderzoek zullen wij anoniem en vertrouwelijk verwerken. De gegevens en resultaten zullen uitsluitend voor analyse gebruikt worden en niet aan derden worden verstrekt. U behoudt zich het recht voor om op elk moment zonder opgaaf van redenen de deelname aan dit onderzoek te beëindigen.

Mocht u verder nog vragen hebben dan kunt u mij te allen tijde bereiken via:
l.h.hamelers@student.utwente.nl of 0651566916.

Ik hoop u hiermee voldoende geïnformeerd te hebben.
Lieke Hamelers

Verklaart u:
- op een duidelijke wijze te zijn ingelicht over de methode en het doel van het onderzoek;
- geheel vrijwillig deel te nemen aan deze enquête;
- volledig in te stemmen met bovenstaande informatie?

- ◯ Ja
- ◯ Nee

Algemene gegevens

Wat is je precieze functie?

Hoeveel jaar ervaring heb je in het vak?

Hoe werk je met de accountants samen?

Ervaring werken met accountants

Wat is je ervaring om met accountants te werken?

Zljn er bepaalde factoren waar je voor accountants rekening mee moet houden?

Hoe komen jullie erachter wat accountants willen?

Hoe reageren accountants op de technologie die jullie aanbieden?

Uitlegfaciliteit- eigen indruk

Waarom heeft een accountant een uitlegfaciliteit nodig?

Welke onderdelen zijn minimaal nodig voor een accountant om zijn werk te kunnen doen middels een uitlegfaciliteit?

De volgende doelen heb ik gevonden in literatuuronderzoek. Welke is van toepassing voor accountants?

- [ ] Transparency = Explain how the system works
- [ ] Scrutability = allow users to tell the system it is wrong
- [ ] Trust = Increase users' confidence in the system
- [ ] Effectiveness = Help users make good decisions
- [ ] Persuasiveness = help users to try or buy
- [ ] Efficiency = help users make decisions faster

☐ Satisfaction = increase the easy of usability or enjoyment

Waarom?

Stel: jij mag nu een uitlegfaciliteit ontwerpen voor accountants. Wat zou je erin doen? En waarom?

Uitlegfaciliteit- voorbeelden

Tijdens mijn onderzoek heb ik verschillende categorieën van uitleg gevonden. Binnen elke categorie bestaan er verschillende mechanismen hoe je uitleg kan geven. Ik ben vooral geïnteresseerd hoe dit uitleg geeft naar de accountant toe, niet de werking erachter. Ik zal ze een voor een doorlopen aan de hand van de voorbeelden. Denk bij elk voorbeeld of het bijdraagt aan de doelen, duidelijke uitleg geeft, begrijpelijk is voor de accountant en de juiste dingen weergeeft.

SLIDE 1. Als je de categorieën ziet, wat zijn relevante voor accountants?

SLIDE 2. Wat denk je van deze mechanismen? Draagt het bij aan de doelen van de faciliteit? Is het begrijpelijk voor de accountant?

SLIDE 3. Wat denk je van deze mechanismen? Draagt het bij aan de doelen van de faciliteit? Is het begrijpelijk voor de accountant?

SLIDE 4. Wat denk je van deze mechanismen? Draagt het bij aan de doelen van de faciliteit? Is het begrijpelijk voor de accountant?

SLIDE 5. Wat denk je van deze mechanismen? Draagt het bij aan de doelen van de faciliteit? Is het begrijpelijk voor de accountant?

SLIDE 6. Wat denk je van deze mechanismen? Draagt het bij aan de doelen van de faciliteit? Is het begrijpelijk voor de accountant?

SLIDE 7. Wat denk je van deze mechanismen? Draagt het bij aan de doelen van de faciliteit? Is het begrijpelijk voor de accountant?

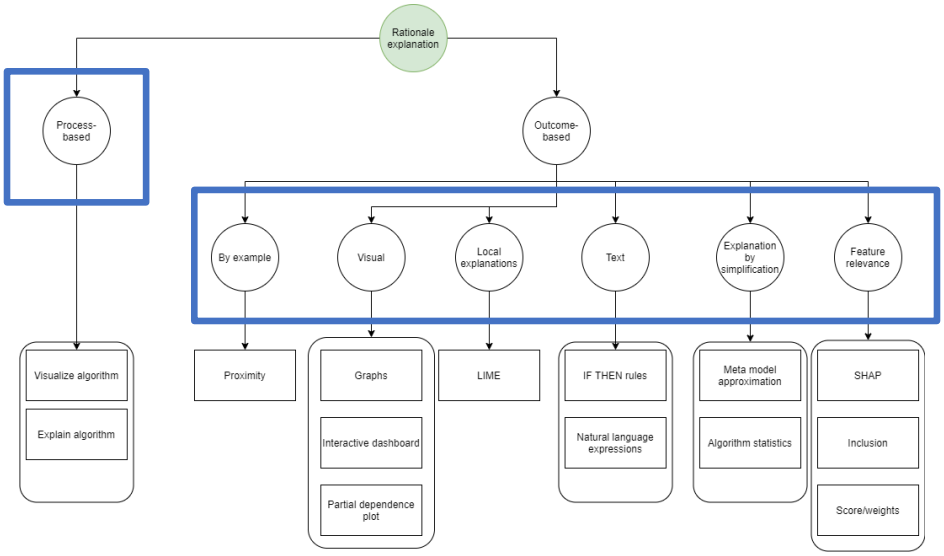
SLIDE 8. Wat denk je van deze mechanismen? Draagt het bij aan de doelen van de faciliteit? Is het begrijpelijk voor de accountant?

