The Creation of an Explainable Artificial Intelligence Model to Enhance Interpretability and Transparency for ING in Their Fight Against Transactional Fraud

Stijn van der Pol

Industrial Engineering and Management - Financial Engineering and Management,
University of Twente, Drienerlolaan 5, Enschede, The Netherlands.

Abstract

This thesis presents the development of an Explainable Artificial Intelligence (XAI) model aimed at enhancing the interpretability and transparency of fraud detection systems at ING. By integrating XAI techniques, the model provides clear and actionable insights into the decision-making processes of machine learning classifiers, thereby bridging the gap between high model accuracy and the need for understandable outputs. Through extensive experimentation and validation with real-world data, this research demonstrates the practical applicability of XAI in transactional fraud detection. The insights gained from expert interviews and use case testing further reinforce the model's effectiveness in improving stakeholder communication and decision-making. This work contributes to the ongoing efforts to develop AI systems that are not only powerful but also transparent and trustworthy in combating financial fraud.

First Supervisor: Jörg Osterrieder
Second Supervisor: Patricia Rogetzer
External Supervisor: Remco Stam

Keywords: Explainable Artificial Intelligence (XAI), Fraud Detection, Model Interpretability, Model Transparency, Transactional Fraud

Contents

List of Figures

List of Tables

List of Acronyms

XAI Explainable Artificial Intelligence
ML Machine Learning
ROC Receiver Operating Characteristic
AUC Area Under the ROC Curve
RF Random Forest
SVM Support Vector Machine
GB Gradient Boosting
GDPR General Data Protection Regulation law
DARPA Defense Advanced Research Projects Agency
ACM Association for Computing Machinery
FAT Fairness Accountability, and Transparency
MO Modus Operandi

1 Introduction

Within this chapter, an overall introduction about the content of this research is provided. In each chapter, a small table of contents is provided to enhance readability. In this chapter, the reader is provided with an introduction to fraud, the company ING, the structure of this research and an understanding of the methodology this research will follow.

Contents

1.1 General Information

The utilisation of advanced technology, particularly in the area of fraud detection, has become increasingly more dominant. Currently, there is an increase in online transactions Skibińska-Fabrowska (2023) , a growth of data Pascua, Prado, Solis, Cid-Andres, and Cambiador (2019) and the day-to-day work tasks of people is changing rapidly Ramos (2023). It is a time where new innovations are are happening constantly Pascua et al. (2019). The number of processed data and, therefore, information has been increasing at a very rapid pace. Every sector has to adapt to these changes, and so does the banking sector. With the increasing focus on online banking, the emergence of new client features, advancements in technology to enhance system efficiency, and the growing potential for criminal exploitation of these developments, a need arises for robust mechanisms to combat fraudulent activities.

With the fourth industrial revolution having arrived and the use of Big Data these days, our society has been widely concerned with the large numbers of personal information that is all over the world. There has been a shift within institutions where the protection of data and, therefore, privacy has been more and more important. An example has been the introduction of the General Data Protection Regulation law (GDPR). These rules include ensuring accuracy, data minimisation, and data security, all of which must be adhered to when using personal data for, for example, fraud detection purposes. Especially within domains like fraud detection, personal data is of extreme importance. To comply with GDPR, fraud detection models must be designed and developed with these principles in mind. This includes ensuring transparency and accountability for the processing of personally identifiable information (PII) Boiarskaia, Albert, and Lee (2019).

ING has been actively improving their models (for example transaction monitoring for fraud detection) to make them more resilient against the constantly changing fraudulent tactics performed by criminals, while also making sure they comply with regulations such as the GDPR. Implementing machine learning into their models has proven to be helpful for the scenario provided above. However, the addition of machine learning often results in complex, unclear, and high-dimensional models, making it harder for programmers and model validators to ensure transparency expectations are met. As complexity arises, there are also evolving technologies which aim to provide a solution to reduce this.

One of these technologies is Explainable AI (XAI). Essentially, XAI serves as a means to comprehend the decision-making processes of complex artificial intelligence algorithms Kumar (2022). In simpler terms, it helps to understand why AI systems make the decisions they do. This thesis aims to delve into the application of XAI in the context of model validation for fraud detection. Against the backdrop of evolving regulatory frameworks and the escalating volume of data, this research seeks to highlight the importance of transparent AI methodologies in ensuring the integrity of financial systems.

By exploring how XAI contributes to the validation of AI models, particularly in the domain of fraud detection, this study sheds light on its significance in adapting to changing regulatory environments and harnessing the potential of Big Data. This thesis provides insights into the dynamics at play in the intersection of XAI, model validation, and fraud detection.

1.2 Company Introduction

ING Group, or ING Groep, is a prominent Dutch multinational banking and financial services corporation headquartered in Amsterdam, with its origins traced back to the 1991 merger between Dutch insurer Nationale-Nederlanden and the national postal bank NMB Postbank I.N.G. (2023a). ING has a significant presence across Europe and beyond, it serves approximately 37 million individuals, corporate entities, and financial institutions in over 40 countries. The company employs 60,000 people globally and primarily offers a range of financial products and services, including savings, payments, investments, loans, and mortgages in its retail markets I.N.G. (2023b). In terms of market presence, ING is a market leader in the Netherlands, Belgium, and Luxembourg.

ING Research Influence

This research focuses on the added value the explanations of an XAI model can provide to departments in the field of fraud within ING. These explanations are qualitatively validated by experts currently working in these departments. In addition, the dataset this model is trained and tested on consists of historical real life transactions which have taken place in the bank. Moreover, an additional dataset is provided containing false positive fraudulent transactions, meaning these transactions were originally flagged as fraudulent but turned out to be correct. The qualitative validation and multiple datasets have all been facilitated by ING. In addition, the explanations resulting from the XAI model can potentially improve real life workload within the bank. This research focuses on the implementation and potential added value of XAI for all relevant departments. Without the influence of the bank, this research would not have been possible.

1.3 Fraud Detection

Before going into the fraud detection domain within ING, let's first introduce the concept of fraud itself. Fraud can be defined as an intentionally deceptive action designed to provide the perpetrator with an unlawful gain or to deny a right to a victim Chen (2024). It involves the false representation of facts, whether by intentionally withholding important information or providing false statements to another party for the specific purpose of gaining at the expense of the victim. Fraud can be both ethically wrong and a criminal act, and its consequences may include fines, imprisonment, and civil litigation to recover monetary damages Chen (2024).

The most widely accepted explanation for why some people commit fraud is known as the Fraud Triangle Inspector General (2021). This explanation hypothesises three factors that cause fraud: financial pressure, the identification of opportunity, and rationalisation for their actions. As illustrated in figure 1, these three components together form the Fraud Triangle, suggesting that the presence of all or some of them increases the likelihood of a person engaging in fraudulent activities.

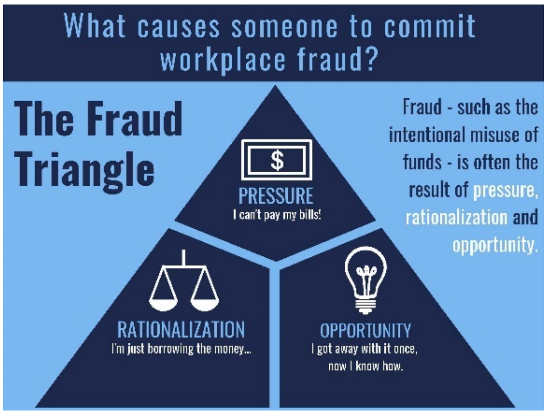

What causes someone to commit workplace fraud?
The Fraud Triangle
$
PRESSURE
I can't pay my bills!
RATIONALIZATION
I'm just borrowing the money...
OPPORTUNITY
I got away with it once, now I know how.
Fraud - such as the intentional misuse of funds - is often the result of pressure, rationalization and opportunity.

Fig. 1: The Triangle of Fraud

Within ING, a lot of models for fraud detection still follow the rule-based decision-making approach van der Pol (2023). Rule-based Machine Learning (ML) models in fraud detection adhere to a set of predefined rules to classify transactions as legitimate or fraudulent. These rules, usually created by domain experts, are based on known patterns of fraudulent behaviour. For instance, a rule might flag transactions that exceed a certain threshold or occur in a specific geographical location. Such a threshold can vary across divisions and even models. The advantage of rule-based models lies in their transparency and ease of interpretation. However, they may struggle to adapt to new types of fraud that deviate from the predefined rules Saiful Islam (2023).
In addition to a rule-based pattern discovery, more extensive and complex Machine Learning models have also been implemented. Traditional ML models learn to detect fraud by being trained on historical data, as illustrated in figure 2. These models can encompass various techniques such as random forest, decision tree, naive Bayes, and logistic regression Arrieta and Dıaz-Rodrıguez (2019). They can identify complex patterns in the data that may be challenging for humans to detect Sarker (2021), potentially making them more effective at identifying new types of fraud.

The process depicted in figure 2 illustrates how ML models can predict fraudulent patterns using historical data. As shown on the left side of the figure, when a transaction is requested, either rule-based or ML models predict whether the transaction is fraudulent based on certain thresholds defined by the policies of the banks which are implemented by model developers. When processed, many transactions are initially predicted as fraudulent Weerts (2019) due to the high number of features marking a transaction as potentially fraudulent. However, this is not the final judgement. Explanations for the marking of the transaction are retrieved, and a fraud analyst will have a final look before approving or disapproving the transaction. This final look is a manual step, meaning there is human interference in this process as well.

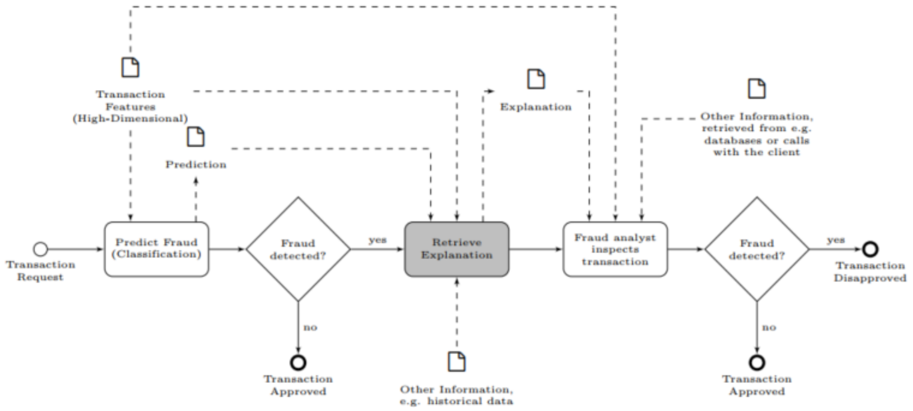

Transaction Features (High-Dimensional)
Prediction
Explanation
Other Information, retrieved from e.g. databases or calls with the client
Transaction Request
Predict Fraud (Classification)
Fraud detected?
yes
no
Retrieve Explanation
Fraud analyst inspects transaction
Fraud detected?
yes
no
Transaction Disapproved
Transaction Approved
Other Information, e.g. historical data
Transaction Approved

Fig. 2: Fraud Detection process Weerts (2019)

As shown in figure 2, the ”Retrieve Explanation” box is coloured grey. This is the area where Explainable AI (XAI) can be most effective and helpful. Making models more transparent and showing the analyst how a prediction came to be can enhance the analyst’s productivity and reduce potential bias. In other words, this is the scope of this research because it can be particularly helpful in this area.

1.4 Problem Context

As mentioned in section 1 so far, the models used by model validators have become more extensive, with higher dimensionality and less transparency Arrieta and Dıaz-Rodrıguez (2019). Machine learning models have grown in accuracy over time due to advancements in computational power, the availability of large datasets, and the development of new algorithms Cylynx (2020). However, this increase in accuracy often comes with an increase in model complexity.

The evolution of machine learning dates back to the 1950s, with significant advancements in the 1990s when the data-driven approach emerged Synetics (2017). Initially, ML models such as the rule-based models are a lot less complex compared to the new ML models. However, as the field evolved, models became more intricate, capable of capturing complex patterns and relationships in data Dong (2017). For instance, deep learning, a subset of machine learning, utilizes neural networks with multiple layers to capture complexity. The more layers are used in the neural network, the more complexity it can capture making it more accurate Dong (2017). Models with high accuracy in capturing complexity are often considered black boxes, where input and output are known, but the model’s internal workings are unknown. There is typically a trade-off between model accuracy and interpretability, as shown in figure 3.

Model complexity refers to the number and type of parameters, features, and interactions a model uses to learn from data Dong (2017). Complex models can capture more nuances and patterns in the data but may be more prone to overfitting, learning too much from noise in the data and failing to generalise new data Kumar (2022). The complexity of a model depends on the structure, relationships, and attributes of the data entities involved in the design. Factors contributing to model complexity include entity relationships, data dependencies, and scalability considerations to name a few Synetics (2017). This increase in complex capabilities enables models to become accurate in solving complex problems.

Model interpretability, on the other hand, refers to the ability to understand the internal workings of a model or how it makes predictions Barceló, Monet, Pérez, and Subercaseaux (2020). Interpretable models, like linear regression or decision trees, are easier to understand and explain but may not capture complex patterns as effectively as more complex models. For very complex problems, these interpretable models are less accurate. The interpretability definition will be further explained in section 1.5, where the core problem is formed.

As models become more complex, they also become harder to interpret. Explainable AI (XAI) aims to make predictions of complex models more understandable Kumar (2022). Transparency can function as an argument to increase trust, which is fundamental in the banking industry Arrieta and Dıaz-Rodrıguez (2019). As mentioned by Arrieta and Dıaz-Rodrıguez (2019), Figure 3 below illustrates the trade-off between interpretability and accuracy, often associated with complexity. The figure shows where XAI holds potential in terms of model explainability.

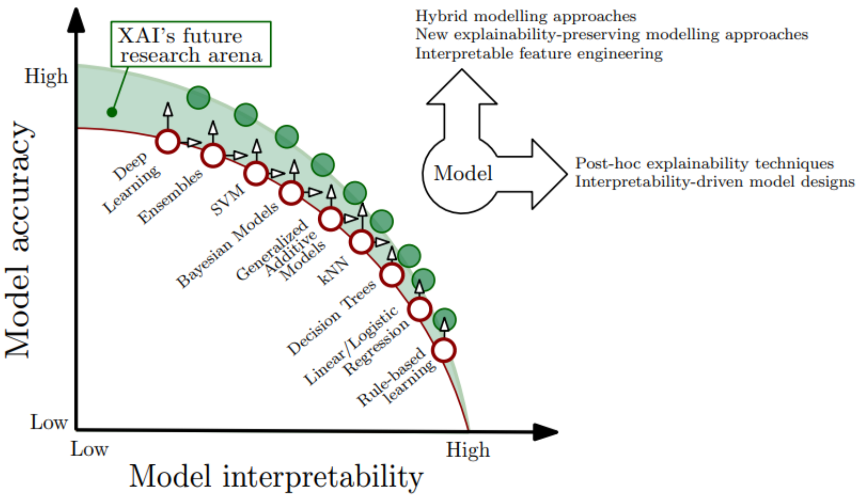

XAI's future research arena
High
Low
Model accuracy
Low
High
Model interpretability
Deep Learning
Ensembles
SVM
Bayesian Models
Generalized Additive Models
kNN
Decision Trees
Linear/Logistic Regression
Rule-based learning
Hybrid modelling approaches
New explainability-preserving modelling approaches
Interpretable feature engineering
Model
Post-hoc explainability techniques
Interpretability-driven model designs

Fig. 3: Trade-off between accuracy and interpretability Arrieta and Dıaz-Rodrıguez (2019)

1.5 Core Problem

Combining all context provided above, the core problem that this research aims to solve can be identified and phrased as follows:

The escalating complexity of machine learning models in fraud detection leads to a troubling trend: they are increasingly veering towards black box methods, decreasing users' ability to comprehend decision-making processes, thereby hindering accurate performance evaluation and potentially compromising fraud detection efficacy.

In the literature, various terms are used to describe the opposite of the ”black box” nature of AI and machine learning models, particularly deep learning models. To clarify the core problem, a start of the main concepts for the XAI terminology is explained. Later in the thesis, when XAI models are used for explaining that fraud flags in transaction data, these terms will be further addressed to explain obtained results.

Transparency: A model is considered transparent if it has the potential to be understandable by itself. Transparency serves as the opposite to the black-box nature of certain AI and machine learning models Arrieta and Dıaz-Rodrıguez (2019). In terms of fraud detection, a transparent model offers stakeholders insight into its internal mechanisms, allowing for a clear understanding of how input data translates into output predictions.

Interpretability: As highlighted in section 1.4, Interpretability refers to the capacity of a model to provide interpretations in terms that are understandable to a human Barceló et al. (2020). In other words, an interpretable model allows stakeholders to understand the underlying logic and reasoning behind its predictions Bau and Gilpin (2019). This is what makes it different from model transparency. While transparency refers to the openness and visibility of the model, it does not particularly emphasise on providing human-understandable explanations Arrieta and Dıaz-Rodrıguez (2019). For instance, in the domain of fraud detection, explainability techniques such as attention mechanisms and graph-based explanations have been employed to provide rationales for the model's predictions, thereby enhancing the trust and understanding of the model's decision-making process Qin and Liu (2022). By using these techniques, machine learning models can provide transparent and human-interpretable explanations for why a certain activity is classified as fraudulent, enabling stakeholders to validate and trust the model's decisions.

1.6 Research Problem

With the constructed core problem from section 1.5 in mind, the following main research question has been constructed:

How can Explainable Artificial Intelligence (XAI) techniques be utilized to enhance the transparency and interpretability of machine learning models in the detection of fraudulent transactions across all involved departments in a bank?

This research question directly addresses the core problem by exploring how XAI can mitigate the issues associated with the complexity of machine learning models in fraud detection. By investigating the application of XAI techniques, this research aims to enhance the transparency and interpretability of these models, allowing users to better understand the decision-making processes. This, in turn, could lead to more accurate performance evaluations and potentially improved fraud detection efficacy. The question implies a focus on both the technical improvement in detection and the human aspect of interpretation, thereby ensuring that the solutions are not only effective but also comprehensible and actionable by the users. This main research question is split into multiple sub questions.

Sub-question 1: What are the key concepts, properties, models, model outcomes, and challenges in Explainable AI and its application in fraud detection?
This sub-question will be addressed in section 2 by reviewing existing literature on machine learning models and XAI including their relevance and limitations in fraud detection. A stepwise approach is created which concludes in a complete overview of XAI as shown in section 2.3.1. The literature review will also cover various types of fraud, and the research that has been done for the use of XAI in fraud.

Sub-question 2: What is the taxonomy of XAI techniques, and how can they be categorized to address different explainability needs for different use cases?
This sub-question will be explored by developing a comprehensive taxonomy of XAI methods, distinguishing between different types of explainability categories and techniques. The taxonomy will help in understanding the various approaches to making AI models interpretable and their applicability to fraud detection. This sub-question will be tackled in section 3.

Sub-question 3: How can XAI methods be implemented to create an effective and efficient model for detecting fraudulent transactions?
This sub-question will be answered in section 4 by detailing the development and implementation of the XAI model, including data processing, testing multiple ML classifiers and the application of specific XAI methods. The chapter will also present the results of the model's effectiveness in fraud detection.

Sub-question 4: How do fraud detection experts evaluate the effectiveness and interpretability of the XAI model and what are the practical insights and implementation opportunities identified from this evaluation?
The first part of the sub-question will be addressed in section 5 by focusing on human-grounded and application grounded evaluation metrics. Fraud detection experts are classified in this research as the departments that are involved with fraud detection. Within ING, this means Data Science,

Rule Writing and Alert Handling. All departments are interviewed and in section 5, department-specific focus points are created. In addition, test cases are constructed in this chapter to test some specific use cases for the XAI model within fraud. Section 6 presents the results of interviews conducted with fraud-related departments. The findings will highlight the model's potential and any identified weaknesses or areas for improvement.

1.7 Research Methodology

The chosen research methodology for this research is the Cross-Industry Standard Process for Data Mining (CRISP-DM) research methodology. It is a widely used process framework for data mining and analytics. It consists of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. These are connected in the way made visual in 4. This method is beneficial for XAI in the field of model validation for fraud detection because it provides a systematic approach to data analysis, which is essential for building transparent and interpretable machine learning models Schröer and Kruse (2021). By following CRISP-DM, this research can ensure that the process of developing and validating the models is well-structured and comprehensive, ultimately leading to more reliable and explainable results.

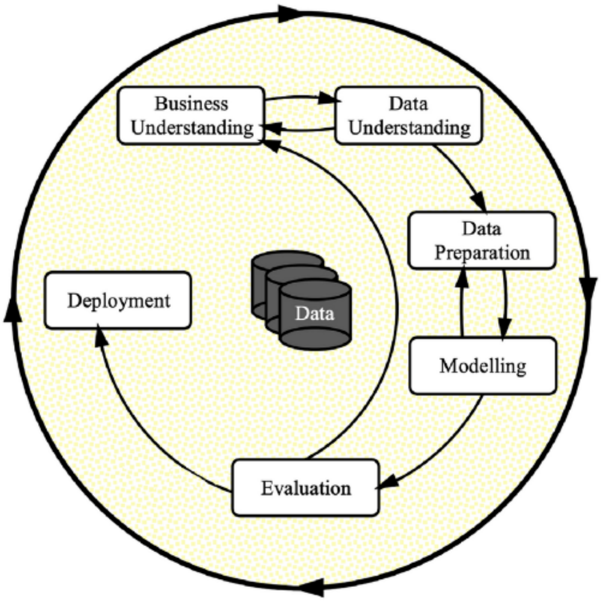


Fig. 4: CRISP-DM process Johnson and Khoshgoftaar (2023)

The CRISP-DM model is particularly relevant in the context of fraud detection, as it emphasizes the importance of understanding the business context and the specific requirements of the problem at the outset of a project Schröer and Kruse (2021). This is crucial for developing AI models that are not only accurate but also interpretable and aligned with the needs of domain experts and stakeholders Carneiro and Junior (2010). Additionally, the evaluation phase of CRISP-DM encourages thorough model validation, which is essential for assessing the performance and reliability of fraud detection models, especially in high-stakes applications such as financial services Belharet, Bharathan, and Dzingina (2020). By following the CRISP-DM process, researchers and practitioners can ensure that their AI models are not only accurate but also transparent, interpretable, and well-aligned with the specific needs of the domain Carneiro and Junior (2010).

CRISP-DM Plan of Approach

As mentioned in section 1.7, there are multiple phases within CRISP-DM with each enabling a different part within this research. Again, it consists of six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. Below, the content of each phase is described.

Business Understanding: This is the first step in the CRISP-DM process. In the context of XAI for fraud detection, this involves understanding the objectives of the project, such as improving the accuracy of fraud detection models, enhancing trust in AI models, and ensuring regulatory compliance Carneiro and Junior (2010). A large part of the business understanding phase has been documented within section 1. A comprehensive understanding of XAI is documented in both sections 2 and 3.

Data Understanding: The next step in the CRISP-DM process is understanding the selected data. As shown in 2, data in this research will consist of historical transaction data. The dataset provided by ING will be explained and structured in section 4.

Data Preparation: This phase involves cleaning the data and transforming it into a suitable format for analysis. For this research, it covers how to deal with the imbalance of the dataset that is provided. Multiple ML classifiers are tested. This is documented in section 4.

Modeling: This involves selecting and applying various modelling techniques and combining all to create one comprehensive and extensive model for ING. As with the data understand and preparation phase, this is all documented in section 4.

Evaluation: This involves assessing the model to ensure it meets the business objectives defined in the first step. This is only of the biggest parts of this research, being present in both section 4 and 6. Here, the XAI model's performance and implementation possibilities are evaluated from both a qualitative and quantitative technical and business points of view. Departments that are closely involved with fraud detection are interviewed to facilitate this evaluation.

Deployment: This involves implementing the model in the operational environment of ING and monitoring its performance. This involves deploying the model in a real-time fraud detection system and monitoring its performance in detecting fraudulent transactions. This is outside the scope of this research.

1.8 Data used

The dataset provided by ING consists of multiple e-banking transactions which have been processed by the bank. The dataset has been anonymised completely to prevent the doom scenario of confidential and personal data leaking.The total size of the dataset consists of 75 million rows which are all individual transactions. The content of the dataset consisted of 50 anonymous features (e.g. feature 1 and feature 39) as columns, shown in appendix A. Each row is an individual transaction which has a value for each feature. The exact content for the determining of the features will not be discussed in this research. However, the dataset itself is describable as a binary classification in which the result of the features determine whether a transaction is considered fraud (it has been labelled with a 1 in the "label" column) or correct (it has been labelled with a 0).

1.9 Scope

The scope of this research is comprehensive and has multiple phases. The research focuses how XAI techniques can enhance transparency and interpretability in machine learning models used for fraud detection. Within the scope of the research, the deployment phase is out of reach given the following circumstances. Deployment considerations involve exploring the practical implications of deploying XAI-enhanced models in real-world banking environments. This includes assessing factors like scalability, performance, usability, and integration with existing systems and processes. This could be a significantly long research on its own. Given these arguments, and the central role of confidentiality within ING, the deployment phase is out of the scope for this research.

2 Literature Review

This section focuses on the literature to clarify the ML models, terminology and techniques surrounding XAI. First, a definition of the technology is explained and emphasised. After which a more in depth explanation is be provided. In the end, a terminology will be created to gain a full overview of the characteristics and relevant challenges. In addition, limitations of XAI are analysed and discussed.

Contents

2.1 An Explanation About Machine Learning

Before a literature review about XAI is given, an explanation about Machine Learning is provided in order to prevent confusion and improve readability of the document. The reason for this is how close both terms are related. Machine learning is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and statistical models, enabling computers to improve their performance on a specific task through experience Kubat (2017). It involves the use of large amounts of data to train models, allowing them to identify patterns, make decisions, and improve their performance over time without being explicitly programmed for each task Janiesch, Zschech, and Heinrich (2021).

Machine learning is extensively used in banking for fraud detection. It helps in the early detection of fraudulent activities, thereby mitigating the associated losses Hanae, Youssef, and Saida (2023). These techniques analyse large volumes of transactional data to detect patterns and anomalies that may indicate potential fraud. Additionally, machine learning models contribute to the development of efficient and accurate fraud detection systems, thereby safeguarding both customers and financial institutions from financial losses and reputation damage Hashemi, Mirtaheri, and Greco (2023).

Several research papers and studies have highlighted the application of machine learning in banking fraud detection. For instance, a study by Achary and Shelke (2023) proposed a machine learning-based approach to successfully contribute to fraud detection in banking transactions. Another study by Gopavaram and Vinothiyalakshmi (2023) focused on the use of machine learning and deep learning algorithms, such as Convolutional Neural Network (CNN), Support Vector Machine (SVM), and Artificial Neural Networks (ANN), to predict fraudulent transactions in a cloud-based credit card fraud detection system. Furthermore, the use of graph-based machine learning models for banking fraud detection has also been explored, demonstrating improved performance in detecting potentially fraudulent transactions Bukhori and Munir (2023).

2.2 Types of Machine Learning

Within the introduction to machine learning mentioned above, algorithms like Convolutional Neural Networks and Support Vector machines are mentioned. These algorithms are both supervised machine learning models.Machine learning is already a subfield of AI. However, this subfield can be categorised into its own subfields. The three most known machine learning types are: supervised learning, unsupervised learning and reinforcement learning.

2.2.1 Supervised Machine Learning

Supervised learning is a type of machine learning where the model is trained on a labelled dataset, meaning it is provided with input data along with the corresponding correct outputs Hastie, Tibshirani, and Friedman (2008). The goal is for the model to learn to map the input to the output. It learns from the dataset by adjusting its internal parameters through optimisation. Once the model is trained, it can be used to make predictions on new, unseen data Barrionuevo, Ramos-Grez, Walczak, and Betancourt (2021). Common algorithms that can be used in supervised learning include linear regression, logistic regression, decision trees, random forests, CNN's and SVM's to name a few. The process of supervised learning involves the following key steps:

- **Data Collection and Labelling**: A labelled dataset is collected, where each data instance is paired with the correct output. For example, in a dataset for predicting housing prices, each house's features (e.g., number of bedrooms, square footage) are paired with its actual sale price Ho, Tang, and Wong (2020). The categorising of data is visualised in 5 to increase understandability.
- **Model Training**: The labelled data is used to train the model. During training, the model adjusts its internal parameters to minimise the difference between its predictions and the true outputs in the training data. This process is often carried out using optimisation algorithms such as gradient descent Hastie et al. (2008).
- **Model Evaluation**: Once trained, the model is evaluated on a separate set of labelled data called the test set. The model's predictions are compared to the true outputs in the test set to assess its accuracy and generalisation to new, unseen data.
- **Prediction**: After successful training and evaluation, the model can be used to make predictions on new input data for which the correct outputs are not known.

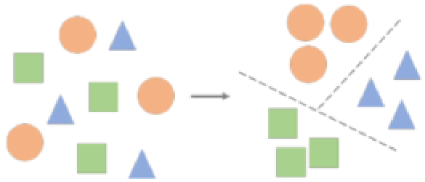


Fig. 5: Supervised Machine Learning: the input data is categorised Sah (2020)

In banking, supervised learning is widely used for credit scoring. By training a model on historical data that includes information about customers and their creditworthiness, such as income, credit history, and loan performance, banks can predict the likelihood of a customer defaulting on a loan. This allows banks to make more informed decisions when approving or denying credit applications Chung and Zhang (2024). Additionally, supervised learning is employed for fraud detection in banking. Models are trained on labelled data that includes both legitimate and fraudulent transactions, enabling them to identify patterns indicative of fraudulent activity and flag suspicious transactions in real time Suriyanarayanan (2020).

2.2.2 Unsupervised Machine Learning

In unsupervised learning, the model is trained on an unlabelled dataset, and the objective is to find hidden patterns or intrinsic structures in the input data Hastie et al. (2008). Unlike supervised learning, there are no correct outputs or labels provided. Instead, the model explores the data to discover its underlying organisation Sarker (2021). Clustering and dimensional reductions are two primary tasks in unsupervised learning. Clustering algorithms, such as K-means and hierarchical clustering, group similar data points together J. Wang and Biljecki (2022). Dimensional reduction techniques, like principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE), aim to reduce the number of features in the data while preserving important relationships. The process of unsupervised learning involves the following key steps:

- **Data Exploration**: The unlabelled dataset is explored to understand the inherent structure and relationships within the data J. Wang and Biljecki (2022). This can involve techniques such as clustering, dimensional reduction, and density estimation.
- **Pattern Discovery**: Unsupervised learning methods aim to identify patterns, similarities, or differences in the data without being explicitly told what to look for Hastie et al. (2008). For example, clustering algorithms group similar data points together based on their features, while dimensional reduction techniques seek to represent the data in a lower-dimensional space while preserving important relationships. This can be seen in figure 6.
- **Model Training and Evaluation**: Unlike supervised learning, unsupervised learning does not have a separate test set with known labels for model evaluation. Instead, the quality of the learned representation is often assessed by human judgement or by its performance on downstream tasks.

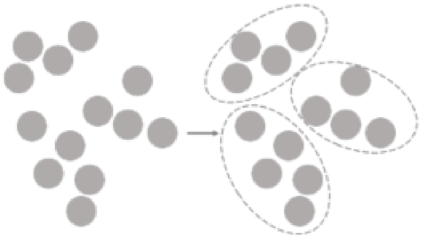


Fig. 6: Unsupervised Machine Learning: the input data is clustered Sah (2020)

Unsupervised learning is utilised in banking for customer segmentation. By analysing transactional data and customer behaviour, unsupervised learning algorithms can group customers into segments based on similarities in their spending habits, investment preferences, or risk tolerance Purohit and Vats (2023). This information can then be used to tailor marketing strategies, develop personalised financial products, and improve customer service.

2.2.3 Reinforcement Machine Learning

Reinforcement learning is a type of machine learning where an agent learns to make decisions by interacting with an environment Sarker (2021). The agent receives feedback in the form of rewards or penalties based on its actions Sah (2020). The goal is for the agent to learn a policy that maximises the cumulative reward over time. Reinforcement learning is commonly used in scenarios such as game playing, robotics, and autonomous systems. Key components of reinforcement learning include the agent, environment, actions, rewards, and exploration-exploitation trade-off Polydoros and Nalpantidis (2017). Popular algorithms in reinforcement learning include Q-learning, deep Q-networks (DQN), policy gradients, and actor-critic methods.

The process of reinforcement learning involves the following key components Sah (2020):

- **Agent**: The entity that learns and makes decisions within the environment. It is shown in Figure 8 how an agent performs actions based on the scanned environment. These actions hope to result in an optimal reward.

- **Environment**: The external system with which the agent interacts and from which it receives feedback. In a game-playing scenario, the environment would be the game itself.
- **Actions**: The set of possible moves or decisions that the agent can make within the environment. For instance, in a game, the actions could be moving in a particular direction or taking a specific strategy.
- **Rewards**: The feedback signal that the agent receives from the environment after taking an action. Rewards can be positive, negative, or zero and are used to reinforce or discourage certain behaviours.
- **Policy**: The strategy or decision-making process that the agent uses to determine its actions based on the current state of the environment.

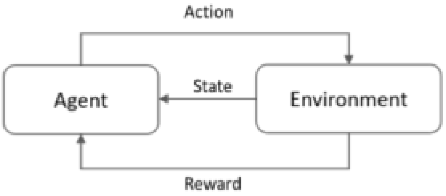


Fig. 7: Reinforcement Machine Learning: the Agent Performs Actions in an Environment to Optimise Rewards Sah (2020)

Reinforcement learning can be applied in banking for portfolio management. By treating the selection of financial assets as a sequential decision-making process, reinforcement learning algorithms can learn to optimise investment portfolios over time. These algorithms continuously adapt to changing market conditions and new information, aiming to maximise returns while managing risk. Lia and Huang (2023)Nakayama and Sawaki (2023)

2.3 Explainable Artificial Intelligence

The interpretation of AI is something that has been discussed quite a long time a go already. In Swartout and Moore (1993), the first architectures for the capturing of needed explanation knowledge and then the creation of ‘’powerful explanation generators”. For those paying attention, this can be seen as the first development of an XAI model. The research managed to create models where rules were explained to clarify the derived decision. Even though they did not know the terminology of yet, the added value was clear: “the explanations they offer are richer and more coherent, they are better adapted to the user’s needs and knowledge, and the explanation facilities can offer clarifying explanations to correct misunderstandings” Swartout and Moore (1993).

The terminology Explainable Artificial Intelligence was formulated much later. Lent, Fisher, and Mancuso (2004) described the need for an explanation generation system which explains the internal processes of training system used by the U.S. Army. Here, they label the Artificial Intelligence that explains the internal choices in the military simulation as Explainable AI. This is one of the first researches where the now known terminology is used. The definition of it however, still varies quite a bit. The following section will focus on the formulation of a widely excepted terminology.

2.3.1 Explainable AI Terminology

Within literature, a lot of studies use different terminologies for Explainable AI because the way of explaining it can be dependent on the models, scope and specificity of technology. As stated by Minh, Wang, and Nguyen (2021), there was a need for a need for a comprehensive standard definition which covers the characteristics. The reason for the prevention of such a unified definition is mostly to blame to the increase in different researches trying to define it. As a result, a lot of studies try to define a term related to transparency and trust instead of a standard concept. After many researches, the overall definition from multiple studies boils down to the following: Explainable AI refers to the set of techniques, methodologies, and approaches aimed at enhancing the transparency, interpretability, and accountability of artificial intelligence systems Arrieta and Dıaz-Rodrıguez (2019). It focuses on making AI systems more understandable to end-users, helping them to comprehend, trust, and manage these systems effectively Saeed and Omlin (2022). XAI enables AI systems to provide understandable explanations for their decisions and actions, especially in complex models often viewed as ”black boxes” Hase and Bansal (2020). By offering insights into how AI systems arrive at conclusions, XAI enhances trust, accountability, and usability in various applications Hase and Bansal (2020).

As mentioned here above, the definition is different due to the broad landscape of explainability. Due to the many applications and therefor differences in use, the audience is incredibly important to take into consideration. To further showcase the broad landscape, another study done by Minh et al. (2021), shows that the definition has been defined by two well-known institutions who both had a large influence within the domain of Explainable AI.

The first institution is the Defense Advanced Research Projects Agency (DARPA). This institution states that the goal of the XAI program is to develop a set of machine learning techniques that will allow human users to comprehend, properly trust, and effectively manage the new breed of artificially intelligent partners D.A.R.P.A. (2020). These techniques should also produce more explainable models while maintaining a high level of learning performance (prediction accuracy) D.A.R.P.A. (2020). This definition is further visualised below in figure 8 Gunning and Aha (2019). Here, it becomes clear how XAI can help end-users to clarify how certain decisions were derived from the data. It presents a clear problem description within the upper part of the figure. This practical example shows how many questions can appear from only receiving a result, while having the explanation how the result came to be solves these questions.

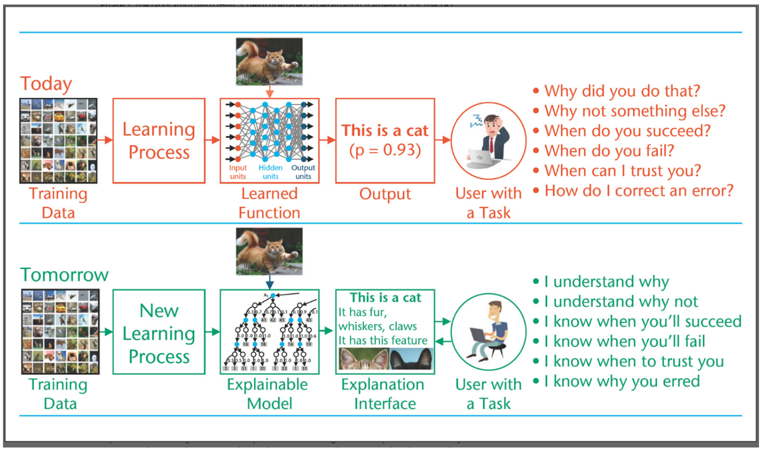


Fig. 8: Explanation of Explainable AI visualised Gunning and Aha (2019)

However, another well-known institution within the domain of machine learning is the Association for Computing Machinery (ACM) A.C.M. (2024). At their Conference on Fairness Accountability, and Transparency (FAT) they stated the following: Explainable Artificial Intelligence is the study of transparency and explainability for sociotechnical systems Hildebrandt and Castillo (2020). This further emphasises the large landscape in which XAI is present. With multiple definitions provided by many different studies, a overall and widely terminology can be formed.

2.3.2 Explainable AI Characteristics

As shown in figure 8, XAI can answer a lot of questions regarding the processes that lead to a certain output. With the increase of the complexity of model, transparency is often needed to clarify decision making. However, it is important to note that this is not always the case. Two studies provided cases where explainability was not the one of the requirements for a model. Freitas (2014) shows applications where users question the use of explainability and mostly care about the performance of the predictability. In other words, they don't care for the reasoning, only the result. Bunt, Lount, and Lauzon (2012) mentions that cost is also an important factor to take into account because, for systems who thrive for the lowest possible costs, it often prefers to save the costs saved on explainability as it outperforms potential benefits. However, the above mentioned scenarios are scenario specific and don't speak for all systems or situations. There is indeed a lot of potential for XAI and research has shown some of the most important goals and motivations for it. The following seven desired properties and outcomes of XAI have been formulated: Correctness, Accountability, Justification, Fairness, User acceptance, Discovery of new knowledge and Safety.

Improving Correctness

Explainability helps in understanding why a model makes specific predictions, enabling stakeholders to assess the reliability of its outputs and take appropriate actions. Knowing the reasoning behind a model helps to evaluate if the model is correct and does what it is supposed to do. Explanations can used for many different evaluation factors which determine the overall quality of a model. When a system is intelligible, it can represent to its users what it knows, how it knows it, and what it is doing about it Belotti (2004). This understanding allows users to identify when the system may be acting incorrectly and provides them with the opportunity to intervene and correct the system's actions. This validation of the model's predictions is essential for building trust and ensuring that the model is making decisions based on correct and fair reasoning Abdul, Vermeulen, and Wang (2018). The validation can be used for multiple factors of the model. Think of its safety, robustness and reliability for example Doshi-Velez and Kim (2017)Adadi and Berrada (2018). Additionally, explainability supports the iterative improvement of models. As users interact with the model and receive explanations for its decisions, they can provide feedback that can be used to refine and improve the model's performance Lipton (2016). This feedback loop is crucial for correcting inaccuracies and enhancing the model's correctness over time.

Overall, the correctness and therefor the quality of a model is of high importance. The above mentioned arguments for the involvement of explainability in improving the correctness of a model can be summarised: validating, feedback iterations which lead to improving and overall decision support.

Increasing Accountability

The accountability of models has gained significant importance in the context of data-driven technologies due to the implementation of stringent regulations like the General Data Protection Regulation (GDPR), the proposed AI Act and the right to be forgotten Villaronga and Kieseberg (2018). These regulations are designed to ensure that AI systems operate ethically, transparently, and responsibly, especially when they impact individuals' rights, freedoms, and privacy. Accountability for models refers to the ability to determine whether a decision made by an AI system was in accordance with procedural and substantive standards, and to hold someone responsible if those standards are not met Kortz and Doshi-Velez (2016). This is crucial in the context of potential harms that may arise from algorithmic decisions, as it allows for redress and the safeguarding of public interest Lepri and Oliver (2017).

Explainability serves as a critical enabler in achieving accountability in AI models. When algorithms can explain their decisions, it becomes easier to identify and correct biases, ensuring that the systems are fair and accountable Abdul et al. (2018). In addition, it also facilitates compliance with regulations that mandate transparency and accountability in automated decision-making Porayska-Pomsta and Rajendran (2019). Explainability also supports the enforcement of user accountability by providing mechanisms that make users aware of the system's actions that impact others and by ensuring that users are visible and accountable for their own actions Belotti (2004).

Justification

As mentioned in section 2.3.2,increased regulation has pushed models and organisations to increase reasoning of models. By requiring organisations to justify their AI-driven decisions and be accountable for their outcomes, these regulations aim to enhance transparency, fairness, and trust in AI technologies Belotti (2004). Fairness and trust will be talked about later but the justification of choices for a model help to enable this. Justification for models involves providing reasons or rationale behind the predictions or decisions made by these models Lepri and Oliver (2017). Explainable AI helps organisations to justify these models by providing clear and understandable explanations of how a model reaches its conclusions or predictions, explainability helps stakeholders - including data subjects, regulators, and developers - comprehend the decision-making process of AI systems Kortz and Doshi-Velez (2016).

Fairness

It has already been mentioned as the result of the characteristics above, but assessing fairness in a model is one of the most common mentioned benefits when looking at the increase of explainability in a model. The fairness of a model refers to the absence of any prejudice or favouritism toward an individual or a group based on irrelevant characteristics, which is crucial in decision-making processes, especially in high-stakes domains like healthcare and finance Chari, Seneviratne, and Gruen (2020). It is important because unfair models can lead to discrimination and inequality, potentially causing harm to individuals and society Arrieta and Dıaz-Rodrıguez (2019).

Explainability can help promote fairness in several ways. It can assist in identifying and mitigating decision biases Rosenfeld and Richardson (2019), ensuring fair decision-making Liao, Gruen, and Miller (2020). To further focus on this potential threat of bias in models, there has been a lot of research trying to find and tackle biasness in models. While these efforts have been made to measure and mitigate bias, many studies focus on result-oriented bias, neglecting bias in the decision-making procedure Zhao and Wang (2022). This limitation hinders the ability to fully debias a model. Explainable machine learning provides insights into the decision-making process, helping identify procedure-based bias. By bridging fairness and explainability, a novel perspective of procedure-oriented fairness based on explanations has been introduced Balagopalan and Zhang (2022).

Again, XAI enables the measurement of explanation quality gaps between different groups, leading to the development of optimization objectives to mitigate procedure-based bias. Comprehensive Fairness Algorithms (CFA) Zhao and Wang (2022) have been proposed to simultaneously improve traditional fairness, ensure explanation fairness, and maintain utility performance. Lastly, as also mentioned in the part for accountability and justification, explainability can improve a system's regulation and policy goals. These goals often include fairness as a one the key objective Testart, Fruchter, and Gilpin (2018).

User Acceptance

A lot of the above mentioned desired characteristics of models have an influence on the usage of them by end-users. In a lot of cases, models are used, monitored, altered or validated by humans. It is in the human nature that people are tentative to use something they don't understand. The same goes for models. It has even lead to the creation of models that describe how users embrace new technologies: the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT) Menant and Gilibert (2021).

User acceptance can be defined as the degree to which users are willing to employ and rely on a system for the tasks it is designed to support Hoffman, Johnson, and Bradshaw (2013). It is of high importance because it will determine how much a system can be implemented to enhance the end-user's workflow. In other words, even though your system is as good as it gets, it might never be utilised, misused or be underutilised if there is bad user acceptance. A possible result of this can be a failure or inefficiencies Hoffman et al. (2013). In addition, the success of a system on the long run (or the sustainability of it) is affected by the usage of users and their acceptance as well

Lim (2011). As mentioned in 8, XAI helps end users solve appearing questions which stimulates their acceptance.

The importance is shown but the measurement of it should be highlighted as well. Studies have shown multiple methods factors which define user acceptance: the quality of explanations, the actual satisfaction of users with the provided explanations, how well the user understands the AI systems and the measurement of the user's trust in the system Mueller and Hoffman (2018)Weld and Bansal (2019). XAI influences the acceptance of users due to providing clarity and precision in explanations, which can lead to a functional understanding of the AI system.

Discovery of new Knowledge

As a result of the characteristics and desired results of the so far mentioned XAI features, new discoveries are made leading to new knowledge. This knowledge is gathered due to XAI providing insights into how the model makes decisions. By understanding the reasons behind a model's predictions, researchers can uncover hidden patterns, relationships, or biases in the data that were previously unknown Freitas (2014). This transparency can help identify areas for further investigation or refinement of the model, leading to new discoveries or improvements in existing knowledge.

An example of such a unexpected relationship discovery was shown in Lipton (2016). Here, the study focuses on different features in a model regarding malfunctions. It delved into features that were focused on damage control in the case of a malfunction and features that were focused on both functioning and malfunction. Within the model, the difference in these features is important for repairing attributes and the discovery of a new mechanism. As it turned out, explanations about the underlying mechanisms between both resulted in unexpected relationships in the data Lipton (2016).

Furthermore, Miller (2017) discusses how Explainable AI can help researchers validate existing theories or generate new hypotheses by providing interpretable insights into complex models. By examining the factors influencing a model's predictions, researchers can test hypotheses and explore new avenues of research based on the model's explanation.

Safety

The safety of models and AI in general, is of course highly overlapping with the characteristics mentioned so far. When models are correct, justified, fair, accountable and accepted by users, they are highly likely to be safe. At least, it is likelier compared to models that don't have these characteristics. Safety for models and AI refers to the measures and approaches taken to ensure that AI systems operate correctly, reliably, and without causing harm to users or the environment Butin and Markova (2021). By understanding the model's reasoning and being able to know how the decision-making process went (like for example figure 8 again), it is possible to create a detector which alert when the model is likely to fail. This creates safe fail strategies where the system can degrade or notify the user to take control Mohseni and Pitale (2019). It becomes possible to anticipate and prevent problems, even those not previously encountered Butin and Markova (2021). An example of a practical use case of transparent XAI models in healthcare is shown by Benrimoh, Israel, and Fratila (2021). These models are explainable, so it becomes easier to identify and correct errors, thus reducing the risk of harm due to incorrect predictions or decisions.

Overall Terminology

The characteristics mentioned in section 1.5 and 2.3 can be summarised to create a complete overview of the terminology for Explainable Artificial Intelligence. In section 1, each of the characteristics can be divided into three types: a main concept, an outcome or a property. For the main concepts, it is meant that this characteristic is one of the core characteristics on which the other characteristics are built on. XAI is responsible for certain characteristics that the model delivers (outcomes) or that the model has on its own (properties).

Table 1: Terminology Explainable AI

Characteristics	Type	Explanation	Supportive literature
Interpretability	Main concept	Not only deliver accurate predictions but also present them in an understandable way. It involves the capacity of a model to provide interpretations in terms that are understandable to a human.	Coma-Puig and Carmona (2021), Arrieta and Dıaz-Rodrıguez (2019), Qin and Liu (2022), Minh et al. (2021), Hase and Bansal (2020), Gunning and Aha (2019), Freitas (2014)
Transparency	Main concept	The model is understandable by itself.	Arrieta and Dıaz-Rodrıguez (2019), Lent et al. (2004), Minh et al. (2021), D.A.R.P.A. (2020), Gunning and Aha (2019)
Improving Model Correctness	Property	Explainability improves the correctness and quality of a model through validating, feedback iterations, and overall decision support.	Belotti (2004), Abdul et al. (2018), Doshi-Velez and Kim (2017), Adadi and Berrada (2018), Lipton (2016)
Accountability	Property	The ability to determine whether a decision made by an AI system was in accordance with procedural and substantive standards, and to hold someone responsible if those standards are not met.	Kortz and Doshi-Velez (2016), Lepri and Oliver (2017), Abdul et al. (2018), Porayska-Pomsta and Rajendran (2019), Belotti (2004)
Justification	Property	Providing reasons or rationale behind the predictions or decisions made by these models.	Belotti (2004), Lepri and Oliver (2017), Kortz and Doshi-Velez (2016)
Fairness	Property	The absence of any prejudice or favoritism toward an individual or a group based on irrelevant characteristics.	Chari et al. (2020), Arrieta and Dıaz-Rodrıguez (2019), Rosenfeld and Richardson (2019), Liao et al. (2020), Zhao and Wang (2022), Balagopalan and Zhang (2022), Testart et al. (2018)
User acceptance	Outcome	The degree to which users are willing to employ and rely on a system for the tasks it is designed to support.	Menant and Gilibert (2021), Hoffman et al. (2013), Lim (2011), Mueller and Hoffman (2018), Weld and Bansal (2019)
Discovery of new knowledge	Outcome	By understanding the reasons behind a model's predictions, researchers can uncover hidden patterns, relationships, or biases in the data that were previously unknown.	Freitas (2014), Lipton (2016), Miller (2017)
Safety	Outcome	The measures and approaches taken to ensure that AI systems operate correctly, reliably, and without causing harm to users or the environment.	Butin and Markova (2021), Benrimoh et al. (2021), Mohseni and Pitale (2019)

2.4 Concept-based Explainable AI (C-XAI)

In addition to the above mentioned terminology for XAI, there is also a new type of XAI rising up in the last couple of years: Concept-based Explainable AI (C-XAI). It can best be defined as a subgroup of XAI which focuses on enhancing transparency by incorporating human-understandable concepts into AI decision-making processes Dreyer and Achtiba (2023). C-XAI models are often called Concept Bottleneck Models (CBMs) which map inputs onto a set of interpretable concepts, known as "the bottleneck," to make predictions Koh and Nguyen (2020). To clarify, these concepts help to identify known features. It has been used in image recognition where an input picture can be split up into known concepts (wheels, doors and headlights for a car for example A. Ghorbani and Wexler (2019)) which together help to construct a final output stating the concept of the input picture Sawada and Nakamura (2022). However, CBMs have practical limitations as they require dense concept annotations in the training data to learn the bottleneck effectively Dreyer and Achtiba (2023)Poeta and Ciravegna (2023). Due to its literature in the field of image recognition and the required concept annotations in the training data, this will not be further in the scope of this research.