SLIDE 9. Wat denk je van deze mechanismen? Draagt het bij aan de doelen van de faciliteit? Is het begrijpelijk voor de accountant?

Zljn er nog mechanismen die je hebt gemist die veel kunnen toevoegen voor een accountant?

Powered by Qualtrics

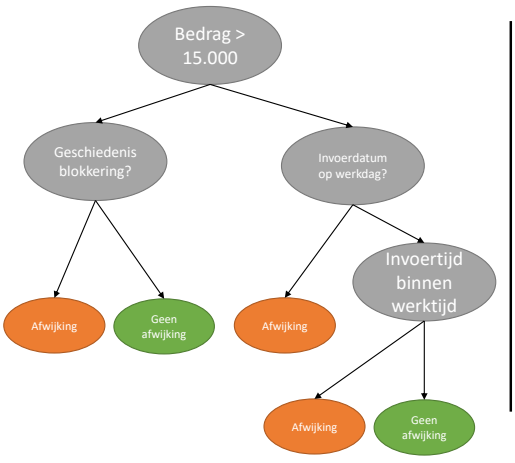
9.7 Examples Explanation Mechanisms Round 1



Explanation Mechanisms

Case: transactie 1007 is beoordeeld als afwijking door het algoritme. De volgende soorten uitleg worden gegeven.

1.1 Visualisatie model



1.2 Uitleg werking model (tekst)

Decision Tree
Een beslissingsboom, beslisboom of alternatievenschema is een boomstructuur voor de weergave van de alternatieven en keuzen in een besluitvormingsproces, en is een techniek uit de besliskunde. Het is een bijzonder geval van een stroomdiagram, namelijk een zonder cykels, en met als enige actie steeds het kiezen van een tak.
Het is een gerichte graaf met een startpunt, met bij elke knoop een vertakking. In of bij de knoop staat een vraag, en bij de takken staan mogelijke antwoorden. Soms staat, aanvullend, in de knoop een omschrijving van het resultaat tot aan dit punt (de categorie gevallen of uitkomsten, die bij elke stap verkleind wordt tot een subcategorie, uiteindelijk leidend tot één bepaalde uitkomst).

Meer uitleg:
https://www.youtube.com/watch?v=7VeUPuFGJHk

Process-based

2.1 Proximity

Transactie 1007
Afwijking = 0.8

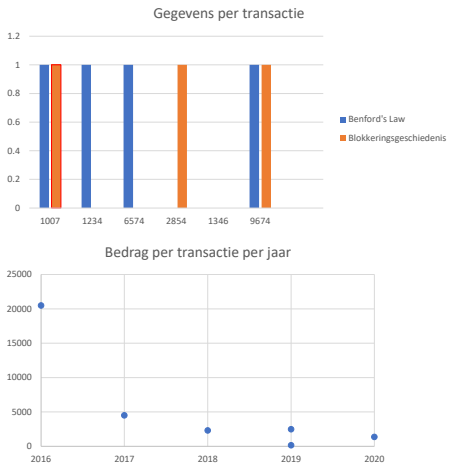
De volgende transactie komt voor 83% overeen met transactie 1007 en is de dichtstbijzijnde.

Transactie 1007		Transactie 1234	
Bedrag	14.321	Bedrag	16.321
Tijd invoer uur	15	Tijd invoer uur	15
Tijd invoer jaar	2019	Tijd invoer jaar	2019
Benford's law	1	Benford's law	1
Betaalmethode	AA	Betaalmethode	AA
Afwijking	0.8	Afwijking	0.8