2.5 Limitations and Challenges of XAI

Apart from the characteristics of XAI which have a positive influence on models, there are also downsides to XAI. One significant concern revolves around data privacy and security, as XAI often requires access to sensitive data to explain AI decisions, raising potential privacy breaches and security vulnerabilities Bruijn, Warnier, and Janssen (2021). Shokri, Stronati, and Song (2017) showed privacy leakage risks that appear when XAI provides explanations for certain model decision boundaries. Based on the explanation of certain predictions, private information could be retrieved which could have significant consequences when it were to fell in the wrong hands.

The complexity of AI models also poses a significant challenge for explainability Mueller and Hoffman (2018). Even though it is the goal of XAI to make very complex models more explainable, the continuous evolution and sophistication of AI models Deloitte (2022) make them challenging to interpret, meaning there need to be ongoing advancements in XAI systems to keep pace with these complexities Yang, Wei, and Wei (2023). An example of this was the case for Z. Ghorbani and Kazemi (2022) where the risk of Deep Learning explanations were shown. Two identical pictures with minuscule changes were given different explanations.

Another challenge mentioned by Bruijn et al. (2021) is the lack of expertise to fully understand the actual explanation that is provided by XAI. Arrieta and Dıaz-Rodrıguez (2019) also stresses the need for explanations that are understandable by many different stakeholders so no mistakes are made and everything is compliant with Commission (2019).

It has been briefly mentioned in section 1.5, but the trade-off between explainability and performance is a challenge for XAI as well. The curve in figure 3 shows different ML models and on which side of the spectrum they lie in term of explainability and accuracy Arrieta and Dıaz-Rodrıguez (2019). The performance of models can decrease when they also need to be explainable M. Ribeiro and Singh (2016). The existence of hidden patterns in some domains suggests that complex, opaque models are sometimes necessary to achieve good performance and explainability can hinder this Crook and Schluter (2023). However, an influential paper by Rudin (2019) suggests that explainable models can match or outperform opaque models with "black-box" characteristics in most real-world scenarios. In other words, there is a possibility that this challenge is no longer challenge moving 10 years from now.

Table 2: Challenges and Limitations of XAI

Challenge/Limitations	Description	Supporting Literature
Data privacy and security	XAI explanations may lead to privacy breaches and security vulnerabilities.	Bruijn et al. (2021), Shokri et al. (2017), Saraswat and Bhattacharya (2022)
Model complexity	XAI models need to keep up with the complex developments of AI models.	Mueller and Hoffman (2018), Deloitte (2022), Yang et al. (2023), Z. Ghorbani and Kazemi (2022)
Lack of expertise	The lack of expertise to fully understand the provided explanation especially as many stakeholders are involved.	Bruijn et al. (2021), Arrieta and Dıaz-Rodrıguez (2019), Hulsen (2023)
Explainability-Performance trade-off	The choice between opaque models which are less understandable but have a higher performance and simpler models that are more explainable.	Arrieta and Dıaz-Rodrıguez (2019), M. Ribeiro and Singh (2016), Crook and Schluter (2023)

2.6 Explainable AI in Fraud Detection

During the literature review, multiple papers have been found that make use of XAI to explain ML models that help against fraud. The most important findings have been summarised.

Cirqueira, Nedbal, Helfert, and Bezbradica (2020) introduced a scenario approach aimed at understanding the needs of fraud experts, focusing on qualitative and quantitative research to delve into their opinions and recommendations for user-centric explanations. With the use of SHAP and expert interviews, researchers have identified thirteen cognitive tasks import for fraud experts when analysing cases. Quantitative results about the importance of each task has been left out from the research however, making it more of a qualitative research than a statistical analysis.

Rao et al. (2021) developed xFraud, an explainable fraud transaction prediction framework tested on eBay transaction records. The framework focuses on a detector and an explainer to explain the reasoning behind the detector. The framework uses LIME as a hybrid explainer for task-aware measures from general neural networks (GNN) to explain the complexity of the created transaction graphs. xFraud represents a significant contribution to the field of fraud detection by addressing critical challenges in the domain and providing a scalable and explainable solution that is validated through extensive experiments on real-world transaction networks. Although the framework is extensive, it is missing a qualitative evaluation to make sure xFraud meets the requirements and that the values shown are correct.

Psychoula et al. (2021) evaluates the performance of various ML models using precision, recall, F1-score, and AUC curve, considering the highly imbalanced nature of fraud datasets. The research stresses the importance of explainability in sensitive domains like fraud detection. The GDPR is cited as a regulatory framework that emphasizes the need for transparency in automated decision-making. In terms of practical implications, the study suggests that SHAP could be used for regulatory compliance and retrospective model accuracy examination.

2.7 Validation of the techniques used

As seen within the discussed literature, the current gap is mostly present in the combination between both quantitative and qualitative validation for models. As shown in 1.7, the fifth phase of CRISP-DM is evaluating the created models. This can be done both quantitatively and qualitatively. In order to fully answer the main research question this is done in both ways.

2.7.1 Quantitative Validation

Quantifying the performance of the used ML model is an essential part of this research as shown in section 1.7. A distinction can be made between the validation of the ML model that predicts the potential of fraud, as shown in figure 2, and the validation of the XAI techniques that are used to do this research. In section 2.6, research has shown that well substantiated ML validation metrics are the precision, recall, AUC and F1-score of the model Psychoula et al. (2021). These metrics show that the performance of ML models is often measured in terms of the accuracy of the prediction that has been made. First, precision, recall and the F1 socre are explained

Precision measures the proportion of true positive results in the set of all instances classified as positive Powers (2008). It is calculated as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

Recall measures the proportion of true positive results in the set of all instances that should have been classified as positive. It is also known as sensitivity Powers (2008). It is calculated as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

The F1 score is the mean of precision and recall, providing a balance between themPowers (2008). It is particularly useful when the class distribution is uneven. The F1 score is calculated as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

AUC and ROC

To extend the evaluation of classification models beyond precision, recall, and the F1 score, the Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC) are used as well. These metrics are useful for evaluating the performance of binary classification models across different threshold settings.

The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied Fawcett (2006). The curve is created by plotting the True Positive Rate (TPR, also known as recall) against the False Positive Rate (FPR) at various threshold settings. By plotting TPR against FPR for various threshold values, the ROC curve provides a view of the trade-off between true positive rate and false positive rate.

The Area Under the ROC Curve (AUC) quantifies the overall ability of the model to discriminate between positive and negative instances across all threshold values. The AUC ranges from 0 to 1, where:

- An AUC of 1 indicates a perfect model that correctly classifies all positive and negative instances.
- An AUC of 0.5 suggests a model with no discriminative ability, equivalent to random guessing.
- An AUC less than 0.5 indicates a model performing worse than random guessing, which is a rare situation in practice.

The AUC is useful because it is independent of the classification threshold and provides a single measure of a model's performance across all possible classification thresholds. This makes it an excellent metric for comparing different ML models. A higher AUC indicates a model with a better performance at distinguishing between the positive and negative classes.

Train Test split

Train-test split is a fundamental technique in machine learning used for evaluating the performance of a predictive model. It involves dividing the dataset into two distinct subsets: the training set and the testing set. The training set is used to train the model, while the testing set is reserved for evaluating the model's performance on unseen data. This method provides an initial understanding of how well the model generalizes to new, independent data Tibshirani and Friedman (2008).

The train-test split method typically involves the following steps:

- **Data Partitioning:** The dataset is randomly partitioned into two subsets. A common practice is to allocate 70-80% of the data to the training set and the remaining 20-30% to the testing set. The exact split ratio can vary depending on the size of the dataset and the specific requirements of the problem Tibshirani and Friedman (2008).
- **Model Training:** The model is trained using the training set, where it learns the underlying patterns and relationships within the data. During this phase, various model parameters and hyperparameters are adjusted to optimize performance.
- **Model Evaluation:** Once the model is trained, it is tested on the testing set. This evaluation provides an unbiased estimate of the model's performance since the testing data was not used during the training phase.

The train-test split helps in the prevention of overfitting. By reserving a portion of the data for testing, the train-test split helps in detecting overfitting, where the model performs well on the training data but poorly on unseen data. This ensures that the model has not simply memorized the training data but has learned to generalize from it. In addition, it provides a clear indication of how the model will perform in real-world scenarios. By evaluating the model on the testing set, researchers and practitioners can estimate its accuracy and robustness on new, independent data.

K-Fold Cross validation

Cross validation is a robust statistical method employed in model evaluation to assess the generalizability and performance of predictive models Tibshirani and Friedman (2008). It is widely recognized for its effectiveness in mitigating overfitting and ensuring that the model performs well on unseen data. K-Fold cross validation is the most commonly used method where the dataset is randomly partitioned into k equal-sized folds. One fold is retained as the validation set, and the remaining k-1 folds are used for training the model. This process is repeated k times, with each fold used exactly once as the validation set. The k results are then averaged to produce a single estimation. This method helps in providing a more accurate measure of model performance as it utilizes the entire dataset for both training and validation purposes Tibshirani and Friedman (2008).

2.7.2 Qualitative Validation: Expert Opinion

For the qualitative validation, an expert opinion is needed. One of the academic contributions of this research is a confirmation of the correctness of the obtained results from experts in the field of Fraud Detection to see if the explanations are aligned with their knowledge and expertise. This expert opinion is formed by conducting interviews with Data Science, Rule Writing and Alert Handling. The construction of the interviews is done with the main concepts, outcomes and properties of XAI in mind, as described in table 1. Each of these characteristics are implemented in the construction of department specific questions. In addition to the department specific questions, several qualitative metrics and factors are considered. These can be broadly categorised into application-grounded metrics and human-grounded metricsZhou, Gandomi, Chen, and Holzinger (2021). The construction of the qualitative validation can be found in section 5, and the results in section 6.

3 Explainable AI Taxonomy

An overall terminology of XAI has been researched and clarified. However, as shown in the literature review, the term for explainable models can still be very broad. In other words, determining the right XAI technique to tackle the present core problem is still not clear. Therefor, a complete taxonomy for Explainable AI is created. Here, distinctions between models based on characteristics are made, leading to a clear overview of all possibilities.

Contents

3.1 Model Concept Distinctions

An extensive literature review has been performed and, based on this review, a taxonomy for Explainable AI has been constructed van der Pol (2024). Based on Speith (2022), Arrieta and Dıaz-Rodrıguez (2019), Korelc (2020), Rawal and McCoy (2022), Minh et al. (2021), Linardatos and Papastefanopoulos (2020) and Chou and Moreira (2021), an understanding of the different distinctions has been formed. The final taxonomy can be seen in figure 11. Here the taxonomy has been divided into Model Concepts, Model Types, Explainability Categories, Explainability Principles and Techniques Examples. Following this taxonomy, the correct technique for each given situation and combined required output can be formulated. For clarification, the first big distinctions that can be made are shown below in figure 9. In literature, there is a clear difference between transparent and opaque models. They can be distinguished by their way of explaining which will be delved into further below. These two types of models are seen as different model concepts. In addition, Opaque models have another classification difference between Model-Agnostics and Model-Specific explanations.

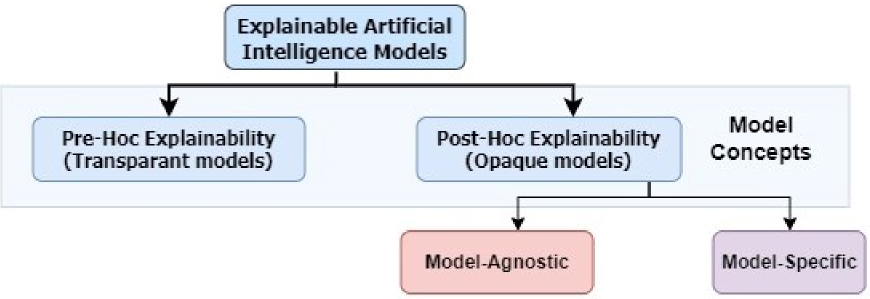

Explainable Artificial Intelligence Models
Pre-Hoc Explainability (Transparant models)
Post-Hoc Explainability (Opaque models)
Model Concepts
Model-Agnostic
Model-Specific

Fig. 9: First Classification in Explainable AI Taxonomy

3.1.1 Pre-Hoc vs Post-Hoc Explainability models

The first classification distinction is between Pre-hoc and Post-hoc explainability models. The difference between both types of models have to do with the timing, model complexity, dependence on surrogate models and computational burden Rawal and McCoy (2022) Chou and Moreira (2021) Acun and Nasraoui (2023). The definition of both types is explained by their name. Post-hoc explainability refers to methods that aim to explain the decisions of AI models after the model has already made a prediction. In other words, after the model has already been trained Chou and Moreira (2021). These methods are often applied to opaque, or "black-box," models, which are complex and not inherently interpretable. Pre-hoc explainability, on the other hand, involves designing models that are inherently interpretable and transparent from the outset Minh et al. (2021). These models are constructed in such a way that their decision-making process is understandable to humans without the need for additional explanation methods. To finalize, opaque and transparent models can be classified based on:

- **Timing**: Post-hoc explainability is applied after the model has made a prediction, while pre-hoc explainability is integrated into the model from the beginning Rawal and McCoy (2022).
- **Model Complexity**: Post-hoc methods are often used with complex models that are not inherently interpretable, whereas pre-hoc methods involve designing simpler, transparent models Chou and Moreira (2021).
- **Dependence on Surrogate Models**: Post-hoc explanations may rely on separate models to approximate the original model's decisions, which can be different from the actual model being explained Acun and Nasraoui (2023).
- **Computational Burden**: Post-hoc methods can add computational overhead because they require additional processing after the model's prediction Acun and Nasraoui (2023).

3.1.2 Post-Hoc Explainability: Model-Agnostic vs Model-Specific models

Post-Hoc explainability can be further classified into Model-agnostic and Model-specific explanations. The names of both explanations explain their characteristics. To clarify, Model-agnostic approaches are designed without the need to understand or access the internal workings of the models they are applied to, making them highly versatile and adaptable Ai and Narayanan.R (2021). Model-specific methods, in contrast, are tailored to the particularities of a given model or a class of models Linardatos and Papastefanopoulos (2020). They leverage knowledge about the model's internal structure to provide explanations, optimisations, or adaptations that are closely aligned with the model's functioning Ai and Narayanan.R (2021). An example of their biggest difference is the customisation level of the explanation. Model-agnostic approaches offer broad applicability across different models at the cost of potentially less precise or efficient solutions, while model-specific methods provide tailored solutions that can leverage the unique features of a specific model for better performance Minh et al. (2021). In addition, it was already mostly shown in the definition, but Model-agnostic methods are more adaptable due to their ease of integration Ai and Narayanan.R (2021).

3.2 Different Model Types

The classification in section3.1.1 results in different model types based on their characteristics as shown in figure 11. For the sake of the scope of this research, each model type is briefly described to get an understanding what is meant with the term model type and the differences each model has.
Neural Networks are computational models inspired by the human brain's structure and function. They consist of layers of interconnected nodes or "neurons" that can learn complex patterns through training Linardatos and Papastefanopoulos (2020). They are particularly effective for tasks like image and speech recognition.
Support Vector Machines are supervised learning models used for classification and regression tasks. They work by finding the hyperplane that best separates different classes in the feature space, maximising the margin between the closest points of the classes, which are called support vectors.
Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is known for its high accuracy, robustness,

and ease of use.
Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It predicts a continuous output Arrieta and Dıaz-Rodrıguez (2019). Logistic Regression is similar to linear regression but is used for binary classification problems. It models the probability that a given input belongs to a particular category using a logistic function.
K-Nearest Neighbours is a simple, non-parametric algorithm used for classification and regression. It classifies data points based on the majority vote of their neighbours, with the object being assigned to the class most common among its k nearest neighbours Bau and Gilpin (2019).
Bayesian models are statistical methods that apply Bayes' theorem for prediction and inference. They are useful in estimating the probability of outcomes, incorporating prior knowledge along with new evidence Sah (2020).
Rule-Based Systems are a type of artificial intelligence that use a set of "if-then" rules to derive conclusions or make decisions. They are straightforward and easy to understand but can become complex as the number of rules grows.
Decision Trees are a non-parametric supervised learning method used for classification and regression. The model predicts the value of a target variable by learning simple decision rules inferred from the data features Arrieta and Dıaz-Rodrıguez (2019). It is represented as a tree structure, with branches representing decision paths and leaves representing outcomes.
Generalised Additive Models are a flexible generalisation of linear models that allow the linear predictor to depend linearly on unknown smooth functions of some predictor variables. They are used to model nonlinear relationships in a data-driven way.

3.3 Explainability Categories

Even though Post-hoc explanations can be further classified in model agnostic and model specific explanations, they both have similar explainability categories. These categories are seen as an overall term to give an idea how these models or concepts are actually explained. Explanations, as shown in figure 11, can be further classified in the following categories: local explanation, explanation by simplification, feature relevance explanation and visual explanation.

3.3.1 Feature Relevance

Feature relevance explanation refers to the assessment of how important individual features are to the predictions made by a machine learning model. This provides insight to the contribution and influence of individual features by ranking and explaining and showing them Linardatos and Papastefanopoulos (2020). However, it is not to be mistaken with the explanation principle feature importance. While feature importance is about quantifying the influence of each feature on the model's output, feature relevance is more about understanding and explaining the contribution of each feature to the model's predictions Arrieta and Dıaz-Rodrıguez (2019). Feature importance is typically represented by scores or rankings, whereas feature relevance often involves decomposition methods and is closely tied to visualisation techniques for interpretability Arrieta and Dıaz-Rodrıguez (2019).

An example where feature relevance has played a crucial role is in Hayashi and Takano (2020). This research aimed to achieve transparency and conciseness in credit scoring. Their approach allowed for the extraction of rules from credit scoring datasets, which are characterised by similar attributes. By identifying the most relevant features and translating them into understandable rules, the study enhanced the interpretability of credit scoring models. This has proven to help the accuracy-interpretability dilemma in deep learning like mentioned before in figure 11, but also supports financial institutions in making more transparent credit decisions.

Another use case is Tritscher and Wolf (2023). Here, it highlighted the importance of identifying key features that contribute to the detection of malicious activities within banking networks. They managed to focus on traffic features and patterns that are indicative of intrusion attempts. By identifying these features, banks can improve their cybersecurity measures and protect their customer data and financial assets. This application of feature relevance not only improves the explainability of intrusion detection models but also allows for the optimisation of security protocols by focusing on the most significant indicators of threats.

3.3.2 Visual Explanations

It has been mentioned in section 3.3.1, but feature relevance can also be seen as a form of visual explanations. As shown in figure 11, certain explainability principles can derive from both visual and feature relevance explanations. Results shown in this figure shows how the importance of each feature is shown visually in order to improve understanding. Visual methods use graphical representations to explain information about the model's decisions Linardatos and Papastefanopoulos (2020). These visual explanations can show the features which influenced the predictions of the model, therefor being closely related to feature relevance. Heatmaps, saliency maps, and other visualisation techniques can illustrate which parts of an input are most influential in a model's prediction Speith (2022). Model-specific visual explanations can be deeply integrated with the model, such as using layer activations in neural networks to generate saliency maps. Arrieta and Dıaz-Rodrıguez (2019) mentions that visual explanations are less commonly used for model agnostic explanations because of its wide applicability. To clarify, the visual explanations needs to be created from inputs and outputs which can be hard given the complexity of opaque models.

3.3.3 Explanation by Simplification

Explanation by simplification in Explainable AI involves simplifying complex models into more interpretable forms, focusing on providing clear and concise explanations for model predictions Arrieta and Dıaz-Rodrıguez (2019). As one of the main research components of this research delves around the interpretability of a model, making explanations human understandable, explanation by simplification is very relevant. One of the main capabilities of this approach is to make the explanations also understandable for people without technical backgrounds Chromik, Eiband, Buchner, Krüger, and Butz (2021). This is also one of the main challenges for this category because, due to complexity of some ML models, it can be hard for a model to be flexible enough to explain it easily and accurately at the same time Korelc (2020).

3.3.4 Local Explanations

Local explanation is a critical aspect that focuses on providing explanations for individual predictions made by AI models Minh et al. (2021). In other words, it describes the decision-making process behind specific instances. These explanations are tailored to clarify why a particular prediction was made, enabling users to understand the model's reasoning at a granular level. Despite their advantages, local explanation methods come with challenges such as parameter dependency, sampling variability, and difficulties in comparing results across different techniques Le, Prihatno, Oktian, Kang, and Kim (2023). An example of a technique used within the category local explanations is anchors. This technique translates the model's decision into rule based explanations making it easier for the user to understand M.T. Ribeiro, Singh, and Guestrin (2018). These rules try to show the most essential features Korelc (2020).

global vs local

The scope of the local explanations category, as the name implies, is set at a local level. Previous categories in this chapter could differ in this scope. Some techniques used in these categories could either be at a local or global level. It is important to make a clear distinction here. As mentioned, local explanations are tailored to clarify the decision-making process behind a particular prediction. Global explanations aim to provide an overarching understanding of the model behavior as a whole, offering insights into how the model functions across its entire dataset or domain. Global explanations provide a broader perspective on the overall behavior and functioning of the AI model, while local explanations delve into the specifics of individual predictions, highlighting the distinction between understanding individual instances and comprehending the model's behavior at a larger scale Chromik et al. (2021). An example of a global explanation would be: which functions are important for the model because they contribute to minimising costs which is the goal of the model. An example of a local explanation would be: Stijn should not receive money because his income is lower this month than usual.

3.4 Explainability Principles

The taxonomy should serve as a complete overview of Explainable AI. Therefore, explainability principles are created to further explain the classification Minh et al. (2021). It concretizes each explainability category which makes it clearer to see the practical process behind it. As shown in figure 11, an example of an explainability principle is the clarification of visual explanations as described in section 3.3.2. There are many ways a model can visually explain something. For model-specific explanations, this differs for each model type. Neural Networks and Support Vector Machines can both apply a sensitivity analysis to visualise their explanation. To clarify, a sensitivity analysis changes the input and analyses the output changes after which it ranks the input based on these changes Cortez and Embrechts (2011). It measures the effect on the output when inputs are varied through their range of values, gradient and variance. Other visualisation principles are the visualisation of internal processes or showing a dependency plot. As mentioned in section 3.3.4 and 3.3.3, linear and rule based explanations are examples of local and simplification explanations. An addition to this is a decision tree or even multiple decision trees (random forest). These four principles are examples how a model can be made more explainable, interpretable and transparent.

3.5 Explainability Technique examples

As shown in figure 11, multiple techniques can be used to provide an explanation. It became clear in section 2.6 that the most commonly used techniques for explaining ML models used for fraud detection are SHAP, LIME and Anchors. Given the scope of this research, these three techniques will be further explained instead of possible explainability techniques.

3.5.1 SHAP

The SHAP (SHapley Additive exPlanations) technique stands out for its ability to offer insights into the contribution of each feature to a model's prediction, enhancing transparency and interpretability. The SHAP framework is used to explain by attributing the impact of each feature on a specific prediction, allowing users to understand the reasoning behind model outputs at a granular level. As shown in figure 11, It is inspired by cooperative game theory, specifically the Shapley value method Lundberg and Lee (2017). Cooperative game theory is a branch of game theory that focuses on how players can cooperate to achieve a common goal and distribute the gains fairly among themselves. In cooperative games, players form coalitions and work together to maximize their collective payoff Merrick and Taly (2020). The Shapley value method, derived from cooperative game theory, is used in SHAP (SHapley Additive exPlanations) to attribute the impact of each feature to a model's prediction Hung, Xu, Wang, and Chen (2023). Properties used to calculate this prediction can be accuracy and missings Strumbelj and Kononenko (2010). The Shapley value (ϕ_i^{SH}) for player i in a cooperative game is calculated using the following formula Narahari (2012):

$$\phi_i^{SH} = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \tag{4}$$

where:

ϕ_i^{SH} = the Shapley value for player i
N = the set of all players in the game
S = a coalition of players that does not include player i
$v(S)$ = the value of coalition S
$|S|$ = the cardinality (number of elements) of set S

The formula computes the weighted average of the marginal contribution of player, or in the case of XAI feature, i to all possible coalitions in the game or model. For each subset S of features excluding i, the difference between the value of the coalition with feature i ($v(S \cup \{i\})$) and without feature i ($v(S)$) is calculated. This difference is then weighted by the number of ways to form coalition S and feature i out of all feature, relative to the total number of possible coalitions Narahari (2012). This way, the importance of a feature can be calculated.

One key benefit of SHAP is its ability to quantify the contribution of individual features to model predictions, providing a clear understanding of each feature's impact on the outcomes derived from data Y. Wang, Cheng, and Liu (2019). Additionally, SHAP allows for contrasting explanations Molnar (2023). This means that instead of comparing a prediction to the average of the entire model, it can compare it with subsets or specific data points of the model. This is something a local technique such as LIME or Anchor can not do. As stated above, SHAP is based on the cooperative game theory which is considered a very strong theoretical foundation Molnar (2023). However, one notable disadvantage is its computational complexity, especially when dealing with a large number of featuresMerrick and Taly (2020). This complexity can hinder practical implementations, requiring significant computational resources. Additionally, SHAP ignores possible feature dependenceBertossi, Kimelfeld, Livshits, and Monet (2023). Lastly, SHAP is no prediction model like Anchor or LIME. It shows feature importance but can't predict how much a prediction would change if the input changes Molnar (2023).

3.5.2 LIME

The name perfectly explains the taxonomy of this technique: Local Interpretable Model-agnostic (LIME) explanations, created during the research of M. Ribeiro and Singh (2016). As shown in figure 11, LIME is both a technique that can be classified as the Local and simplification explanation category Minh et al. (2021). The reason for this is that LIME creates linear models or decision trees Molnar (2023) which make the complex ML models easier to interpretable M. Ribeiro and Singh (2016). It works by altering the input data and observing how the model's predictions change, allowing it to generate explanations that are understandable to humans Aldughayfiq, Ashfaq, Jhanjhi, and Humayun (2023). In other words, they simplify the model locally. As the definition of model agnostic implies, LIME can be integrated with any black box ML model and data formats, such as text, tabular, and image data. One of the differences between LIME and SHAP is the approach of the explanation. As shown in section 3.5.1, SHAP focuses on the marginal contribution of each feature which is derived from the Shapley value. Lime uses linear models for the output and uses this to examine the processes of the model locally.

To best explain how LIME captures the local importance of features, a step by step approach is created Ferdib-Al-Islam et al. (2023).

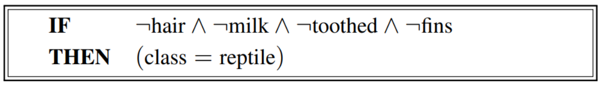
- Choose a specific data point for which you want to explain the model's prediction.
- Create alterations around the selected data point by randomly perturbing its features while keeping the target feature constant.
- Obtain predictions from the machine learning model for each altered data point.
- Fit an interpretable model, such as linear regression, to the perturbed data points and their corresponding predictions. This model approximates the behavior of the complex model locally.
- Calculate the importance of each feature based on how much the predictions change when the feature is perturbed. Features with a significant impact on the prediction are considered important.
- Use the interpretable model to generate explanations by highlighting the features that had the most influence on the model's prediction for the selected data point.
- Provide a local interpretation of why the model made a specific prediction for that particular data point, making the decision-making process more transparent and understandable to users.

An advantage of LIME over SHAP is its significantly faster when getting explanations from single instances Psychoula et al. (2021). However, the paper questions if it is at the cost of reliability. It is also easy to use and very interpretable but because it is done locally, it may not capture the global trends Y and Challa (2023). Another very important disadvantage of LIME is the scope of the neighbourhood around the local data point you're going to compare it with Molnar

(2023). The width of this neighbourhood is not preselected and it is up to the user to see if the explanations make sense or not. This is step two in the step by step approach. The determining of this neighbourhood can lead to manipulation and the covering of biasness by the researchers Slack, Hilgard, Jia, Singh, and Lakkaraju (2020).

3.5.3 Anchors

After LIME, a successor named Anchors was created two years later by the same researchers M.T. Ribeiro et al. (2018). As the taxonomy describes, Anchors is a local model-agnostic technique which is based on the ruled-based learner principle. What this means is that Anchors works by creating simple if-then rules that capture the behavior of the model for a particular prediction. It determines an IF - THEN decision rule which will be anchor a prediction, meaning the alterations of other features does not change the prediction Molnar (2023). It does this by altering the features to see if this prediction changes. This is a form of reinforcement learning, discussed in section 2.2.3, which constantly tries to improve until it has found the final decision rule Demajo, Vella, and Dingli (2020). These rules are designed to be easily understandable by humans, making it easier to trust and verify the model's outputs. By focusing on specific features that significantly impact the model's decisions, Anchors provide a clear explanation of why a certain prediction was made. Figure 10 shows a basic example of an anchored IF - THEN decision rule created in Anchor which is based on a animal classification model Ignatiev (2020).



IF $\neg$hair $\wedge$ $\neg$milk $\wedge$ $\neg$toothed $\wedge$ $\neg$fins
THEN (class = reptile)

Fig. 10: A very basic example of an anchored decision rule

Just like LIME, Anchors' biggest advantage is that it very interpretable with easy to understand decision rules Ignatiev (2020). Anchor computes very efficiently and fast and offers subsettable explanations which can help to fully understand the model Molnar (2023). However, also just like LIME, the determining of the features for the local prediction can be very domain-specific which can be a threat for the prediction itself Ignatiev (2020). In addition, some form of discretisation is required for predictions done by Anchor, due to the possibility of too specific results, the risk of overfitting is present Molnar (2023).

Table 3: Overview of Advantages and Disadvantages of the used XAI techniques

XAI Technique	Advantages	Disadvantages	Supporting Literature
SHAP	Quantifying individual feature contribution. Contrastive explanations. Strong Theoretical foundation	Computational heavy. Feature Dependency.	Lundberg and Lee (2017), Y. Wang et al. (2019), Molnar (2023), Merrick and Taly (2020), Bertossi et al. (2023)
LIME	Computationally fast. Easy to use. Very interpretable	Missed global trends. Neighbourhood scope. Manipulation allowance	Psychoula et al. (2021), Y and Challa (2023), Molnar (2023), Slack et al. (2020)
Anchor	Easy to understand. Efficient. Subsettable explanations	Domain-specificity. Discretisation needed	Ignatiev (2020), Molnar (2023)

3.6 The final XAI Taxonomy

Each part of the taxonomy has been described in this chapter in order to get a good understanding of the different techniques, categories and principles. The final overview and therefore complete taxonomy is created in figure 11.

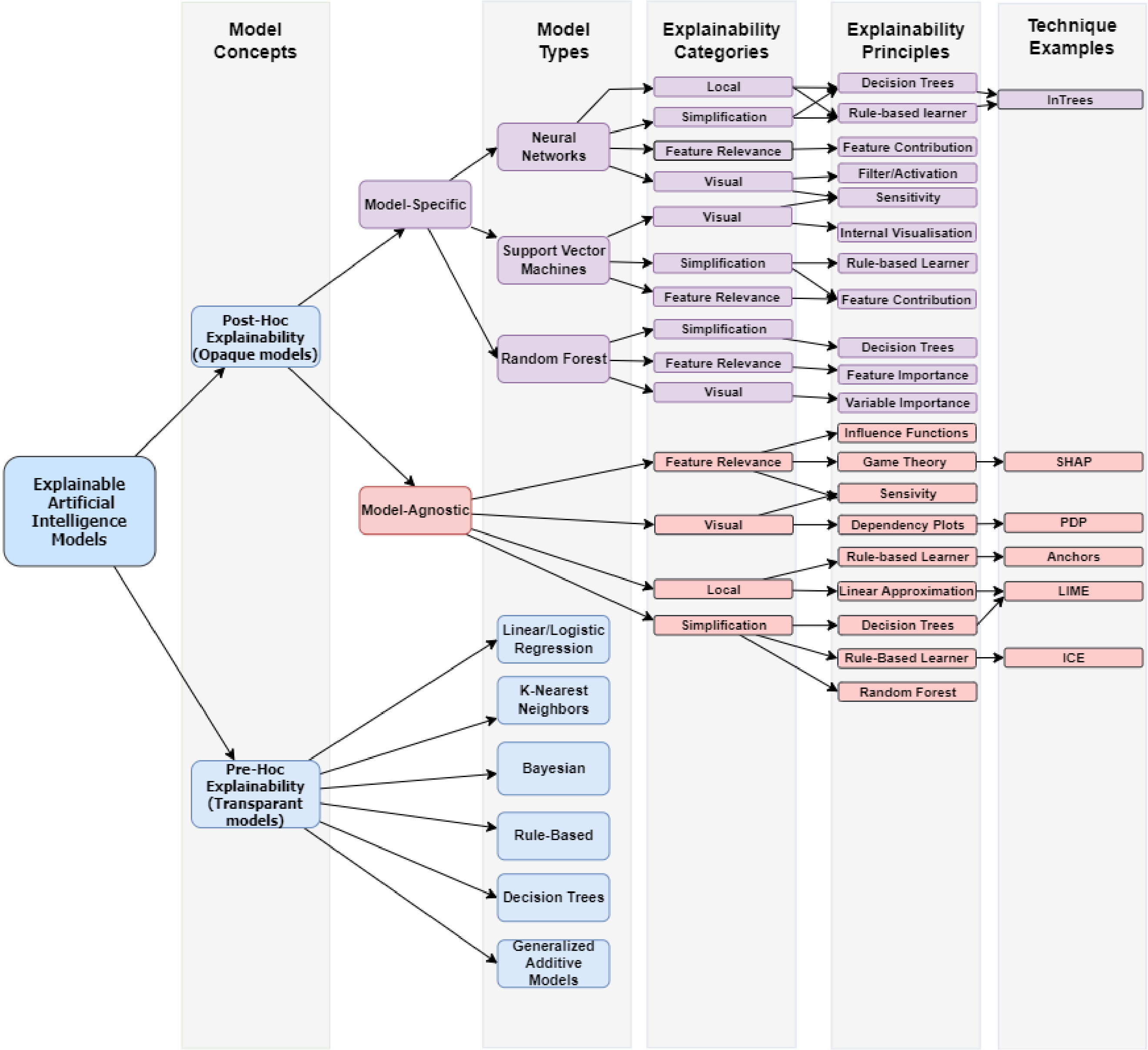


Fig. 11: Full Taxonomy of Explainable AI

4 The XAI Model for Transactional Fraud Detection

Within this section, the creation of the XAI model has been split into a step wise approach. This approach involves a number of CRISP-DM steps: Data Understanding, Data Preparation, Modelling and also a part of the Evaluation phase. Each step has been explained in section 1.7. This section starts by understanding the dataset that has been provided by ING. The content of this section will focus on the characteristics of the dataset and how it is processed in order to fully address the research question. Next, different ML classifiers with varying datasets are tested and their performance is evaluated. A XAI technique is chosen such that the requirements of this use case are aligned with the XAI model characteristics, as set up in the taxonomy in figure 11. A last, the results of the XAI model are shown and evaluated.

Contents

4.1 The Dataset of the XAI Model

The dataset provided by ING consists of multiple e-banking transactions which have been processed by the bank. The dataset has been anonymised completely to prevent the doom scenario of confidential and personal data leaking. The content of the dataset consisted of 50 anonymous features (e.g. feature 1 and feature 39) as columns, shown in appendix A Each row is an individual transaction which has a value for each feature. The exact content for the determining of the features will not be discussed in this research. However, the dataset itself is describable as a binary classification in which the result of the features determine whether a transaction is considered fraud (it has been labelled with a 1 in the ”label” column) or correct (it has been labelled with a 0).

One of the important characteristics about this dataset to take into account when processing it, is the fact that it is highly imbalanced. From a society perspective, this is a good thing. Meaning, there is a significantly lower number of fraud cases than correct transactions. Important to know is that these fraud transactions have been blocked in real life and are therefor no real transactions anymore. For the research however, this imbalance in the dataset causes the dataset to become very big in size when considering a sufficient number of fraud cases.

The total size of the dataset consists of 75 million rows which are all individual transactions. There are a total of 50 different features that are shown in the dataset as different columns making it a high dimensional dataset. Each row for these features is a float between 0 and 1 except for two features, shown in appendix A. Because the provided dataset is the result of calculations in another dataset, a lot of preprossessing has already been done. Scaling is not needed due to all values having a similar maximum and minimum. All relevant values are a float so there is no need to encode for example categorical data. Last, there are no missing values. Again, because the dataset is the result of processing another dataset, there are no values missing. It has been discovered in section 6 that missing or incomplete information in the original dataset is indicated in this dataset with the help of a fixed value. To clarify with an example, if some characteristics of a transaction were to be missing, the features give this a value through the written rules in the model.

Sample Ratios

Throughout the rest of this section, datasets are indicated with either a 1:100 or 1:1000 ratio. This shows the ratio of the taken sample in relation to the original dataset. To clarify, for the 1:100

sample, a total of 750000 transactions are taken from the original 75 million transactions, meaning the sample has a 1:100 ratio in relation with the original dataset. For all samples, the total number of fraud cases from the original dataset are taken. This is done to see if the used ML classifiers and the XAI model would perform better if it trains on a less imbalanced dataset. For both the 1:100 and 1:1000 ratio sample, different variations in the dataset are made to see a potential difference in performance. In this research, a dataset containing only the top 10 most important features and a dataset having no heavy correlating features are created.

False Positives dataset

In addition to the total dataset described in this section, another dataset has been provided. This contained the transaction id's of transactions that were originally flagged as fraudulent, but turned out to be falsely accused in the end. To clarify, these transactions are shown as not fraudulent in the original dataset because the final result was a non fraudulent verdict. These are not the false positives of the ML classifiers indicated by their confusion matrix shown in section 4.3. Instead, these are false positives that were documented by fraud experts within ING. For each of these transactions, an investigation has been initialised by Alert Handling, which allows for the addition of information outside of the scope of the XAI model. This false positive information is used for the construction of two of the four test cases, as documented in section 5.2.3.

The reason why these transactions are specifically targeted within this research, is because of the potential clarification XAI explanations can provide. Knowing the impact of each of the involved features for these specific examples, and analysing a potential identifiable pattern allows for an increase in the concepts, properties and outcomes of XAI models, documented in table 1. A potential increase in justification or fairness are of high importance for ING. However, in the case of false positives specific, the discovery of new knowledge is especially important.

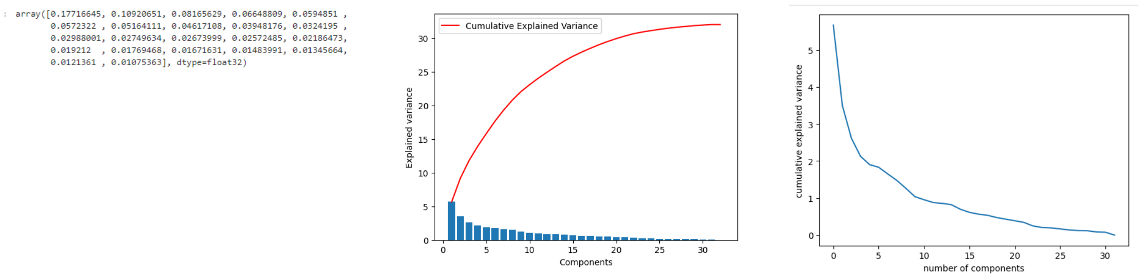
4.2 Data Transformation

The dataset is the result of calculations on another dataset. This means that a lot of the data is already processed and transformed. However, there were still some steps to be taken in the this phase. One of the steps within processing the dataset consisted of checking for highly correlated features. As shown in figure A3, there are some features that have a high correlation. Looking at the heatmap and the histograms, it becomes clear that features 11, 12, 13, 14, 15, 16, 17, 18 and 19 are heavily correlated with each other. In addition, feature 5, 22, 30, 32, 33, 34, 43 and 44 are also correlated with other features in the dataset. To rule out the influence of correlating features on the explanations, a dataframe with all features and a dataframe with the high correlating features removed have been taken into account for this research. In addition, a dataframe containing the top 10 most contributing features is created and analysed as well. However, this dataframe was created after the first iteration of the XAI model. This is due to the current obscurity of the most important features.

Principal Component Analysis for Feature Selection

A Principal Component Analysis (PCA) transforms a dataset with many features into a new set called principal components. These principal components are linear combinations of the original variables and are ordered in such a way that the first principal component captures the maximum possible variance in the data, the second principal component captures the maximum remaining variance, and so on Jolliffe and Cadima (2016). The goal is to reduce the number of dimensions without losing significant information. Given the high dimensionality of this dataset due to high number of features, a reduction can be helpful and therefor, a PCA has been done within this research.

The first step of a PCA is to apply a scaler in order to reduce the difference in values between the features within the dataset. This would be helpful when one feature's values are ranged in the millions while the other's are values between one and zero. However, in the provided dataset almost all values are already between 0 and 1. Still, a standard scaler is applied for the dataset. The results for the explained variance in each of the created principle components are mapped in figure 12. In the array left of the figure, each value represents the captured percentage of variance in the principle component. In this case, the highest captured variance in a principal component is 17,72% which is not so much. The second highest variance capturing principal component captures 10,92%. Mapping this array in the graphs, it becomes clear that the principal components do not capture too much variance. Each blue bar in the middle chart is a principal component and the red line shows the total cumulatively explained variance with the number of principal components. The total explainable variance is 32. Given that the first principal component explains 17,72%, it is mapped as the value 5,67.The graph in the right shows that there are a significant number of principal components needed to capture the most varaince in this dataset. Therefor, the dataset is taken is a whole for this research.



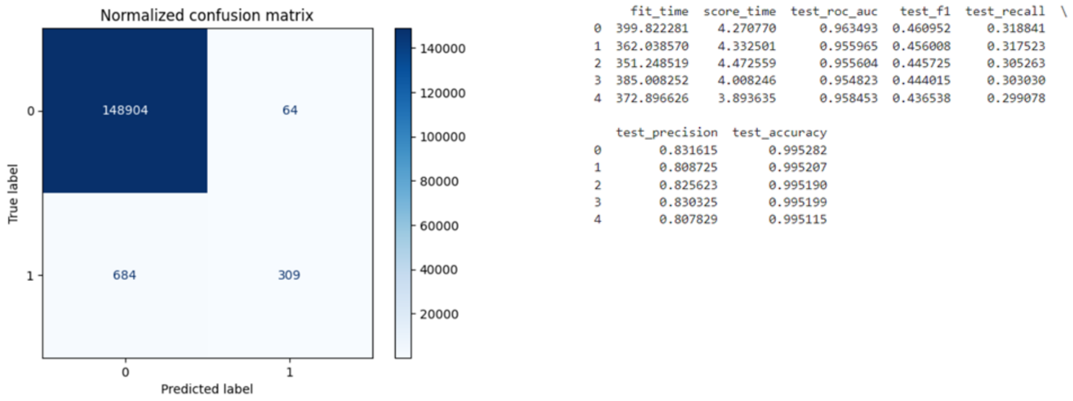
```
: array([0.17716645, 0.10920651, 0.08165629, 0.06648809, 0.0594851 ,
         0.0572322 , 0.05164111, 0.04617108, 0.03948176, 0.0324195 ,
         0.02988001, 0.02749634, 0.02673999, 0.02572485, 0.02186473,
         0.019212  , 0.01769468, 0.01671631, 0.01483991, 0.01345664,
         0.0121361 , 0.01075363], dtype=float32)
```

Fig. 12: PCA Results

4.3 ML Classifiers

Upon receiving the dataset, ING noted that it was of no use for them to test whether or not Random Forest was the best classifier to implement. Apart from standard decision ruling models, Random Forest (RF) is the current classifier being used by the more advanced models. This has been validated by data scientists before constructing the models. However, to improve objectivity and test whether or not the results could potentially have been improved, other classifiers are tested as well. The classifiers being used and validated in this are RF, Support Vector Machine (SVM) and Gradient Boosting (GB). All classifiers have been cross-validated with a total of 5 k-folds.

As mentioned in section 4.1, samples with a ratio of 1:100 and 1:000 when comparing with the complete dataset are taken. An example of the cross validated results for the 1:00 RF classifier is shown in 13. Here the confusion matrix, AUC, f1-score, recall, precision, accuracy and computing time are shown. For the train test split, a ratio of 80/20 is used which is widely accepted as a sufficient ratio Joseph (2022). The split ratio provides enough train data to sufficiently train the model while also preventing the possibility of overfitting. It is important to stress the number of fraudulent transactions in the 1:100 sample in relation to the total number of fraudulent transactions. From the 750.000 transactions in this sample, around 4500 are fraudulent which accumulates to 0.6%. The imbalance also shows in the confusion matrix. Because the number of fraudulent transactions is the same for the 1:1000 sample, this dataset is less imbalanced. Given that the dataset is focused on fraud detection, the number of false negatives should be as low as possible. It is a threat for fraud experts when the model classifies transactions as not fraudulent while they are. As documented in section 2.7.1, the metric to determine the score in terms of false negatives is recall. Comparing the average recall and auc for all datasets and classifiers, table 4 is formed. All individual cross validation results and confusion matrix such as figure 13, can be found in appendix C.



```
     fit_time  score_time  test_roc_auc   test_f1  test_recall  \
0  399.822281    4.270770      0.963493  0.460952     0.318841
1  362.038570    4.332501      0.955965  0.456008     0.317523
2  351.248519    4.472559      0.955604  0.445725     0.305263
3  385.008252    4.008246      0.954823  0.444015     0.303030
4  372.896626    3.893635      0.958453  0.436538     0.299078

   test_precision  test_accuracy
0        0.831615       0.995282
1        0.808725       0.995207
2        0.825623       0.995190
3        0.830325       0.995199
4        0.807829       0.995115
```

Fig. 13: 1 : 100 Random Forest Results

The cross-validation results indicate that for the 1:100 dataset, the GB classifier achieved the highest average recall (0.3405), while the RF classifier with non-correlating features and the standard RF both achieved the highest average AUC (0.956). For the 1:1000 dataset, GB also demonstrated superior performance with the highest average recall (0.7491) and AUC (0.980), followed closely by Random Forest classifiers, particularly those using non-correlating features. The SVM classifier showed a terrible performance for the 1:100 dataset. Figure C8 showed that it had 993 false negatives out of the 993 fraudulent cases, therefor having a recall of 0. In order to test the robustness of the classifier, a random dataset was created and tested. It's results are further discussed in section 4.4.1. As written in section 2.7.1, having an AUC score of around 0.5 means the classifier is scoring equivalent to random guessing which would be correct in this case. However, SVM has a similar score for the 1:100 sample, making it the worst performing classifier out of the three. It is important to mention that while GB performs slightly better than RF, it's computation time is a lot longer. This is clearly shown in the fit time column shown in appendix C. For this reason, and the recommendation by ING, RF is used for this XAI model.

Table 4: Results Cross Validation

Dataset and Classifier	**Average Recall**	**Average AUC**
1:100 Random Forest	0.3053	0.956
1:100 Gradient Boosting	0.3405	0.939
1:100 SVM	0	0.6091
1:100 Top 10 Features Random Forest	0.3373	0.942
1:100 Non Correlating Features Random Forest	0.334	0.956
1:1000 Random Forest	0.7021	0.976
1:1000 Gradient Boosting	0.7491	0.980
1:1000 SVM	0.6394	0.9561
1:1000 Top 10 Features Random Forest	0.7002	0.969
1:1000 Non Correlating Features Random Forest	0.7234	0.973
1:1000 Random Numbers Random Forest	0	0.4991

4.4 XAI Techniques

Within section 4.3, the conclusion is to implement the Random Forest ML classifier within the XAI model due to its reduction in computing time. Next, the XAI techniques discussed in 3.5 are analysed to test which is the most suited for the given problem. First, LIME has been tested due to it being computationally fast and its clear interpretability. The local explanations for a random transaction in the test dataset is provided in appendix B in figure B4. This explanation is similar to the local waterfall explanation documented in section 4.4.2. In the top left of the plot, the final probability score is given. The middle graph shows the contribution of each feature for this verdict. If a feature is pointed to the right, it means this feature pushed the prediction towards the fraudulent side. If to the left, to the not fraudulent side. The right side of the graph provides an additional overview of the feature contributions. In this case, the colour orange means the feature was pushed towards the fraudulent side.

One of the disadvantages LIME has, is the limitation to only provide local explanations. This is also the case for Anchors. SHAP does not have this limitation. Following the path in the taxonomy created in figure 11, SHAP fits the given problem significantly well. An XAI technique which shows both local and global feature relevance and is capable of doing this for different type of ML classifiers. The last characteristic is important to take into account when looking at potential model changes in the future. Currently, Random Forest is chosen to be the most suited ML classifier, but there is a possibility that this is no longer the case in the future. The explainability category feature relevance is of high importance for this XAI model, therefor SHAP is chosen as the XAI technique.

Within section 3.5.1, an explanation of SHAP, the advantages and disadvantages and the corresponding SHAP values are explained. However, in order to understand the results, a brief explanation how to read the plots is given. Given the anonymity of the dataset, some examples from literature are provided to make the plots less abstract. The reduction in abstraction makes it more clear how the value of features contribute to the prediction of the model, instead of showing only values between 0 and 1.

4.4.1 Global Explanations: SHAP Bar and Beeswarm Plots

Bar and beeswarm plots are ways to visualise global explanations. In this research, they show how each feature has contributed overall. In other words, when looking at the entire test dataset, it shows the impact of each individual feature.

An example of a use case where a bar plot is used to visualise explanations, is a use case from the official website of SHAP SHAP (2024). Here, the goal is to predict if a person earns more than 50.000 euros. It is a similar example to the dataset of this research in the sense that it is a binary classification. It is either a yes (1) or a no (0). The result in the form of a global bar plot is shown in 29. The features Age, Relationship and Capital Gain are resulted as the three most important. The values in grey on the y-axis are the average values for that feature. The red numbers on the right are the mean SHAP values of the feature. In other words, the feature age had a mean SHAP value of 0.86. This feature has a very high impact on the prediction considering the use case being a binary classification. This does not have to mean that age always pushes the prediction towards one side. It means that the feature age is of high influence on the final prediction, stating that the salary is either above or below 50k euros.

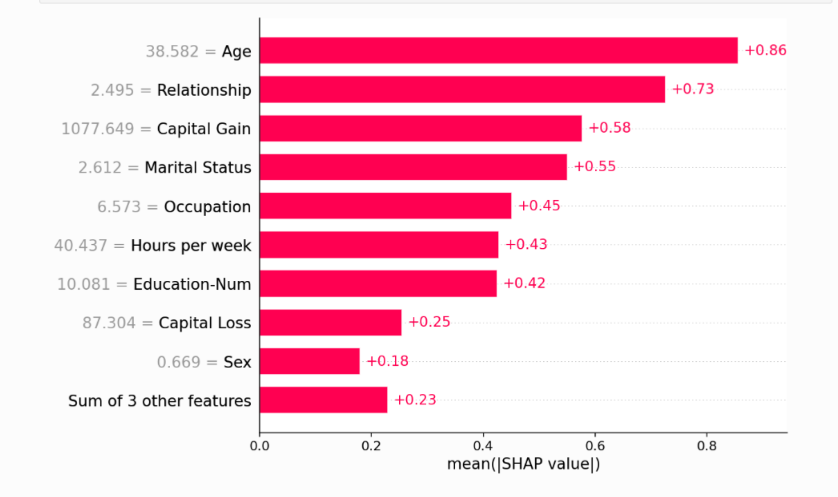


Fig. 14: An example of a global bar plot

Given the anonymity of the dataset and therefor the visualised information, the global bar plot for this research shown in figure 15 is more vague than figure 29. Again, the x-axis represents the mean absolute SHAP value for each feature, which quantifies the average contribution of each feature to the model's output. Higher mean SHAP values indicate greater importance. Figure 15 is the result from the 1:1000 sample dataset. It becomes clear that feature 41 has the highest mean SHAP value of around 0.041. It is the most important feature, followed by feature 42, 32 and so forth, down to feature 44. Even though feature 41 is the most important, 0.041 suggests it has a low average SHAP value. As shown in 4.3, the dataset is highly skewed towards a not fraudulent which is logical considering that the dataset are real historical transactions. Within this model, a prediction indicating a low chance of being fraudulent is close to zero. Therefor, it is logical that the average SHAP values of the shown features is very low.