Outcome-based Example

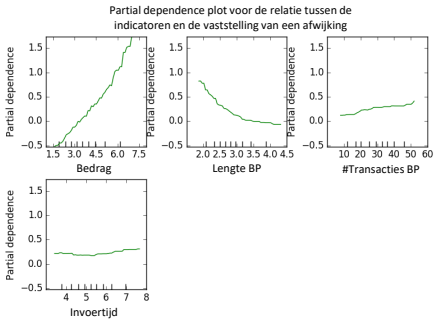
3.1 Graphs

Transactie 1007
Afwijking = 0.8



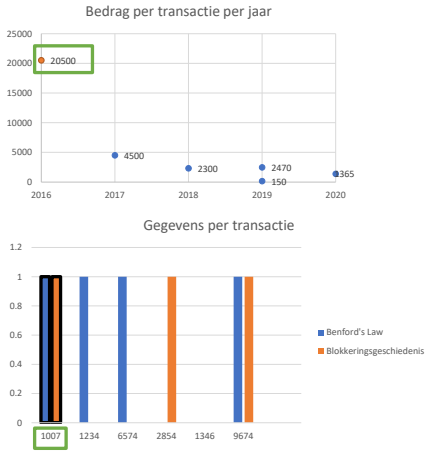
3.2 Partial dependence plots

Transactie 1007
Afwijking = 0.8



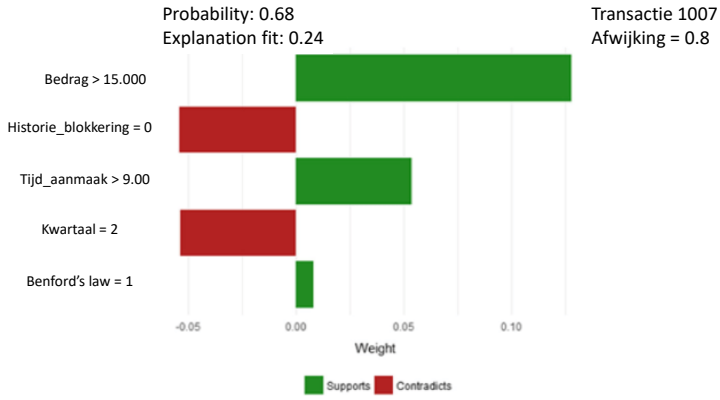
Outcome-based Visual

3.3 Interactive dashboard



Outcome-based
Visual

4.1 LIME



Outcome-based
Local

5.1 IF THEN Rules

Transactie 1007
Afwijking = 0.8

IF (bedrag > 15.000 AND jaar_invoering = 2019)
THEN afwijking = 1

IF (bedrag < 15.000 AND jaar_invoering = 2018)
THEN afwijking = 0

IF (Transacties_BP > 300 AND Tijd_invoering < 18)
THEN afwijking = 0

5.2 Natural language expressions

Deze transactie wordt gezien als afwijkend omdat het bedrag over 15.000 euro is. Hiernaast is de zakenpartner relatief nieuw (23-09-2020) en zijn er nog weinig transacties (4) uitgevoerd.

De zakenpartner heeft geen historie met blokkeringen en het bedrag is niet afwijkend aan de hand van Benford's law.

Outcome-based
Natural language

6.1 Algorithm Statistics

Transactie 1007
Afwijking = 0.8

Confidence = 0.87
Accuracy = 0.94
Recall = 0.73
Precision = 0.69
F1 Score = 0.71

Outcome-based
Simplification

7.1 Inclusion

De volgende indicatoren zijn gebruikt om de beslissing te maken:

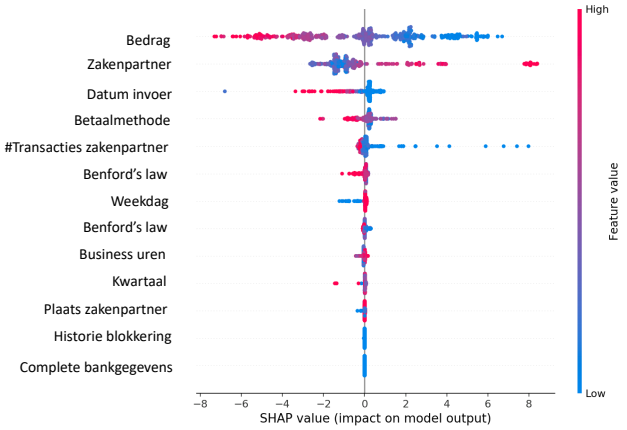
- Bedrag
- Aantal transacties zakenpartner
- Tijd van factuur
- Risicovol land zakenpartner
- Kwartaal

7.2 Scores/Weights

Indicator	Score
Bedrag	0.87
Aantal transacties zakenpartner	0.56
Tijd van factuur	0.34
Risicovol land zakenpartner	0.66
Kwartaal	0.22

Outcome-based
Feature relevance

7.3 SHAP



Outcome-based
Feature relevance

9.8 Treatment validation survey

Several questions were asked during this phase to receive data on certain goals. Below are the questions listed that were asked.

General

1. What is your general impression of this explanation facility?
2. What are some of the advantages of this explanation facility?
3. what are some of the disadvantages of this explanation facility?

Goals

1. Has this facility increased the transparency of the algorithm and outcomes?
2. Has this facility increased your trust in the algorithm and the outcomes?
3. was the facility satisfying to use?

Information

1. Was the offered information sufficient to determine if an invoice was outlying?
2. What information would be beneficial to add to improve the facility?

Ease of use

1. Would you use this facility again? Why?
2. Was it clear where each type of information could be found? Would you rearrange any part of the interfaces?