A SHAP Beeswarm plot provides an information-dense summary of how the top features in a dataset influence the model's output. The x-axis represents the SHAP value, which quantifies the impact of a feature on the model's prediction. Positive SHAP values (towards the right) indicate that the feature contributes to a higher prediction, while negative SHAP values (towards the left) indicate that the feature contributes to a lower prediction. The y-axis displays the features in descending order of importance, similar to the bar plot shown in figure D16. Each data point (dot) represents a single transaction. The horizontal position of the dot corresponds to the SHAP value for that feature and instance. In addition, the colour of each dot represents the original value of the feature for that instance. By default, red indicates a high feature value, and blue indicates a low feature value.

Figures 15 and 16 show the global explanations for both the 1:100 and 1:1000 sample datasets. Comparing both samples shows a shift in the top 4 most important features with feature 41 and 42 being dethroned by feature 32 and 35. This has been discussed in the interviews in section 6, but can not be disclosed within the research without revealing the content of the features. In the beeswarm plot of both samples, it shows how, in general, the value of feature 41, 42, 7, 30 and 50 is positively correlated with the fraud prediction. In other words, a lower value typically results in a not fraudulent prediction while a higher value results in a fraud prediction. For feature 32 and 35, this correlation is more complex. For example, feature 32 shows that a low value results often in a low SHAP value with some exceptions where the value was high and the SHAP value was very low. The provided figures show the top 10 most important features, the complete overview of individual feature importance is shown in appendix D. In figure D17, feature 9 is noticeable for being flipped when compared to other features. Meaning, a low value results in a high SHAP value. This has been taken into account for the interviews and is discussed in section 6.

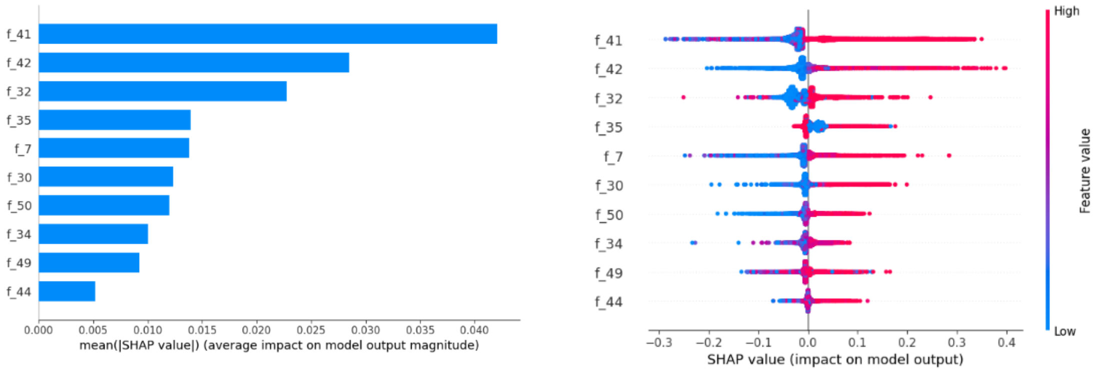


Fig. 15: 1:1000 Global Bar and Beeswarm Plot for Fraud Detection Features

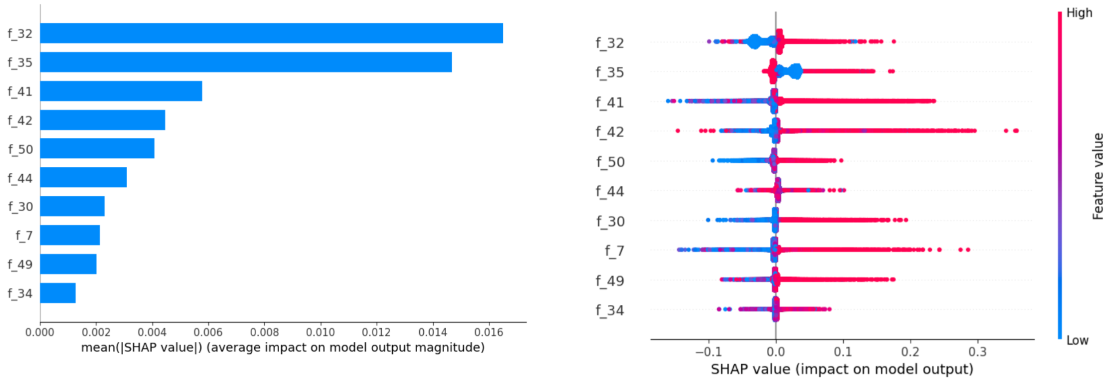


Fig. 16: 1:100 Global Bar and Beeswarm Plot for Fraud Detection Features

In addition, the global explanations for the false positive dataset, the top 10 features dataset and the dataset containing only non correlating features have been analysed. As shown in figure C13, the ranking or distribution of this dataset does not differ significantly when comparing it to the 1:100 and 1:1000 samples containing all features. This is also the case for the top 10 features dataset. As figure 17 shows for these transactions here, the higher the value the higher the SHAP value. Feature 41 and feature 42 show this especially, potentially indicating why the transaction was flagged in the first place. This also becomes more clear when looking at the bar plot in figure E40, where it shows that the mean SHAP value of the features is significantly higher, indicating less transactions with a low fraudulent prediction.

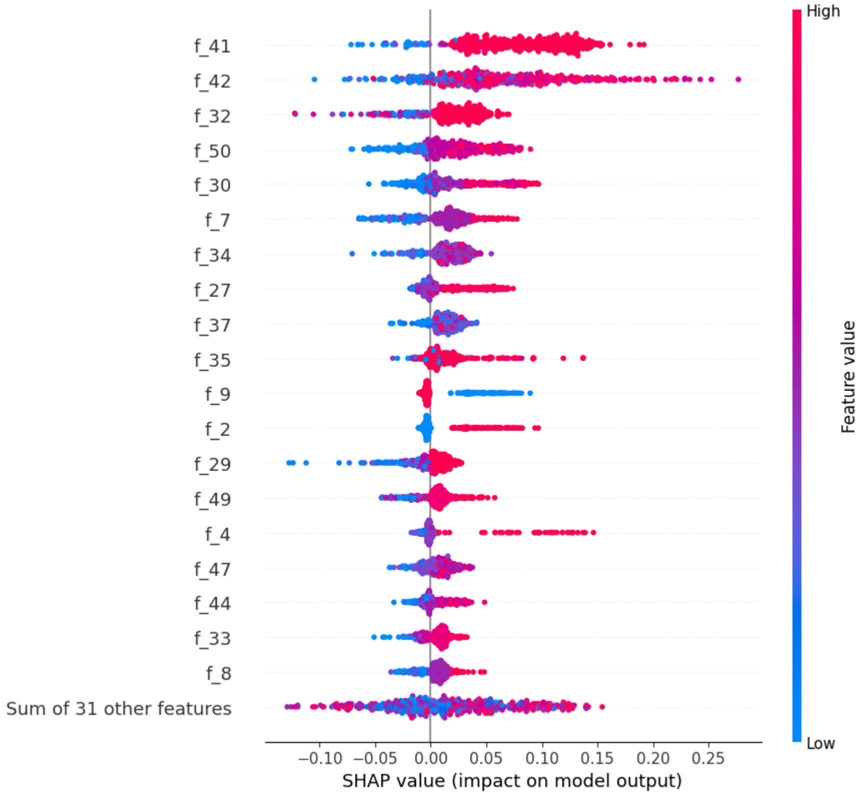


Fig. 17: False Positives SHAP Beeswarm Plot

Robustness Test

Within the quantitative validation for the explanations to make sure the results are actually feasible, a robustness test has been performed. The goal of this test is to see if the models give different output for different data. A complete new and random dataset where each feature consists of a float between 0 and 1, is done in python. The output is shown in appendix F. From the global beeswarm and bar plot in figure F41, it becomes clear that no real conclusion can be drawn from the plots. Each feature's importance and explanation is very similar. The absence of feasible and clear explanations becomes clear when running the model multiple times. Within this research, the model was run four times. As visible in F41, the importance of features changes each iteration and the explanation in the beeswarm plot doesn't change, there are only new features visible. Looking at the top 5 of all iterations, it shows there is no pattern.

4.4.2 Local Explanations: SHAP Waterfall plot

The waterfall plot provides a clear and detailed view of how each feature contributes to the final prediction for a specific instance. An example for this research is shown in 18. To clarify, the base value (E[f(X)]) is the average model output over the training dataset. It represents the expected value of the prediction if no features were known Lundberg and Lee (2017). Each bar in the waterfall plot represents the contribution of a specific feature to the final prediction. Positive contributions (features that push the prediction higher) are shown in one color (red), while negative contributions (features that push the prediction lower) are shown in another color (blue). The order of the features is based on the absolute individual impact. The reason for absolute impact becomes clear in figure 27 , where feature 42 is more important than feature 41 because the absolute impact of the feature is higher. The final prediction is the sum of the base value and all feature contributions. It is shown at the end of the waterfall plot.

This type of local explanation is the foundation for the test cases that are created in section 5.2.3. Figure 18 shows an example of a waterfall plot used in this research. The final score of the model is shown in the top right corner: the number 1 indicates that the model is very certain this transaction is fraudulent. In figure 18, it shows that the final outcome is the result multiple features resulting a potential fraudulent expectation instead of a small number of features leading the model. Meaning, the model takes all features into account to determine its final outcome.

The plot shown in 18 shows the real life problem for fraud experts. The model results in a score that indicates that there is no doubt in this transactions being fraudulent. However, this transaction is present in the false positive dataset, meaning it turned out to be a correct transaction. The essence of the problem is discussed during the interviews and results are documented in section 6. This specific situation of a false positive resulting in high model score is analysed and tested in Case A. The outcome of this case is documented in section 6.4.1.

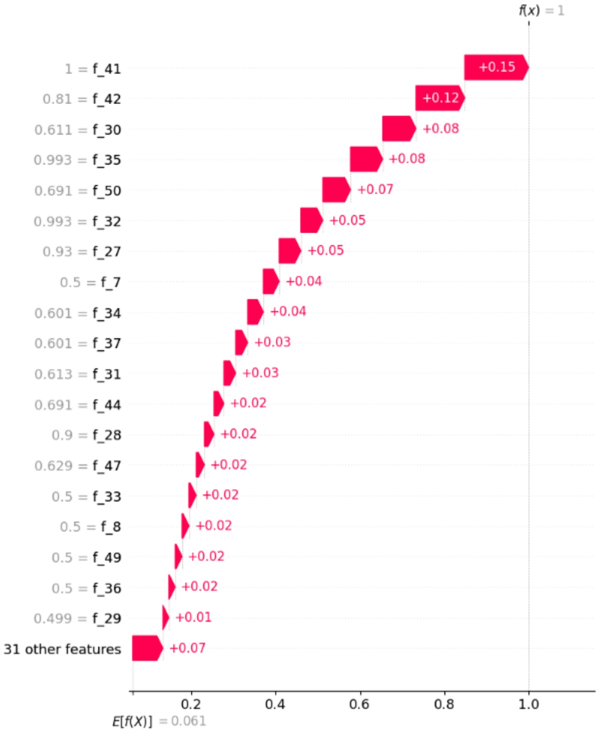


Fig. 18: Local Waterfall Plot for a False Positive Fraud Flag

4.5 XAI model results

The result of the XAI model, are values of data type explanation. This is an umbrella term for an array containing multiple sub arrays. The first sub array consists of the individual feature impact on the prediction of the model, these are the SHAP values. For this research, this means this sub array has 50 values. The second array is the base value of the model, which is explained in 3.3.4]. To calculate the final prediction of the model, the summation of the first sub array is added to the base value. The output of this calculation is a value between one and zero. In the last sub array, the array stores all the original feature values that are needed to plot the complete overview.

In other to get a good overview of the model results, the final model prediction for each transaction has been calculated and stored into a separate array. The correct label indicating if the transaction was indeed fraudulent or not is appended to this array to get the complete picture. Figures 19 and 20 show how the model performs if a threshold of 0.05 (figure 19) or 0.35 (figure 20) is set. Again, this value is the final predicted value of the XAI model for each transaction. The reason for the selected thresholds is because makes it possible to visualise the model's results. As mentioned before, the dataset of the model, is very imbalanced. For the 1:100 sample, the test dataset consists of 149.961 transactions of which 145.922 are given a prediction score lower than 0.05. Out of these 145.922 transactions, 149 are fraudulent. Meaning, the model predicted that these model had a low chance of being fraudulent while they were. With a total of 993 fraudulent transactions in the test set, this means that the model wrongfully predicted around 15% of actual fraud transactions to not be fraud, when taking a threshold of 0.05 into account. The results of Case D in section 6.4.4 show that this can have multiple reasons outside of the model's scope. However, as figure 19 show that a threshold of 0.05 would still give a lot of false positives and therefor a lot of unnecessary work for Alert Handling who process flagged transactions. A total of 4039 transactions are in set, of which 844 are fraudulent. In other words, around 26% of the transactions are fraudulent. Although this is a high score, it would mean that 3195 transactions also have to be processed. A significant finding here shows the decrease in false positives as the prediction score gets higher. In other words, as the prediction score gets higher, the probability of the model being correct increases. This is further showcased in figure 20 when the threshold is increased to 0.35. A total of 665 transactions are left of which 477 are fraud which is a total of 71.7%. Looking at the total of 993 fraud cases, the models captures 48% of them with a threshold of 0.35. To put this into perspective, the total dateset for a threshold of 0.35 encapsulates 0.44% of the total dataset while capturing 48% of fraud. These results have been tested for multiple datasets and have been added to tables 5, 6 and 7.

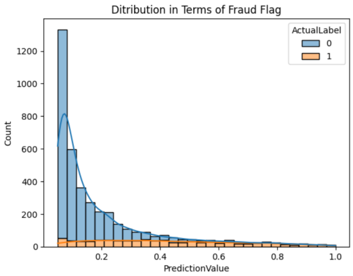


Fig. 19: 1:100 Model results for a threshold of 0.05

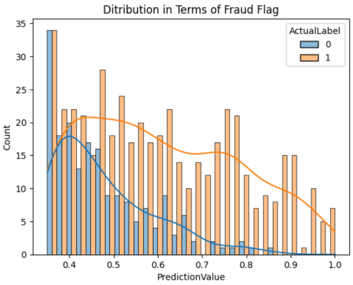


Fig. 20: 1:100 Model results for a threshold of 0.35

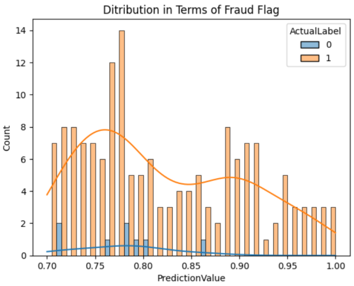


Fig. 21: 1:100 Model results for a threshold of 0.7

The results from tables 5, 6 and 7 form significant insights in the XAI model's performance for the different selected thresholds. To explain the tables, for each test, a different dataset was used. For each test, the total dataset size used in the test set is given, as well as the total number of fraud cases in that dataset. The remaining size of the dataset after the threshold specified in the table is given. This indicates how many transactions are left in the dataset and have therefor received a prediction equal or higher than the threshold. It is important to also show this as a percentage of the whole dataset to understand how many transactions were given a score lower than the threshold. Furthermore, the table shows the number and percentage of fraud cases in the threshold dataset. This can be seen as the true positives of this threshold dataset. Last, the percentage of true positives is calculated with regards to the total number of positives. In other words, the precision is calculated here. Tables 5, 6 and 7 serve as an overview for different dataset performance, but also as an overview for the trade-off between the percentage of fraud captured and the number of false positives that come with it.

Looking at the tables, the first noticeable finding is the performance of the "1:1000 Top 10 Features" and "1:1000 sample" tests. These two datasets perform significantly higher than the dataset variations containing the 1:100 sample. Especially when applying a threshold of 0.7, with a result of 54.5% and 57.5% still being captured while only having 3.5% and 3.9% of the original dataset remaining. The dataset for this threshold consists for 92.8% and 87.1% of fraud cases and therefor decreases the chances of a false positive.

Looking at the dataset variations containing the 1:100 sample, it is clear that the performance is very similar for all variations. The difference in percentage of fraud captured for the normal sample, the sample containing the top 10 features and the sample containing no correlating features is 0.3% for the 0.35 threshold and 0.6% for the 0.7 threshold. The sample containing the top 10 features under performs compared to the other two for the 0.05 threshold with a 5.2% and 5.9% difference. Important to mention for these samples is the reduction in dataset size for the 0.35 threshold and the percentage of true positives in this dataset. For this threshold, there is only 0.4%, 0.5% and 0.43% left of the original dataset while still capturing 48%, 47,7% and 47,7% of fraud. For all cases, more than 70% of the dataset consists of fraud cases. Given that this ratio for fraudulent and non fraudulent transactions is closer to the real life situation, this is an important finding.

Apart from the datasets maintaining a consistent ratio in size for the training and testing dataset, there have also been tests where this is imbalanced. For the "1:100 Train 1:1000 Test" test, the RF classifier has been trained on the 1:100 sample dataset and the explainer has been tested with the 1:1000 testing dataset. This is the other way for the "1:1000 Train 1:100 Test" test. Both tests perform better for all thresholds than the tests done before by having high percentages of fraud captured. One downside for the "1:1000 Train 1:100 Test" is the large number of false positives for the 0.05 and 0.35 threshold. For the 0.7 threshold, the number of true positives (66.9%) is significantly better than the other two thresholds, but under performs compared to the other tests.

The distribution for all tests with the different thresholds is shown in appendix 4.5. The decrease in false positives when increasing the threshold is more abstract when looking at the tables 5, 6 and 7, but clearly shows in these distribution figures. As shown in table 7, the "1:1000 Top 10 Features" test performed significantly better than the dataset variations consisting of the 1:100 samples. However, by plotting the distribution and comparing them, as done in figure 23 and 24, it shows the clear performance difference. The distribution in figure 24 is a lot smaller when looking at the y-axis. Figure 22 and 23 show clearly that, although both tests have similar fraud capturing performance according to table 7, the distribution shows an increase in fraud cases as the model score for the "1:1000 Top 10 Features" test increases. This further confirms the increase in probability of true positives as the model score increases.

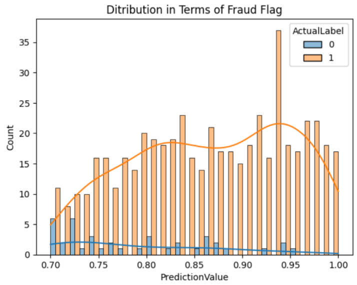


Fig. 22: 1:1000 Model results for a threshold of 0.7

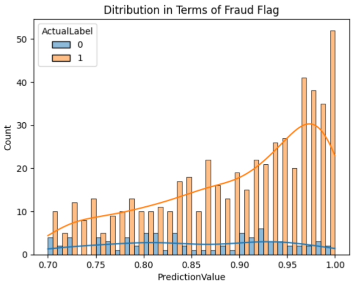


Fig. 23: 1:1000 Top 10 Features Model results for a threshold of 0.7

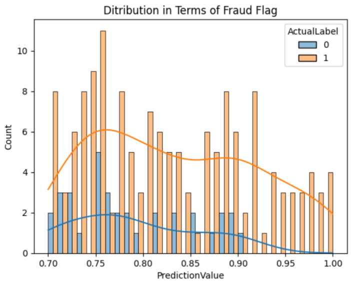


Fig. 24: 1:100 Top 10 Features Model results for a threshold of 0.7

Lastly, the false positives dataset is tested to see how the XAI model scores suspicious transactions. The total dataset consisted of 450 transactions and the results are shown in figure E40. Again, the dataset sees these transactions as non fraudulent because they turned out to be false positive. It shows that the model gave a high score to a significant portion of the dataset. As described in 6, the reason for these transactions to be a false positive can be outside the scope of model. This further confirms that the model should serve as an assisting tool and should not be leading.

Table 5: Results XAI Model with a 0.05 threshold

Dataset Used	Dataset Size	Fraud Dataset Size	Dataset Size at Threshold (Percentage of Total)	Fraud Captured (Percentage of Threshold Dataset)	Percentage of Fraud Captured (Precision)
1:100 Sample	149961	993	4039 (2.69%)	844 (26%)	85%
1:1000 Sample	15917	932	2496 (15.7%)	888 (35.6%)	95.3%
1:100 Top 10 Features	149870	951	3417 (2.3%)	759 (22.2%)	79.8%
1:1000 Top 10 Features	15825	936	2119 (13.4%)	888 (41.9%)	94.9%
1:100 No Correlating Features	148919	951	3865 (2.6%)	817 (21.1%)	85.9%
1:100 Train 1:1000 Test	15825	936	1096 (6.9%)	914 (83.4%)	97.6%
1:1000 Train 1:100 Test	149870	951	15201 (10.1%)	944 (6.2%)	99.2%

Table 6: Results XAI Model with a 0.35 threshold

Dataset Used	Dataset Size	Fraud Dataset Size	Dataset Size at Threshold (Percentage of Total)	Fraud Captured (Percentage of Threshold Dataset)	Percentage of Fraud Captured (Precision)
1:100 Sample	149961	993	665 (0.4%)	477 (71.7%)	48%
1:1000 Sample	15917	932	972 (6.1%)	754 (77.6%)	79.3%
1:100 Top 10 Features	149870	951	744 (0.5%)	454 (61%)	47.7%
1:1000 Top 10 Features	15825	936	988 (6.2%)	744 (75.3%)	79.5%
1:100 No Correlating Features	148919	951	645 (0.43%)	454 (70.4%)	47.7%
1:100 Train 1:1000 Test	15825	936	837 (5.2%)	833 (99.5%)	89%
1:1000 Train 1:100 Test	149870	951	2637 (1.8%)	919 (34.9%)	96.6%

Table 7: Results XAI Model with a 0.7 threshold

Dataset Used	Dataset Size	Fraud Dataset Size	Dataset Size at Threshold (Percentage of Total)	Fraud Captured (Percentage of Threshold Dataset)	Percentage of Fraud Captured (Precision)
1:100 Sample	149961	993	168 (0.1%)	160 (95.2%)	16.1%
1:1000 Sample	15917	932	559 (3.5%)	519 (92.8%)	54.5%
1:100 Top 10 Features	149870	951	184 (0.12%)	150 (81.5%)	15.8%
1:1000 Top 10 Features	15825	936	618 (3.9%)	538 (87.1%)	57.5%
1:100 No Correlating Features	148919	951	168 (0.11%)	156 (92.9%)	16.4%
1:100 Train 1:1000 Test	15825	936	557 (3.5%)	556 (99.8%)	59.4%
1:1000 Train 1:100 Test	149870	951	1183 (0.79%)	792 (66.9%)	83.2%

5 Fraud Expert Opinion

As described in section 1.7, results from a model need to be evaluated. For the created XAI model, this consists of multiple validations, further elaborated in section 2.7. For the qualitative validation, an expert opinion is needed. One of the academic contributions of this research is a confirmation of the correctness of the obtained results from experts in the field of Fraud Detection to see if the explanations are aligned with their knowledge and expertise. This expert opinion is formed by conducting interviews with Data Science, Rule Writing and Alert Handling. The construction of the interviews is done with the main concepts, outcomes and properties of XAI in mind, as described in table 1.

Contents

5.1 Fraud Experts

The evaluation is done with all domains within ING that deal with fraud. Each domain has a different perspective on the implementation of the XAI model and how it has potential added value in their way of working. Relevant domains are the teams that are directly involved with the monitoring, model creation or rule writing within the fraud detection department within the bank. In other words, Data Science, Rule Writing and Alert Handling are all individually interviewed.

5.1.1 Data Science

The Data Science department is responsible for developing and maintaining models that detect fraudulent transactions. These models, which can be both rule-based and machine learning-driven, incorporate various features focusing on different aspects of for example a transaction. The models calculate a score based on these combined features, which rulewriters can use in their rules to trigger alerts when a specific threshold is reached. Regular evaluation and updates are essential to keep the models effective. Explainability, or understanding how the model makes its decisions, is crucial for building trust and making necessary adjustments.

5.1.2 Rule Writing

The Rulewriting department creates and maintains rules to flag potentially fraudulent transactions, often integrating Data Science models and their scores to improve detection. This department handles both real-time fraud monitoring and long-term trend analysis, requiring rulewriters to understand and react to the behavior of various features within fraud detection models. Their daily tasks include examining recent fraud cases, identifying emerging trends, and crafting rules to adapt to these trends. Collaboration is key in this department, allowing rulewriters to share insights and develop more efficient rules. However, the ever-growing ruleset due to evolving fraud strategies can lead to a loss of individual oversight, making it necessary for other teams to explain triggered rules developed by different teams.

5.1.3 Alert Handling

The Alert Handling department investigates flagged transactions and takes appropriate actions, such as contacting customers via phone or email. Their work is dynamic, with new trends and patterns emerging daily. When a transaction is flagged, alert handlers research the reasoning behind it and, if necessary, contact the customer directly for immediate resolution. They often encounter customers unaware of being scammed or in denial, particularly in cases like dating fraud, where a phone call is crucial to clarify the situation. Alert handlers also communicate with other banks for further clarification via email. The ever-changing fraud trends require constant updates to rules and systems to keep pace with emerging technologies and market changes like the rise of cryptocurrency platforms.

5.2 Interview Construction

The interviews with the Data Science, Rule Writing, and Alert Handling departments were structured to gather insights on how the output of the XAI model can enhance aspects such as effectiveness, transparency, explainability and trustworthiness of fraud detection systems. These are examples of he key characteristics, properties and outcomes of XAI which have been thoroughly discussed in section 2.3.2 and outlined in table 1. This has been implemented differntly for each department

In the Data Science department, the interviews are constructed to delve into how XAI principles are incorporated in the development and deployment of fraud detection models. The focus was on examining the importance of creating models that deliver accurate predictions and provide understandable explanations, assessing how explainability contributes to the correctness and quality of these models through validation and feedback iterations, and evaluating the methods used to ensure model transparency and interpretability.

For the Rule Writing department, the interviews aimed to understand how XAI can enhance the creation and maintenance of fraud detection rules. Key areas of focus included ensuring the transparency of rules, providing justifications and rational explanations for the rules, and maintaining fairness and accountability in rule formulation.

In the Alert Handling department, the interviews were tailored to investigate how XAI impacts the processing and management of fraud alerts. The focus was on understanding how XAI influences user acceptance and trust in the alert handling system, facilitates the discovery of new knowledge through the analysis of fraud alerts, and contributes to the safety and reliability of the alert handling processes.

In addition to the department specific questions, several qualitative metrics and factors are considered. These can be broadly categorised into application-grounded metrics and human-grounded metricsZhou et al. (2021). Functionality-grounded metrics are not taken into account for the interviews as they are not suited for human evaluations. As shown in figure 25, Application-grounded and Human-grounded are evaluations that are human-centrered Zhou et al. (2021).

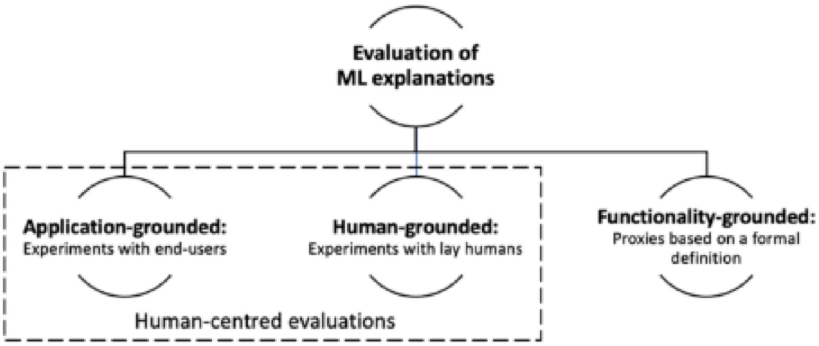

Evaluation of ML explanations
Application-grounded:
Experiments with end-users
Human-grounded:
Experiments with lay humans
Functionality-grounded:
Proxies based on a formal definition
Human-centred evaluations

Fig. 25: Overview of Methods to Evaluate ML Explanations

5.2.1 Application grounded metrics

Application-grounded metrics are designed to evaluate the performance of machine learning models in real-world settings that closely mimic their intended use cases Zhou et al. (2021). These metrics ensure that the models are not only theoretically sound but also practically useful in their deployment contexts.

Additionally, contextual relevance ensures that the data, tasks, and users involved in the evaluation are representative of the actual deployment environment. This is essential for the model's practical utility. An example of this is the task success rate which measures how well the model helps users complete a specific task. It reflects the practical utility of the model's interpretability in achieving desired outcomes. A high task success rate indicates that the model's explanations are not only understandable but also actionable, enabling users to perform tasks more effectively. For the created XAI explanations, it is structured in the interviews such that the impact of the shown features are correct and aligned with the expertise of each department.

Another application grounded metric is user efficiency which evaluates the time taken by users to complete tasks with the help of the model's explanations. Faster task completion times indicate better usability and efficiency of the explanations. This metric highlights the importance of the usability of the model's outputs, emphasizing that interpretability is not just about understanding the model's decisions but also about how efficiently users can leverage these explanations to achieve their goals. This metric is more focused on Alert Handling as they are tasked with investigating the flagged transactions. Alert handlers are asked if the provided explanations would shorten their time to investigate transactions.

Lastly, the decision-making quality. This metric assesses the quality of decisions made by users when aided by the model's explanations. It includes various aspects such as decision accuracy, confidence in decisions, and the ability to detect errors or biases. High-quality decision-making indicates that the explanations are not only clear but also reliable and insightful, helping users make better-informed decisions. During the interview, each department is asked if the provided explanations will enhance their decision making. The detection of errors will be similar to the task success rate because in this case, it means that they are asked if the features are correct. The constructed test cases described in section 5.2.3 will furter elaborate on the detection of potential bias or further incorrectness.

5.2.2 Human-Grounded Metrics

Human-grounded metrics focus on the interpretability and usability of machine learning models from a human perspective. These metrics are crucial for ensuring that the models are understandable and actionable by human users. Psychological constructs play a significant role in this evaluation.

The first metric taken into account here is comprehensibility, which looks at how easily users can understand the explanations provided by the model. High comprehensibility ensures that users can quickly understand the reasons behind the model's outputs which should improve effectiveness. This metric is important for increasing user acceptance and confidence in the model, as indicated in 1. All departments are given a small tutorial explaining the plots that are presented. After this, all are asked if the visible plots are understandable and clear to work with.

Apart from comprehensibility, user trust and confidence is also of high importance. It evaluates the level of trust and confidence users have in the model's explanations. Shown in section 2.3.2, trust and confidence are critical for the adoption and continued use of machine learning models, as they influence whether users will depend on the system in decision-making processes. High levels of trust and confidence indicate that users feel assured in relying on the model's outputs. This is tested with the use of the four test cases structured in 5.2.3. Here, all factors of the transaction are taken into account and based on all the information, all departments judge whether or not they would trust the outcome provided by the model.

Lastly, an important metric to take into account is user satisfaction. This assesses overall user satisfaction with the explanations. This is subjective and is likely to differ for each department. High user satisfaction indicates that users find the explanations not only clear but also beneficial and relevant to their needs. This is the concluding question asking if they are satisfied with the received results.

5.2.3 Test Cases

Apart from the constructed interview content based on the literature in this section, use cases are created to test the potential added value of the model for real life scenarios. During the first section of the interview, examples of global and local explanations are presented to qualitatively validate the results. This is done to explain how the output of the model would look like when it is implemented. These explanations are in the form of plots, explained in section 4.4. The bar and beeswarm plots are presented to test whether a preference can be made between both plots. A small tutorial for both cases is given and impressions are documented.

As with previous content, results will be validated across multiple domains within ING that are in contact with fraud detection. Due to the dataset covering information about the transactions themselves(e.g. the sender, receiver and amount money to name a few), the scope of the model is also bounded by these aspects. Therefore, it is of high added value to test outliers with potential out of scope reasoning to see if the model can really contribute something or if there are errors and weaknesses present. In order to identify suited cases, the array used in 4.5 is used. This serves as the foundation for the construction of the test cases in this section. Within this array, the transaction id is included in the dataset. This transaction id of the four discovered outliers is documented. An employee within Data Science who is not interviewed has completed the case information. This consist of an investigation including the reasoning behind the final verdict for each individual transaction. This information, together with the description of each feature, has been send to Data Science, Rule Writing and Alert Handling in advance. Again, because of the confidentiality of this topic and therefor the dataset, this information is not shared with the researcher.

In order to prevent bias towards certain features in the explanations of the XAI model, the plots containing the local explanations for these features are send before the interview. The goal is for each department to construct an impressions and prediction of the provided scenario before the model's prediction is shown. Again, this is done to prevent bias. If the answers are given in advance, a path to the answer is constructed instead of objectively dealing with the situation. In other words, reverse engineering the answer is prevented.

The first presented case, case A, involves a transaction from the false positive dataset, documented in section 4.1, that received a high fraud prediction score by the XAI model. This case tests whether the XAI model's reasoning aligns with the triggered rules and whether new insights can be discovered. The second case, case B, features a transaction that the model predicts as having a low chance of being fraudulent, despite the rules initially flagging it. The goal here is to determine if the XAI model can provide insights to see if there was even a need for rules to be triggered. Case C presents a fraudulent transaction that the model correctly identifies with a confident prediction score of one. The aim is to validate the model's certainty and reasoning. Lastly, Case D addresses a critical scenario where the model predicts a very low fraud probability, yet the transaction is fraudulent. This case aims to uncover potential weaknesses in the model or identify if the reasoning behind the prediction lies outside the model's scope.

6 Interview Results

During this evaluation phase, results have been qualitatively validated with multiple fraud related departments within ING. As stated in 5, the interview is constructed in such a way that the implementation opportunities of the model have been tested. It is of importance that this can be structured so that it becomes an added value instead of hindering the workflow. In addition, results from the model have been shown and qualitatively validated from both a technical and business perspective.

Contents

6.1 Data Science Department

The Data Science department is responsible for developing and maintaining the models used to detect fraudulent transactions. These models can be both rule-based and ML driven, incorporating various features that focus on different aspects of a transaction, such as the sender, receiver, platforms, and underlying institutions. The models are designed to calculate a score based on the combination of these features. Rulewriters can use such models with their scores in their rules, triggering alerts when a predetermined threshold is reached.

The models are constantly evaluated and updated to ensure they remain effective. Explainability is a key aspect of model validation, helping both data scientists and other stakeholders understand how the model arrives at its decisions. This understanding is crucial for gaining trust in the model and making necessary adjustments to improve its performance. Essentially, explaining how the model's score is structured allows Data Science to better interpret and utilize the score, leading to a deeper understanding of the model's functionality. Interviews with the Data Science department have provided insightful findings about the potential implementation of XAI for their domain, models, and communication with other domains. The first discovery focuses on the benefit XAI provides when it comes to testing the features used in models, and eventually testing and evaluating the models themselves.

6.1.1 Feature Testing

One of the most important promises of the created models is that they are robust and can handle a variety of fraud patterns. This involves using diverse datasets for training and testing the models, as well as employing techniques like cross-validation to ensure the models are generalizable.

Robustness is essential for maintaining the accuracy of fraud detection over time, especially as fraud patterns evolve.

Technical errors in the Data Science models can be detected using the XAI model by detecting structural deviation in feature behavior. Adding additional checks can reveal if a feature fails to contribute when it should, suggesting areas for model improvement. This makes it possible to create a more focused and overarching test for the features in question and the model in general. Currently, a handmade dataset is used to test features, but it is recognized that this method may not cover all scenarios. The goal is to simulate reality as closely as possible, but perfect accuracy is unattainable due to the inconsistencies of humans. People sometimes do things that are really unexpected: buy a car or don't even realise that they're contributing to fraud. This can be correct, but it is very hard to identify such behavior with a model that should cover all scenarios. A line must be drawn at some point: how many fraudulent flags do I take resulting in how many false positives I accept are part of the model. Having explanations for the model that show the influence of features and the score help to identify patterns and again, potential deviation in feature expectations. So in this way, it's possible to figure out what the ultimate impact of the feature is, why doesn't it do what we expected to happen with it? Investigating features that lower the fraud score, rather than increase it, is also valuable. Identifying these features can inform rule creation and model adjustments.

During the interview, especially when looking at local explanations, an insightful findings was discovered. The department has identified interesting cases where features do not perform as expected, often due to edge cases or insufficient data. For example, when historical data is sparse, certain features might yield a preset value, which indicates insufficient data rather than a clear fraud indication. So there is too little data which in itself can also mean something. The explanations during the interview showed that these values also can have significant impact on the fraud prediction of the model. Understanding these anomalies can further provide insights into feature behavior and highlight the importance of manually set values in the overall prediction.

6.1.2 Feature Selection

In addition to feature testing, it can also be very beneficial for feature selection. Feature selection is crucial for efficient model performance. Selecting the top 20 features, for instance, can save costs and computational power, focusing on those with high correlation to fraud or the model specific target. Moreover, you can also run models selecting the lowest performing features to see how they behave in this new environment. This helps to further analyse their behavior.

Pattern recognition tests with newly selected features help outline how features influence the model's predictions. These tests can reveal how different features complement each other, leading to the creation of new, more effective features. This can further improve models.

6.1.3 Improved Stakeholder Communication

Data scientists collaborate closely with rulewriters, especially when deploying new models. Model scores often inform the creation or adjustment of rules, ensuring alignment and effectiveness in fraud detection. Feedback from alert handlers provides real-world insights into model performance, guiding necessary adjustments. This iterative process of feedback and refinement is essential for maintaining an effective fraud detection system.

Regular discussions with rulewriters help data scientists understand their needs and challenges, tailoring models to support decision-making processes. One of the primary challenges is ensuring models are interpretable. The ability to explain model predictions is crucial for trust and transparency, aiding in model validation and governance.

Visualization tools are used to present model explanations in an easily understandable format. These visualizations help less technical stakeholders, such as rulewriters and alert handlers, grasp the model's decisions intuitively. It is also very useful for the Model Validation department who has to validate their models before they can roll into production. Model validation also wants to know all relevant information about the model. This can be done with more ease with the help of the explanations provided by the XAI model. The explanations secures effective communication of model insights which ensures all stakeholders can effectively use the models.

6.1.4 Qualitative Results Validation

During the inspection of the global explanation and the added value of features across the entire dataset, certain insights were noted. For instance, after looking at the beeswarm plot as shown in D, feature F9 behaves contrary to the other features, which aligns with existing domain knowledge. It was noticed immediately and proved to be an interesting finding for the Data Science because it clearly shows that the feature is only of importance when it has a low value.

In terms of readability, Bar plots provide better clarity compared to beeswarm plots, particularly in illustrating the importance of a few significant features. While beeswarm plots lack detailed explanations, bar plots effectively reflect the proportions of feature importanc making them the preferred option. Regarding the number of features visible per plot, there is a debate between displaying all 50 features or limiting the number to avoid clutter. Data scientists prefer all features visible to understand each feature's contribution, facilitating model adjustments. In contrast, alert handlers may prefer to see only the top features to prevent information overload.

Data Science validated the global features displayed, confirming alignment with current expertise, and highlighted the top 3 and bottom 3 features. F41 was recognized as a highly useful feature, frequently employed by rule writers as a fundamental component in their models, justifying its high ranking. F42 was deemed highly relevant to a significant characteristic of the transaction, explaining and supporting its high importance. Similarly, F32, like F41, indicated potential significance. For features of lower importance, F11 was surprisingly low in contribution, performing almost nothing, which was unexpected. F6 and F1 were seen as correctly low impact.

6.2 Rule Writing Department

The Rule Writing department is responsible for creating and maintaining rules that flag potentially fraudulent transactions. As mentioned during the interviews with Data Science, these rules are able to incorporate Data Science models and their scores. The Rule Writing department is responsible for both real-time fraud monitoring and long-term trend analysis. Rule writers are tasked with understanding and reacting to the behavior of various features within fraud detection models. Their daily activities involve examining recent fraud cases, identifying emerging trends, and writing rules to adapt to these trends. This constantly updating their models is essential as fraudsters continually adapt their strategies.

In practice, rule writers combine their own expertise and collaborative knowledge to engineer the rules. It can happen that something is seen, which seems new to one, but someone else has already created an analytical rule to counter this. This collaboration creates more efficient work environment than when everything would be done alone. Because the team is constantly countering newly appearing trends, the ruleset increases significantly which can result in a loss of individual oversight. Triggered rules which are developed by other teams often have to be explained by other teams because of this. Rule writers implement a modus operandi (MO), which refers a criminal activity (i.e. investment fraud), into their rules.

For further readability increase, the difference between a model and features will be explained. A feature is a calculation of a characteristic of the transaction. A model takes multiple characteristics into account to calculate potential suspicious behavior resulting in a score as output. The score of a model, together with different or common features, can form a rule. For example, a model which looks at country and amount of money results in a score based on these two features. This score can be used in a rule together with the feature frequency of money transfer. This way, the rule covers multiple characteristics of a transaction creating a prevention method for these characteristics. If it turns out that the feature amount of money is leading in this model, deciding solely what the output of the model score will be, it would be more wise to add the feature amount of money as a stand-alone feature instead of the score of the model. Because if you use a model that's so controlled by one feature, the model doesn't much content. In fact, it can be dangerous when the leading feature is low while other features in the model do show fraud. In this case, based on that model, you could miss fraud.

The rule writers also stress the importance of understanding feature combinations. Sometimes, specific feature interactions can lead to false positives or false negatives, affecting the overall accuracy of the fraud detection system. By analyzing these interactions, rule writers can refine their rules to better capture fraudulent activities without unnecessarily flagging legitimate transactions.

6.2.1 Model Explanations

Although some rules only include separate features resulting in a very transparent rule, there are MOs where Data Science models are used. Here, XAI tools are highly valued by rule writers because they provide clarity on the model's scoring mechanisms, significantly reducing dependence on data scientists by offering clear explanations for why certain scores are generated. Understanding why a particular transaction received a high fraud score helps rule writers adjust their rules more effectively. During the interviews, Rule Writing mentioned that now it is not clear from time to time what is used in determining the model score. When all features are of high value, but the model results a low score, it is currently unclear how this score came to be. Depending on the output of a threshold, further steps are taken. In this case, Rule Writing could ask Data Science for explanations, but because there are hundreds of millions of transactions that this model should cover, it is out of their scope to provide individual explanations. In addition, it becomes a burden for Data Science when Rule Writing asks them for explanations for every case . With the local explanations created by the XAI techniques, the reliance on Data Science is reduced and the models used for the rules, and their exceptions become more clear.

In addition, when Data Science develops a new model, it is presented to Rule Writing showing all the features that are incorporated and their importance. Based on this presentation, rule writers can decide whether they want to implement this new model into new or existing rules. For instance, there are occurrences where a new model doesn't improve rules because incorporating some individual features in a rule does the same job. Because of this, some models are not used by rule writer at all. However, as fraud techniques change, so do the rules so the need for different models might change as well. By this time, the feature importance overview of each model is not clear anymore due to it being presented far in the past. Having an overview of the inner workings of each model at hand will increase the complete knowledge base of rule writers without having to ask Data Science.

While this improves productivity by minimising the need to consult data science when doubts arise, these explanations are often viewed as a "nice to have" rather than a necessity. If the workload for Data Science increases a lot by implementing this, it should be dropped. This has been mentioned because the current way of working also suffices.

6.2.2 Anomaly Explanations

The work of the rule writers is highly case-specific, focusing on individual instances of flagged transactions rather than the entire dataset. This case-level focus is an example for the need for explainability tools that provide insights on a per-case basis. As stated in 6.2.1, it is possible to ask Data Science, but models use hundreds of millions of records, so it's hard to discover the specific exceptions. This way, as a rule writer, you can see exactly at a local level why features get certain scores and place this next to your own expectation. Understanding why a model score is low despite all indicators being set to true can help rule writers adjust thresholds more effectively. For example, the impact of the score is reduced when a certain combination of features is discovered. This approach ensures that the rules are finely tuned to detect actual fraud while minimising false positives. It also allows for further model bug discoveries because the use case centred approach might discover flaws which were overlooked by Data Science.

However, Rule Writing mentioned a potential threat in this approach. If a rule writer discovers which features work well or poorly, there is a potential need to distance themselves from the created model by tweaking with these individual features. It can become some sort of a game to be a little better than the models written by Data Science. Rule Writing can shift the focus to specifically look at how discover exceptions faster and advising Data Science to implement this. However, the model is designed so that with the combination of all features together, the best possible outcome is achieved. A competition to try and win from the the model, should not be the focus of the job.

6.2.3 Improved Stakeholder Communication

There is a direct and continuous collaboration between rule writers and the alert handlers team. This relationship is vital because rule writers frequently update rules based on feedback from the alert handlers, who are directly in contact with flagged transactions. The rule writers rely on the alert handlers to provide real-world insights into the effectiveness of the rules, which helps in refining them further. Alert Handling needs Rule Writing to increase transparency for their rules in order to enable a efficient and effective research when a transaction is flagged for fraud. In addition, this rules needs to be understandable to improve this research, also allowing for new employees to understand.

Rule writers also work closely with data scientists, especially when new models are implemented. This collaboration ensures that the rules are effectively integrating model scores and adapting to new fraud patterns identified by the rule writers or alert handlers. The feedback loop from alert handlers to rule writers, and subsequently to data scientists, is crucial for refining both rules and models.

Strengthening underlying knowledge and communication with Data Science and Alert Handling, by increasing the explainability and understandability of the models being used across all domains, will be a benefit for everyone's workload and productiveness. XAI techniques can bridge this gap by showing the underlying mechanisms of each model, reducing the need for redundant global explanations and directly focusing on the problem. The overall knowledge base across all domains will increase.

6.2.4 Qualitative Results Validation

Looking at the global explanations in the form of an overview of each feature's importance, further insights were found. The top three was confirmed to be expected. Similar to section 6.1.4, features 42, 41 and 32 are all seen as very important. Rule Writing indicates this by frequently using these features as individual features in there rules. This is aligned with the comments from Data Science stating that these features can be considered to be a foundation for other features. After analysing the bottom three, feature 11 was expected to be a bit higher, but it depends on the MO that is taken into account. In this case, it looks at fraud in general, but it may be that it changes when you specifically look at a particular trend. It is important to note that the data set is a subset of a data set that does not show the complete picture. Apart from that, all features shown make sense to have this little impact. Special case feature 9 is also confirmed to be in place. The ordering of features makes sense.

The alignment of the local explanations with current knowledge and expertise was also positive. The values of features resulting in negative or positive impact has been confirmed to be logical. An interesting finding from Rule Writing was to see that the model showed a cumulative impact on the prediction instead of having a few features that solely decided the output. The graph mentioned here is similar to figure 26 and clearly shows how multiple features influence the prediction. The presence of a top 3 accumulating for 30/40% of the predictions is logical, but it was insightful form them to see that the model looks further, showing what else is important. This teamwork makes the model more interesting for a rulewriter, decreasing the need to implement the high impact features individually.

After discussing the readability of the different plots shown, the following finding was found. The depth of the model is not always within the expertise of a rule writer, and while the beeswarm plot clearly shows that feature 41 is important, its overall readability can be challenging. Understanding the details, such as the significance of data points on the x-axis and the meaning of the red colour, can be difficult for those without prior experience. This lack of clarity suggests a need for additional training. When considering readability, especially for non-technical stakeholders like managers, the beeswarm plot poses challenges in quickly explaining the model's workings. In contrast, a box plot offers clearer insights and makes it easier to understand and communicate the information effectively, thereby making it a stronger tool for interpretation.

6.3 Alert Handling Department

The Alert Handling Department at the bank is tasked with investigating flagged transactions and taking appropriate actions, such as contacting customers via phone or email. The work content changes every day with new trends and patterns appearing. When a flag enters the system, reasoning is researched. If the alert handler is in doubt or wants to take immediate answer, the customer can called to inform them about the current situation. Alert handlers often encounter customers unaware of being scammed or in denial about it such as with dating fraud. Here, a phone call is needed to clarify the situation. Other banks can also be asked for clarification via email.

Fraud trends vary significantly and change quickly, requiring constant attention. For example, previously unlikely items such as Pokémon cards have changed into a frequent occurring method for scams. Scammers often use new merchants and repeating patterns to exploit the system. This dynamic creates a constant "cat and mouse" game, necessitating regular updates to rules and systems to keep up with emerging technologies and market changes, such as the rise of cryptocurrency platforms.

6.3.1 Efficient Investigation

When a transaction is flagged for fraud, an investigation is initialised to separate fraud from a false positive. This process involves pulling up the customer's profile to verify the legitimacy of the transaction. Incorrectly blocking transactions can cause significant inconvenience to the clients, so it is crucial for alert handlers to make accurate judgements based on past interactions and behavior patterns. An example here could be a sudden deviation from the common store the customer goes to. This, combined with for example an increase in spending limit, can be a substantiated decision to call the customer to check for a potential threat. The following example is unfortunately a textbook example. In practice, potential indicators for reasoning require a lot of experience and precision and are often not directly clear.

The alert handlers emphasise the need for explainable AI models to enhance their efficiency in their research. The primary challenge they face is the technical implementation of models that can provide real-time explanations. One of the ongoing challenges in alert handling is balancing the detection of actual fraud with minimising false positives. Understanding the features that contribute to a transaction's fraud score is crucial in achieving this balance. Tools that can interpret these features in real time would allow alert handlers to adjust thresholds more effectively, ensuring that legitimate transactions are not mistakenly flagged, which can lead to customer dissatisfaction. This capability would not only help them understand the rationale behind each alert but also enable them to make quicker, more substantiated decisions. The alert handlers also suggest tools that can provide contextual information about flagged transactions. For example, understanding the customer's transaction history and behavior can help them determine whether a flagged transaction is genuinely fraudulent or a false positive. This contextual information is helpful for making accurate decisions and reducing the number of unnecessary customer interactions.

6.3.2 Effective Training for new Employees

Training new employees the way of working or training more experienced employees new fraud trends is a significant aspect of maintaining an effective fraud detection organisation. When a flagged transaction is picked up, new employees often face the challenge of navigating trough a large amount of data without a clear point of reference. Tools that highlight key features of flagged transactions can provide a starting point for investigations, making it easier for new employees to understand the context and rationale behind alerts. This approach can streamline the training process, enabling new hires to become better at their work more quickly. An example, if a new employee receives an alert, they can focus on the top contributing factors to determine its validity. If the top 10 factors can be easily explained, the transaction may be pushed through, reducing unnecessary investigations. It also eases training more experienced employees, because trends can be shown trough easy to understand plots.

However, Alert Handling clearly mentioned that this tool should be seen as an assisting tool enabling these focus points, instead of leading the decision. The trick with fraud detection is that all safety checks but one could be marked as correct and the only incorrect one results in it being fraud. This tool allows for a clearer overview of the inside of the models that are used, increasing productiveness.

6.3.3 Improved Stakeholder Communication

As stated in section 6.2, Alert handlers provide valuable insights based on their direct interactions with flagged transactions, helping rule writers adjust thresholds and data scientists refine models. This collaboration ensures that the system remains effective and adaptable to emerging fraud trends. Alert handlers and rule writer are working especially close with each other. If they have set a rule too tight, so a threshold too low, the people of alert handling will be drowning in work. This also works the other way around, when there is a lot of damage and no rule present, there is a high need for one. This is called a signal to action.

While data scientists have less direct contact with alert handlers, their work is influenced by the feedback provided to rule writers. Integrating model scores into rules or creating new rules based on model outputs requires close cooperation between these departments. Implementing XAI tools here allows for a more streamlined communication network by increasing transparency and understandability in the models that are being used.

6.3.4 Qualitative Results Validation

Looking at the global explanations, there were no remarks regarding incorrect rankings. Both the top and bottom features were seen as logical and validated to their own expertise. Again, especially feature 42 and 32 were no surprises due to their common presence in the daily work of Alert Handling. However, in contrast to Data Science and Rule Writing, Alert Handling stated that a total of 50 features shown is too much information. It creates noise which draws the attention away from the most important focus points. A total of 20 would be preferred here. In terms of plot readability, the beeswarm plot was seen as quite valuable because it clearly shows the impact variance of each feature. It was noted that a small training is needed for this plot. A beeswarm plot covering only one transaction, as shown in D16, made it more clear how each dot represents an individual transaction. Notable features, such as F4, F9, and F29, stand out in these visualisations, making it a plot that allows for feature distribution insights. Feature F4 was especially interesting for them because it becomes clear that the feature has hardly any impact on the prediction if the value of the feature is low.

Also the local explanations are validated as correct. The first reaction that emerged was the combination of features 41 and 42. Alert Handling stated that, as is shown in the explanations, if both features have a high value, the chances of the transaction being fraudulent increases significantly. After further analysing different instances, cases where the model could actually mislead you were found as well. However, as indicated, if there is a new trend emerging you have to look at the starting patterns of it which the model will likely not pick it up immediately. It should be able to help, but it shouldn't be leading you. It is important to note here that the model is a only showing a small part of the complete picture.

6.4 Testing Use Cases

As documented in section 5.2.3, a total of four different test cases have been constructed to see whether or not the interviewee would have described the addition of XAI explanations as added value or not. Case A and B are cases having a false positive fraud transaction. The goal of the addition if these transactions is to see if, with the use of XAI explanations, the transaction would not have been seen as fraudulent. Case C and D are both cases where the transaction was fraudulent, but the model predicted it differently for each case. The cases were presented to Data Science, Rule Writing and Alert Handling to test the different perspectives on the matter.

6.4.1 Case A: False Positive with High Fraud Prediction

As the title implies, for the first case, a transaction from the false positive subset was retrieved and analysed. The model resulted in a high fraud prediction score as well. During this use case, it is tested if the model aligns with the reasoning behind the rules being triggered, and if new insights in the case would be discovered with the use of the model.

During the analysis of this test case, each domain had the following to say. Data Science noted that for this case, a selection of features are expected to rank very high. This had to do with the description of the case. Without mentioning the scenario, it was confirmed that features 4, 42 and 41 were within the range of expected high impact features, as also shown to be present in figure 26. Especially the value of 41 indicates an almost immediate reason to flag the transaction. Although the reasoning behind the flag for, both the description of the case, as for the XAI model was confirmed to be correct, Data Science noted that Alert Handling can also investigate other aspects. Rule Writing immediately noted that the model could not have covered the most important reasoning behind the flag, this is outside the scope of the dataset. They acknowledged that the top features in this case strongly indicated fraudulent activity. The rule did look at extra things that made the transaction a lot more suspicious. However, the explanations of the model are correct in the sense that it would be considered fraudulent, only less fraudulent than it is currently the case. Rule Writting noted that it also happens that the customer does not know if it is fraud. An example of this is romance scam. The customer thinks that they are in contact with a close relationship or beloved person. Therefore, it is not considered wrong while all signals indicate that it is. After contacting the customer and this is further confirmed by them, it will also be closed under non-fraudulent. Exceptions like these are therefor described as not fraudulent in the dataset, but are very likely to be fraudulent.

Alert handling described that the model would have send the alert handler in the wrong direction, which made sense when looking at the case. It was the case that the dataset only delineates a part of the case. In this case, it is not fraudulent due to ticked off safety checks outside the model's scope. It was concluded to be both a clear flag, but also a clear false positive. It is therefore important to draw a situational picture and not to trust the models output blindly. For implementation purposes, it is of high importance that the model should describe which part of the aspects it covers.

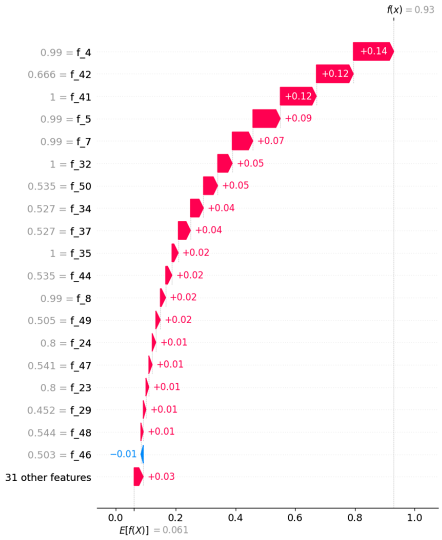


Fig. 26: Case A

6.4.2 Case B: False Positive with Low Fraud Prediction

Case B shows a transaction where the model indicates it to be of low chance to be a fraudulent transaction. Again, it is important to note here that, even though the final results was not fraudulent, the rules were originally triggered for it. In other words, with the use of the XAI model, would it even be needed to flag this transaction. The explanations are shown in figure 27.

Data Science observed that the alert was triggered by a new counter account, which was later determined not to be fraudulent. They noted that F41 was a prominent feature, with F32 also under observation. This suggested that the alert might have been based on a specific rule rather than the model, as such a case would likely not arise under the model's threshold. Rule Writing confirmed this by noting that there's a feature in a rule that's also triggered which is present in the scope of the model. However, feature 41 largely aligns with the case, so it was insightful that the model acknowledged this by pulling it towards the fraudulent prediction. As was the same for feature 42 and 32 in the opposite direction. In fact, it was confirmed that the model effectively demonstrates how the features behave.

Alert Handling stated that the case is clearly not fraudulent. In addition, it mentions that the model visualises the situation accordingly which looks clear. Alert Handling states the model also serves as an additional check for this case. All features push it towards a non fraudulent prediction, except for feature 41, 7 and 34. These features can be investigated further to improve the substantiation of the verdict and therefor minimise the risk.

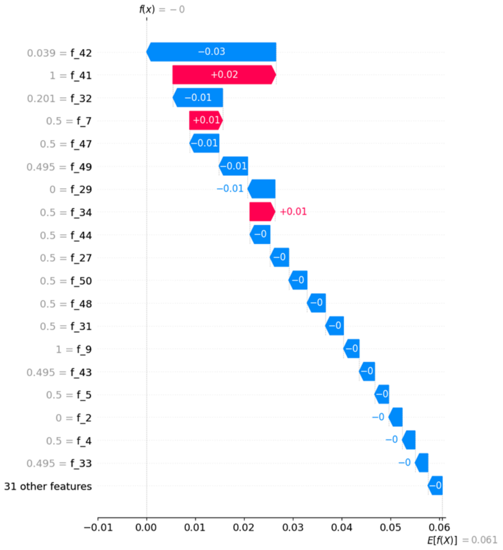


Fig. 27: Case B

6.4.3 Case C: Model Correctly Predicts High Chance of Fraud

Case C is a scenario where the transaction turned out to be fraudulent. The model correctly predicts this and even states that there is no doubt by giving a score of one as output for the prediction. The goal is to see if the model is rightfully indicating a lack of doubt and if the reasoning for it is correct. The XAI model's explanations are shown in figure 28.

Data Science's initial reaction to Case C was that, opposed to Case B, feature 32 now significantly contributed to the fraud score, along with F42 and other similar features, which were consistent with the model's values. Rule Writing agreed, noting that the model clearly indicated fraudulent activity, and the high value of F42 was in line with their expectations. Alert Handling indicates that for this case the situation is clearly fraudulent, it makes sense that feature 42 has a very high impact here. They emphasised that all signals pointed to fraud, and the transactional data supported this conclusion without doubt. All domains agree with the construction of the model's explanation.

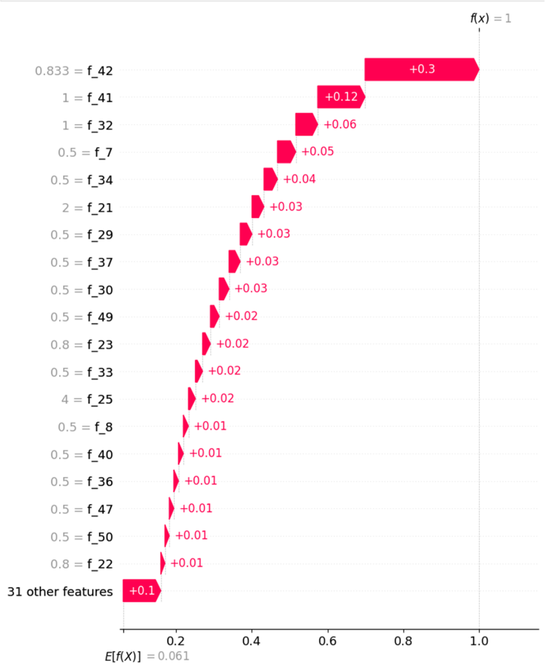


Fig. 28: Case C

6.4.4 Case D: Model Incorrectly Predicts Low Chance of Fraud

Case D represents a very dangerous case where the model's prediction shows a very low chance of fraud. However, the transaction turned out to be fraudulent. With the use of this case, the is goal is to identify potential weaknesses of the model or if it is a possibility that the reasoning is outside the scope of the model.

It was immediately noticed here that there are additional reasons, apart from transactional data, present within the use case. Without seeing the explanations of the XAI model, Data Science predicted the model to result in a low score with feature 42 to have a negative impact, which it did. Alert Handling viewed the case as an example of a difficult scenario where it is understandable that the model classifies the transaction as not fraudulent. They highlighted that while the model, which focuses on transaction data, did not flag it as fraudulent, the broader situational context revealed the criminal intent. A criminal has set up a connection here where he wants the first few transactions to stay low on the radar. Only then does he want to look at the real thing. It is clear that the model based on the dataset does not pick this up.

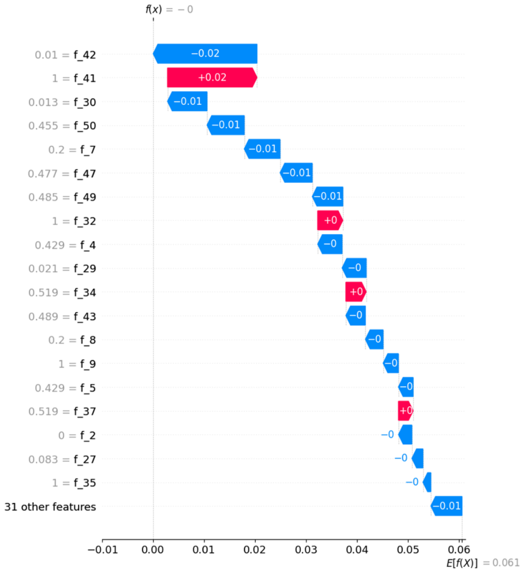

f(x) = −0
0.01 = f_42
1 = f_41
0.013 = f_30
0.455 = f_50
0.2 = f_7
0.477 = f_47
0.485 = f_49
1 = f_32
0.429 = f_4
0.021 = f_29
0.519 = f_34
0.489 = f_43
0.2 = f_8
1 = f_9
0.429 = f_5
0.519 = f_37
0 = f_2
0.083 = f_27
1 = f_35
31 other features
−0.02
+0.02
−0.01
−0.01
−0.01
−0.01
−0.01
+0
−0
−0
+0
−0
−0
−0
−0
+0
−0
−0
−0
−0.01
−0.01
0.00
0.01
0.02
0.03
0.04
0.05
0.06
E[f(X)] = 0.061

Fig. 29: Case D

7 Conclusions

Wihtin this chapter, an answer to the proposed research question in section 1.6 is given. An overview of all findings structured by answering all sub-questions is shown in section 7.1. In addition, the limitations, discussion and future research will be discussed.

Contents

7.1 Conclusion

In section 1.6, the following research question was formed:

How can Explainable Artificial Intelligence (XAI) techniques be utilized to enhance the transparency and interpretability of machine learning models in the detection of fraudulent transactions across all involved departments in a bank?

The main research question has been answered through a structured exploration of XAI in general, its application in fraud detection, and validation through expert opinions and practical implementation. The literature review revealed that Explainable AI (XAI) can be divided into main concepts, properties and outcomes. The focus for the main concepts lied in the enhancement of transparency and interpretability of models.

The taxonomy of XAI developed in this research categorizes models based on their transparency and the type of explanations they provide. This taxonomy serves as a framework for selecting appropriate XAI methods, distinguishing between pre-hoc and post-hoc explainability. After the framework was applied to the use case present in this research, the XAI technique SHAP was chosen as a suitable option. SHAP has been implemented in the development of the XAI model designed for transactional fraud detection which involved processing datasets provided by ING, analysing the performance of Random Forest, Gradient Boosting and SVM, and using SHAP to explain the model's predictions. The robustness of the model was tested through multiple iterations, ensuring that the explanations provided were consistent and reliable.

The results demonstrated that the XAI model could effectively identify and explain fraudulent transactions, offering valuable insights to fraud analysts. The XAI model's results include individual feature impacts (SHAP values), a base value, and original feature values for plotting. The final predictions for each transaction, combined with the correct labels, reveal that a 0.05 threshold results in many false positives and unnecessary work for alert handling. With a 0.35 threshold, the model captures 48% of fraud while encompassing 0.44% of the dataset, demonstrating fewer false positives as the threshold increases. Significant findings indicate that the "1:1000 Top 10 Features" and "1:1000 sample" tests outperform the "1:100 sample" datasets, especially at a 0.7 threshold, capturing over 54% of fraud with less than 4% of the original dataset remaining. Tests with imbalanced training and testing datasets show high fraud capture but also high false positives at lower thresholds. Overall, increasing the threshold enhances the model's precision by reducing false positives and improving fraud detection rates.

Expert opinions from the Data Science, Rule Writing, and Alert Handling departments within ING highlighted the practical benefits of the XAI model from both the technical and business perspective. The Data Science department's focus on feature testing and selection ensured that the most predictive features can be utilised and improve the current way of working. The Rule Writing department's evaluation of model and anomaly explanations provided insights into the model's decision-making process. The Alert Handling department's assessment of the model's contributions to investigation efficiency and training highlighted its potential to streamline fraud detection processes. In addition, all departments indicated that the model can improve stakeholder communication. The qualitative validation confirmed the model's robustness and practical utility.

7.2 Limitations

While this research provides valuable insights into the application of XAI for fraud detection in financial systems, several limitations must be acknowledged. These limitations suggest areas for future research, such as the data limitation and broader regulatory and ethical implications. The following limitations are documented.

- **Data Limitations and Model Scope:** The biggest limitation of this research is the dataset and therefor the scope of the model. The dataset used for developing and testing the XAI model was provided by ING and covers the transactional data which does not represent the full spectrum of a fraudulent scenario. This has been highlighted during the expert opinion showing the model's weakness. The false positives proved this by receiving high model scores. During the expert opinion, the reason for the transaction being a false positive was often outside the scope of the model. In addition, transactions deemed to be fraudulent by the model but (incorrectly) disapproved by the customer are set to non-fraudulent in the system. The model's utility was therefor seen as supporting decision making, but not leading it.
- **Qualitative Validation:** The expert opinions and qualitative validations were obtained from specific departments within ING. While these insights were valuable, they may not capture the full range of perspectives and requirements from the bank across all levels or different regulatory environments. For example, it was outside the scope of this research to ask higher management for further model implementation possibilities.
- **Regulatory and Ethical Considerations:** The research primarily focuses on the technical aspects of XAI and its application in fraud detection. However, broader regulatory and ethical considerations related to the deployment of AI in financial systems, such as data privacy and algorithmic fairness, were not deeply explored. This was also not possible with the anonymity of the dataset and could therefore be included in the future research. In addition, the field of AI and XAI is rapidly evolving. New techniques and advancements could significantly impact the findings and relevance of this research.

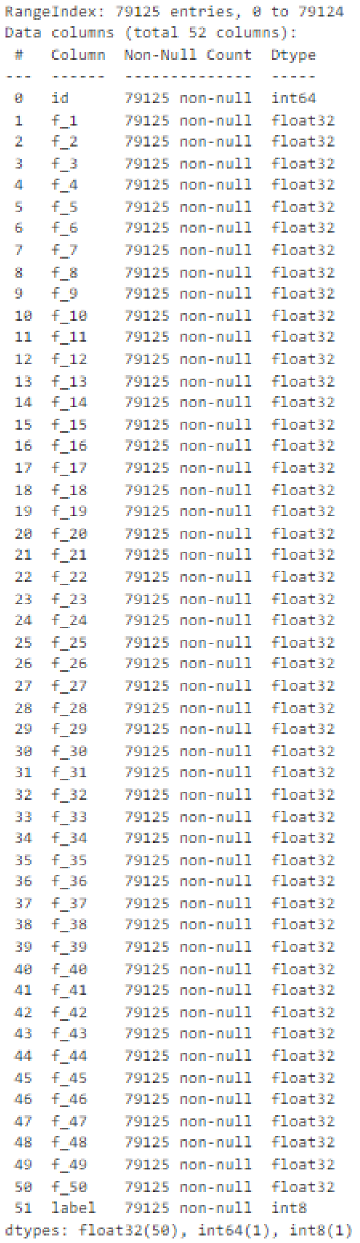
7.3 Future Research

Building on the findings and limitations of this research, multiple possibilities for future research can be identified to further enhance the application of XAI in fraud detection. By addressing these areas, future research can contribute to the development of more effective, transparent, and adaptable XAI models for fraud detection.

- **Broader Dataset Evaluation:** Future research should consider evaluating XAI models using datasets covering more than only transactional data features. This would help in understanding how different types of fraud and transaction patterns affect the model's performance and interpretability. As criminals continuously develop new tactics, it is essential to adapt XAI models to detect emerging fraud techniques. Research should focus on creating adaptive and resilient models that can learn from new data and identify these fraud patterns effectively.
- **Real-Time Implementation and Testing:** Implementing and testing XAI models in real-time environments would provide valuable feedback on their practical utility and integration challenges. Providing Alert Handling with the model and see how they investigate flagged transaction could identify new opportunities or weaknesses. In addition, setting up a training case for Alert Handling could improve this feedback loop as well. Conducting studies to observe the long-term impact of XAI models on fraud detection processes and outcomes enhance potential improvements as well. These studies would help in understanding how the models evolve, their effectiveness, and their influence on organizational practices and policies. Engaging a broader range of stakeholders, including new users such as compliance officers in the evaluation process can provide deeper insights into the usability and effectiveness of XAI models. User-centric studies and feedback loops can help refine the models to better meet the needs and expectations of various stakeholders.

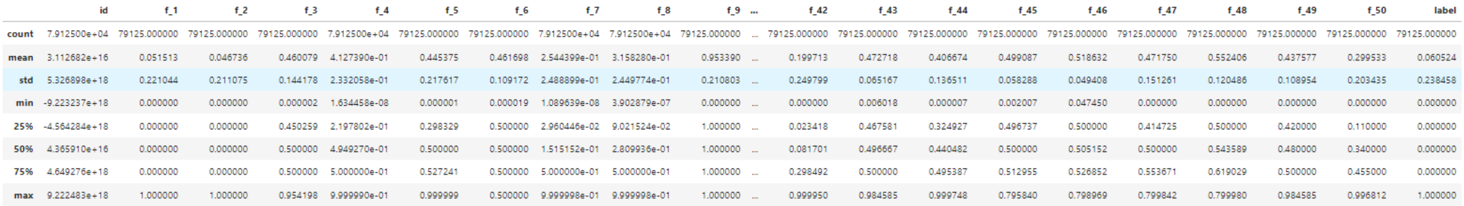
- **Regulatory and Ethical Implications:** Investigating the regulatory and ethical considerations associated with the deployment of XAI in fraud detection is crucial. Future research should explore how to ensure compliance with data privacy laws, mitigate biases, and promote fairness and transparency in AI-driven decision-making processes. Encouraging interdisciplinary collaboration between AI researchers, domain experts in finance, regulatory authorities, and ethicists can maintain a more comprehensive approach to developing and deploying XAI models. This collaborative effort can ensure that technical advancements are aligned with practical, regulatory, and ethical considerations

A Dataset Overview



```
RangeIndex: 79125 entries, 0 to 79124
Data columns (total 52 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   id      79125 non-null  int64  
 1   f_1     79125 non-null  float32
 2   f_2     79125 non-null  float32
 3   f_3     79125 non-null  float32
 4   f_4     79125 non-null  float32
 5   f_5     79125 non-null  float32
 6   f_6     79125 non-null  float32
 7   f_7     79125 non-null  float32
 8   f_8     79125 non-null  float32
 9   f_9     79125 non-null  float32
 10  f_10    79125 non-null  float32
 11  f_11    79125 non-null  float32
 12  f_12    79125 non-null  float32
 13  f_13    79125 non-null  float32
 14  f_14    79125 non-null  float32
 15  f_15    79125 non-null  float32
 16  f_16    79125 non-null  float32
 17  f_17    79125 non-null  float32
 18  f_18    79125 non-null  float32
 19  f_19    79125 non-null  float32
 20  f_20    79125 non-null  float32
 21  f_21    79125 non-null  float32
 22  f_22    79125 non-null  float32
 23  f_23    79125 non-null  float32
 24  f_24    79125 non-null  float32
 25  f_25    79125 non-null  float32
 26  f_26    79125 non-null  float32
 27  f_27    79125 non-null  float32
 28  f_28    79125 non-null  float32
 29  f_29    79125 non-null  float32
 30  f_30    79125 non-null  float32
 31  f_31    79125 non-null  float32
 32  f_32    79125 non-null  float32
 33  f_33    79125 non-null  float32
 34  f_34    79125 non-null  float32
 35  f_35    79125 non-null  float32
 36  f_36    79125 non-null  float32
 37  f_37    79125 non-null  float32
 38  f_38    79125 non-null  float32
 39  f_39    79125 non-null  float32
 40  f_40    79125 non-null  float32
 41  f_41    79125 non-null  float32
 42  f_42    79125 non-null  float32
 43  f_43    79125 non-null  float32
 44  f_44    79125 non-null  float32
 45  f_45    79125 non-null  float32
 46  f_46    79125 non-null  float32
 47  f_47    79125 non-null  float32
 48  f_48    79125 non-null  float32
 49  f_49    79125 non-null  float32
 50  f_50    79125 non-null  float32
 51  label   79125 non-null  int8   
dtypes: float32(50), int64(1), int8(1)
```

Fig. A1: Sample Datatype Overview



	id	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	...	f_42	f_43	f_44	f_45	f_46	f_47	f_48	f_49	f_50	label
count	7.912500e+04	79125.000000	79125.000000	79125.000000	7.912500e+04	79125.000000	79125.000000	7.912500e+04	7.912500e+04	79125.000000	...	79125.000000	79125.000000	79125.000000	79125.000000	79125.000000	79125.000000	79125.000000	79125.000000	79125.000000	79125.000000
mean	3.112682e+16	0.051513	0.046736	0.460079	4.127390e-01	0.445375	0.461698	2.544399e-01	3.158280e-01	0.953390	...	0.199713	0.472718	0.406674	0.499087	0.518632	0.471750	0.552406	0.437577	0.299533	0.060524
std	5.326898e+18	0.221044	0.211075	0.144178	2.332058e-01	0.217617	0.109172	2.488899e-01	2.449774e-01	0.210803	...	0.249799	0.065167	0.136511	0.058288	0.049408	0.151261	0.120486	0.108954	0.203435	0.238458
min	-9.223237e+18	0.000000	0.000000	0.000002	1.634458e-08	0.000001	0.000019	1.089639e-08	3.902879e-07	0.000000	...	0.000000	0.006018	0.000007	0.002007	0.047450	0.000000	0.000000	0.000000	0.000000	0.000000
25%	-4.564284e+18	0.000000	0.000000	0.450259	2.197802e-01	0.298329	0.500000	2.960446e-02	9.021524e-02	1.000000	...	0.023418	0.467581	0.324927	0.496737	0.500000	0.414725	0.500000	0.420000	0.110000	0.000000
50%	4.365910e+16	0.000000	0.000000	0.500000	4.949270e-01	0.500000	0.500000	1.515152e-01	2.809936e-01	1.000000	...	0.081701	0.496667	0.440482	0.500000	0.505152	0.500000	0.543589	0.480000	0.340000	0.000000
75%	4.649276e+18	0.000000	0.000000	0.500000	5.000000e-01	0.527241	0.500000	5.000000e-01	5.000000e-01	1.000000	...	0.298492	0.500000	0.495387	0.512955	0.526852	0.553671	0.619029	0.500000	0.455000	0.000000
max	9.222483e+18	1.000000	1.000000	0.954198	9.999990e-01	0.999999	0.500000	9.999998e-01	9.999998e-01	1.000000	...	0.999950	0.984585	0.999748	0.795840	0.798969	0.799842	0.799980	0.984585	0.996812	1.000000

Fig. A2: Sample Dataset Overview

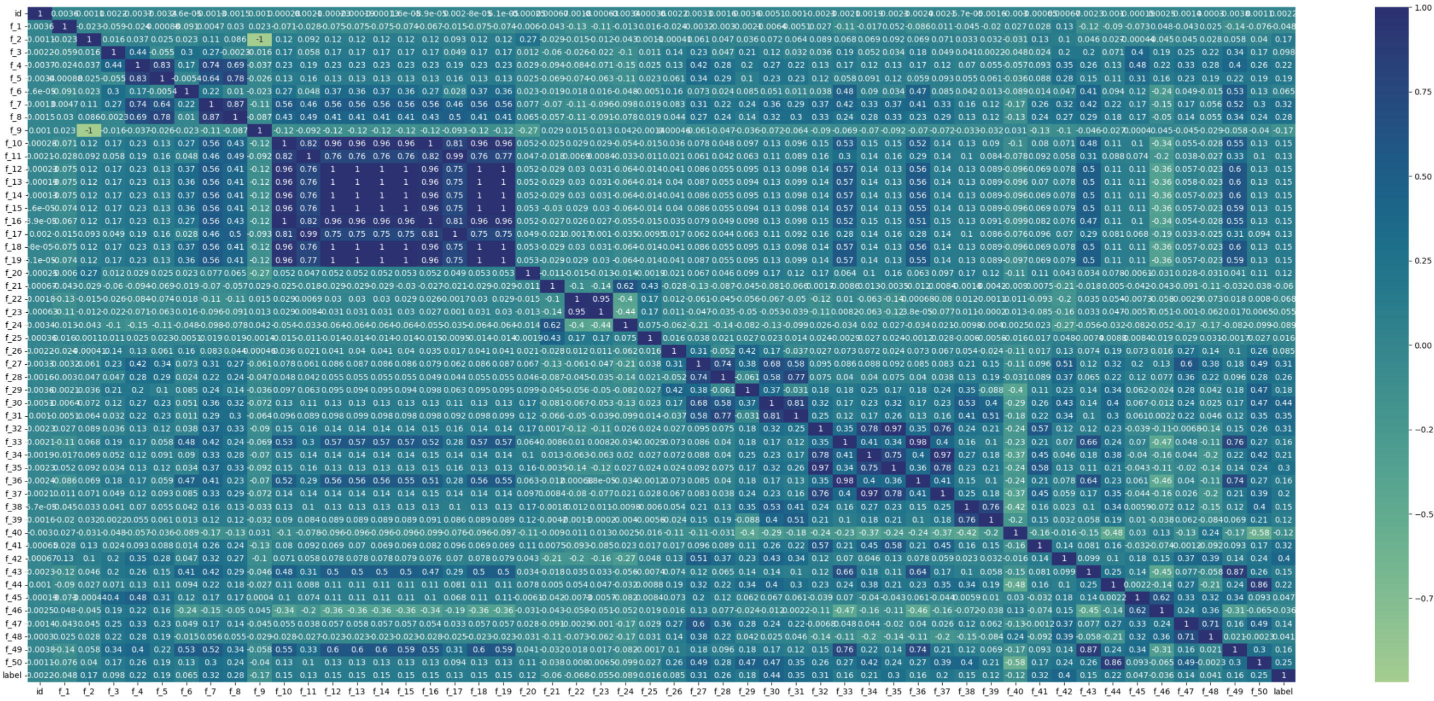


Fig. A3: The correlation heat map for all features

B LIME Result

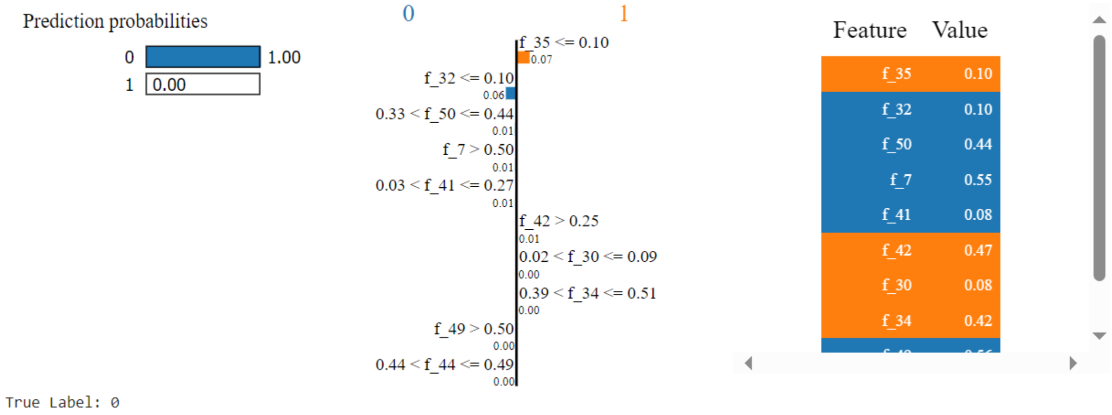
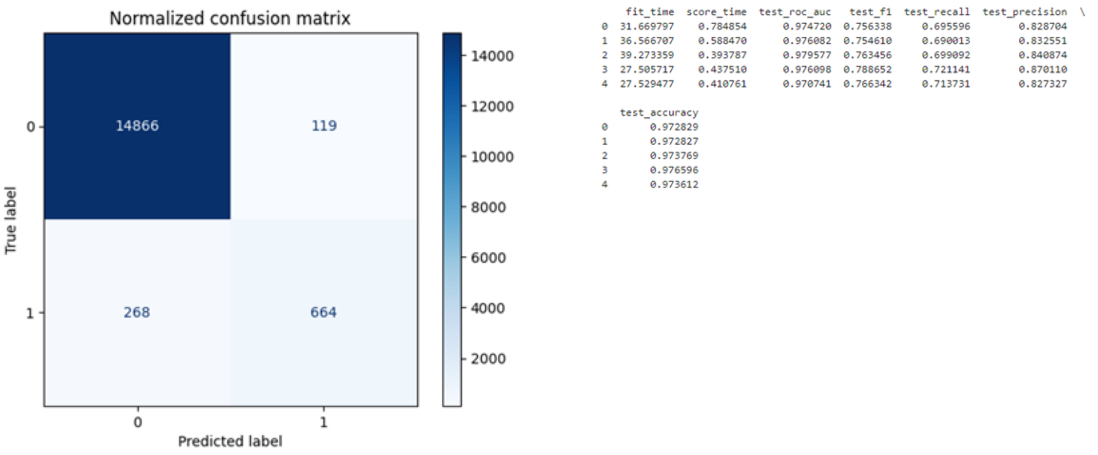


Fig. B4: LIME explanations

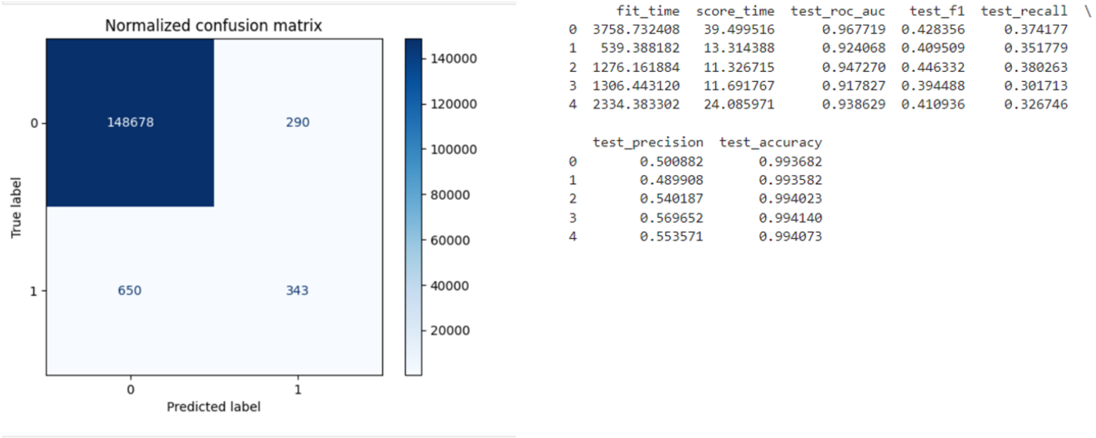
C ML Classifiers Performance



```
    fit_time  score_time  test_roc_auc   test_f1  test_recall  test_precision  \
0  31.669797    0.784854      0.974720  0.756338     0.695596        0.828704
1  36.566707    0.588470      0.976082  0.754610     0.690013        0.832551
2  39.273359    0.393787      0.979577  0.763456     0.699092        0.840874
3  27.505717    0.437510      0.976098  0.788652     0.721141        0.870110
4  27.529477    0.410761      0.970741  0.766342     0.713731        0.827327

   test_accuracy
0       0.972829
1       0.972827
2       0.973769
3       0.976596
4       0.973612
```

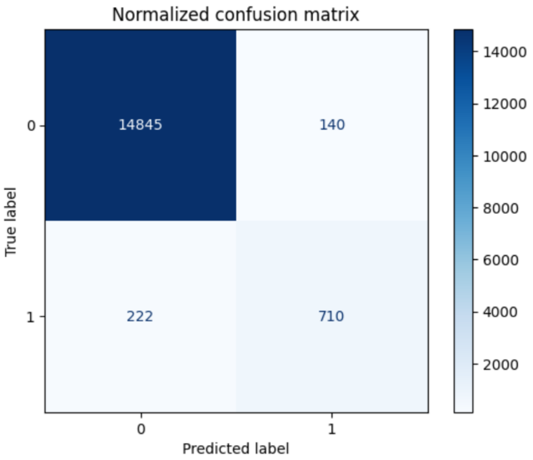
Fig. C5: R1 : 1000 Random Forest Results

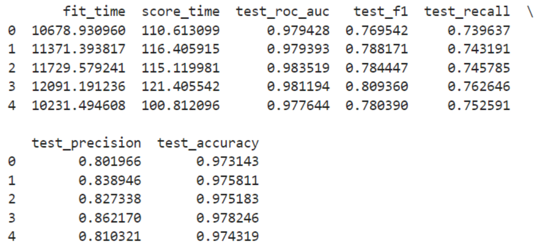


```
      fit_time  score_time  test_roc_auc   test_f1  test_recall  \
0  3758.732408   39.499516      0.967719  0.428356     0.374177
1   539.388182   13.314388      0.924068  0.409509     0.351779
2  1276.161884   11.326715      0.947270  0.446332     0.380263
3  1306.443120   11.691767      0.917827  0.394488     0.301713
4  2334.383302   24.085971      0.938629  0.410936     0.326746

   test_precision  test_accuracy
0        0.500882       0.993682
1        0.489908       0.993582
2        0.540187       0.994023
3        0.569652       0.994140
4        0.553571       0.994073
```

Fig. C6: 1 : 100 Gradient Boosting Results

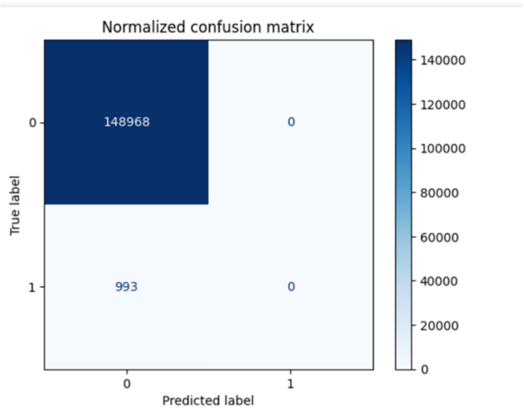


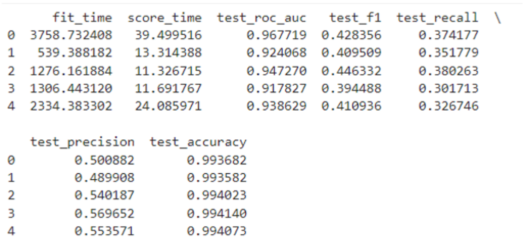


```
        fit_time  score_time  test_roc_auc   test_f1  test_recall  \
0   10678.930960  110.613099      0.979428  0.769542     0.739637
1   11371.393817  116.405915      0.979393  0.788171     0.743191
2   11729.579241  115.119981      0.983519  0.784447     0.745785
3   12091.191236  121.405542      0.981194  0.809360     0.762646
4   10231.494608  100.812096      0.977644  0.780390     0.752591

   test_precision  test_accuracy
0        0.801966       0.973143
1        0.838946       0.975811
2        0.827338       0.975183
3        0.862170       0.978246
4        0.810321       0.974319
```

Fig. C7: 1 : 1000 Gradient Boosting Results

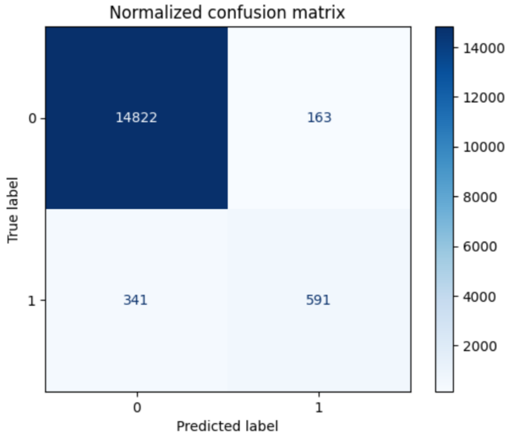


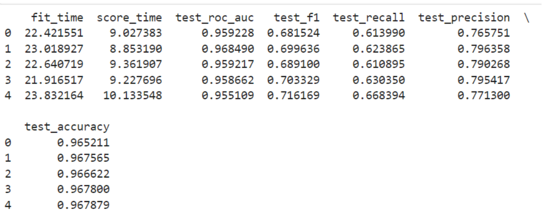


```
      fit_time  score_time  test_roc_auc   test_f1  test_recall  \
0  3758.732408   39.499516      0.967719  0.428356     0.374177
1   539.388182   13.314388      0.924068  0.409509     0.351779
2  1276.161884   11.326715      0.947270  0.446332     0.380263
3  1306.443120   11.691767      0.917827  0.394488     0.301713
4  2334.383302   24.085971      0.938629  0.410936     0.326746

   test_precision  test_accuracy
0        0.500882       0.993682
1        0.489908       0.993582
2        0.540187       0.994023
3        0.569652       0.994140
4        0.553571       0.994073
```

Fig. C8: 1 : 100 Support Vector Machine Results

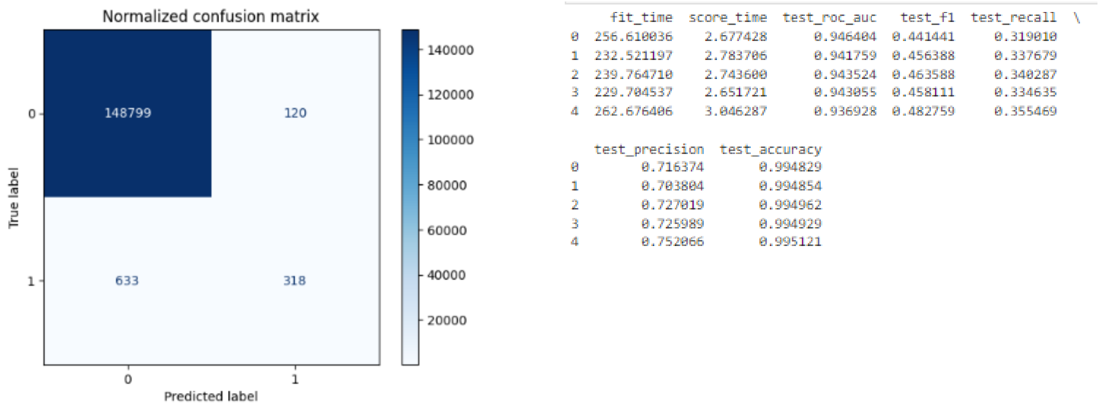




```
    fit_time  score_time  test_roc_auc   test_f1  test_recall  test_precision  \
0  22.421551    9.027383      0.959228  0.681524     0.613990        0.765751
1  23.018927    8.853190      0.968490  0.699636     0.623865        0.796358
2  22.640719    9.361907      0.959217  0.689100     0.610895        0.790268
3  21.916517    9.227696      0.958662  0.703329     0.630350        0.795417
4  23.832164   10.133548      0.955109  0.716169     0.668394        0.771300

   test_accuracy
0       0.965211
1       0.967565
2       0.966622
3       0.967800
4       0.967879
```

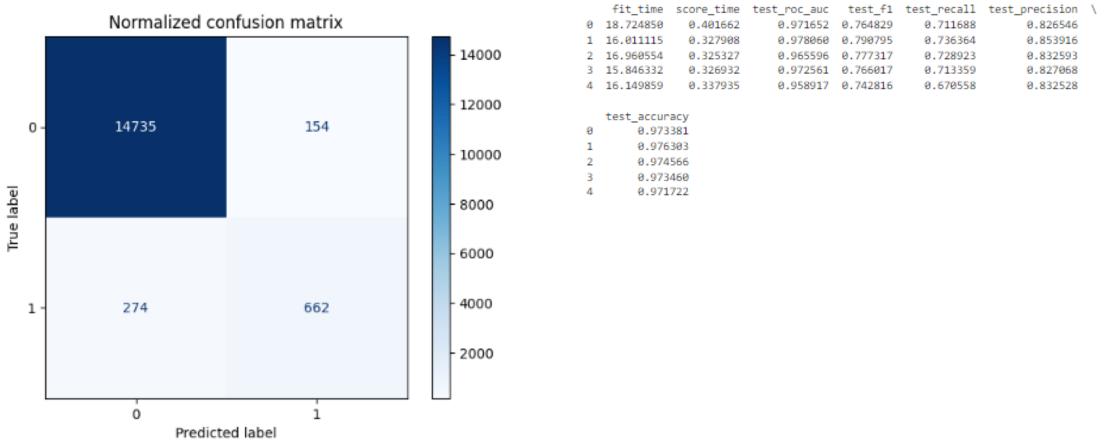
Fig. C9: 1 : 1000 Support Vector Machine Results



```
      fit_time  score_time  test_roc_auc   test_f1  test_recall  \
0  256.610036    2.677428      0.946404  0.441441     0.319010
1  232.521197    2.783706      0.941759  0.456388     0.337679
2  239.764710    2.743600      0.943524  0.463588     0.340287
3  229.704537    2.651721      0.943055  0.458111     0.334635
4  262.676406    3.046287      0.936928  0.482759     0.355469

   test_precision  test_accuracy
0        0.716374       0.994829
1        0.703804       0.994854
2        0.727019       0.994962
3        0.725989       0.994929
4        0.752066       0.995121
```

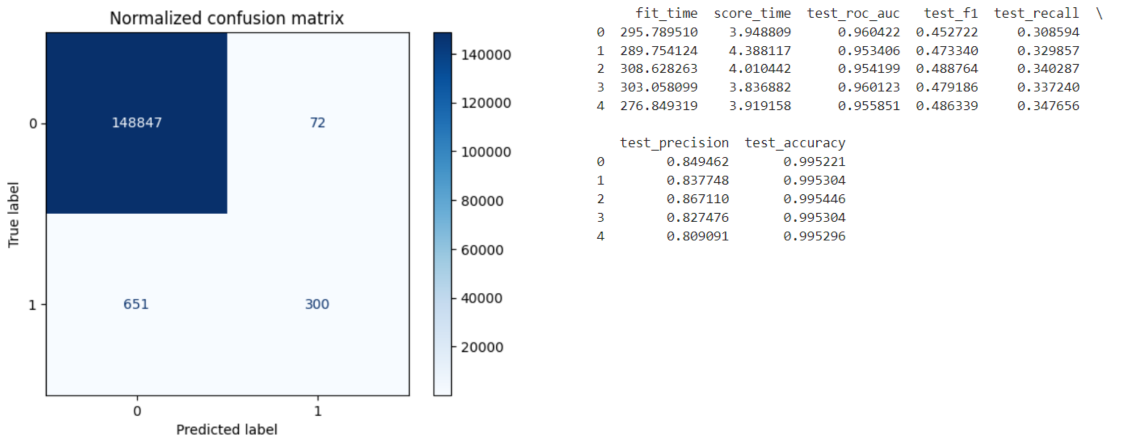
Fig. C10: 1 : 100 Top 10 Features Random Forest Results



```
    fit_time  score_time  test_roc_auc   test_f1  test_recall  test_precision  \
0  18.724850    0.401662      0.971652  0.764829     0.711688        0.826546
1  16.011115    0.327908      0.978060  0.790795     0.736364        0.853916
2  16.960554    0.325327      0.965596  0.777317     0.728923        0.832593
3  15.846332    0.326932      0.972561  0.766017     0.713359        0.827068
4  16.149859    0.337935      0.958917  0.742816     0.670558        0.832528

   test_accuracy
0       0.973381
1       0.976303
2       0.974566
3       0.973460
4       0.971722
```

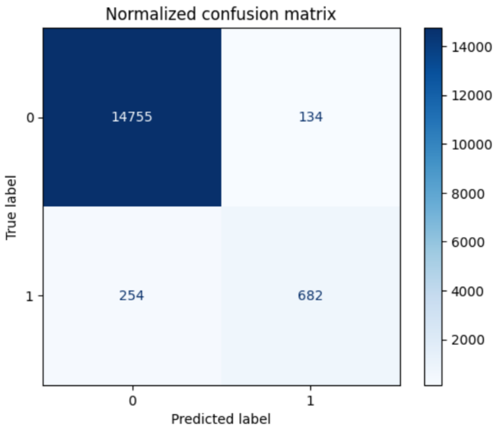
Fig. C11: 1 : 1000 Top 10 Features Random Forest Results

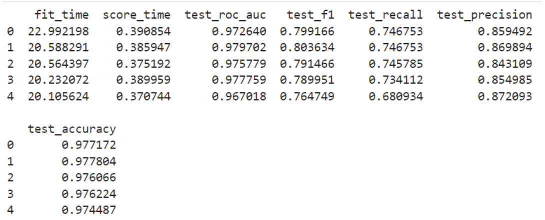


```
      fit_time  score_time  test_roc_auc   test_f1  test_recall  \
0  295.789510    3.948809      0.960422  0.452722     0.308594
1  289.754124    4.388117      0.953406  0.473340     0.329857
2  308.628263    4.010442      0.954199  0.488764     0.340287
3  303.058099    3.836882      0.960123  0.479186     0.337240
4  276.849319    3.919158      0.955851  0.486339     0.347656

   test_precision  test_accuracy
0        0.849462       0.995221
1        0.837748       0.995304
2        0.867110       0.995446
3        0.827476       0.995304
4        0.809091       0.995296
```

Fig. C12: 1 : 100 No Correlating Features Random Forest Results

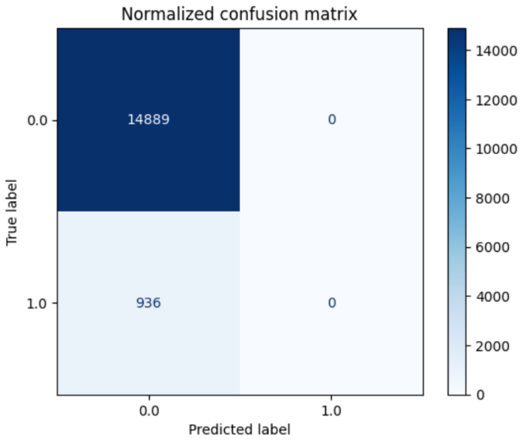


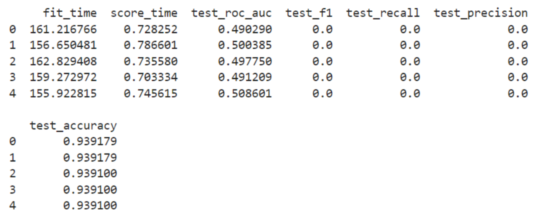


```
      fit_time  score_time  test_roc_auc   test_f1  test_recall  test_precision
0  22.992198    0.390854      0.972640  0.799166     0.746753        0.859492
1  20.588291    0.385947      0.979702  0.803634     0.746753        0.869894
2  20.564397    0.375192      0.975779  0.791466     0.745785        0.843109
3  20.232072    0.389959      0.977759  0.789951     0.734112        0.854985
4  20.105624    0.370744      0.967018  0.764749     0.680934        0.872093

   test_accuracy
0       0.977172
1       0.977804
2       0.976066
3       0.976224
4       0.974487
```

Fig. C13: 1 : 1000 No Correlating Features Random Forest Results





```
      fit_time  score_time  test_roc_auc  test_f1  test_recall  test_precision
0  161.216766    0.728252      0.490290      0.0          0.0             0.0
1  156.650481    0.786601      0.500385      0.0          0.0             0.0
2  162.829408    0.735580      0.497750      0.0          0.0             0.0
3  159.272972    0.703334      0.491209      0.0          0.0             0.0
4  155.922815    0.745615      0.508601      0.0          0.0             0.0

   test_accuracy
0       0.939179
1       0.939179
2       0.939100
3       0.939100
4       0.939100
```

Fig. C14: 1 : 1000 Random Values Random Forest Results

D SHAP Plots

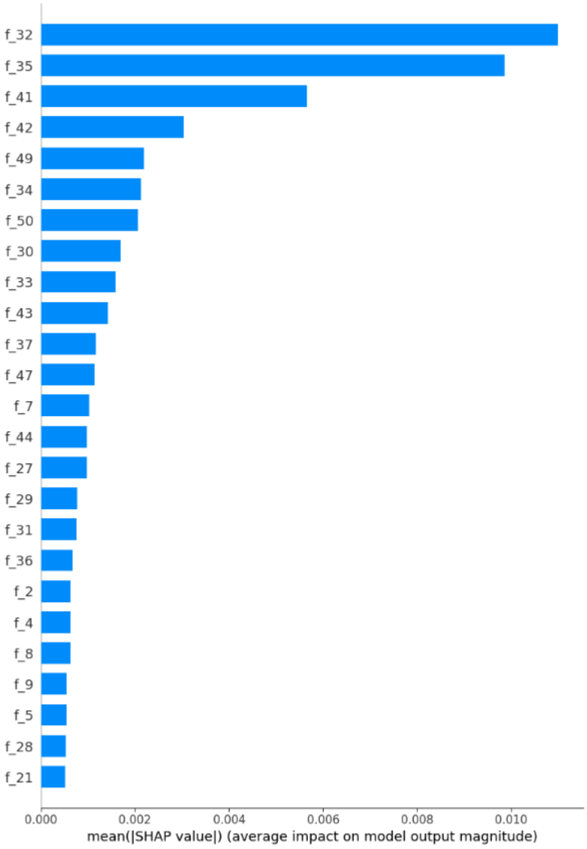


Fig. D15: Global Explanations: Bar Plot with Top 25 Features

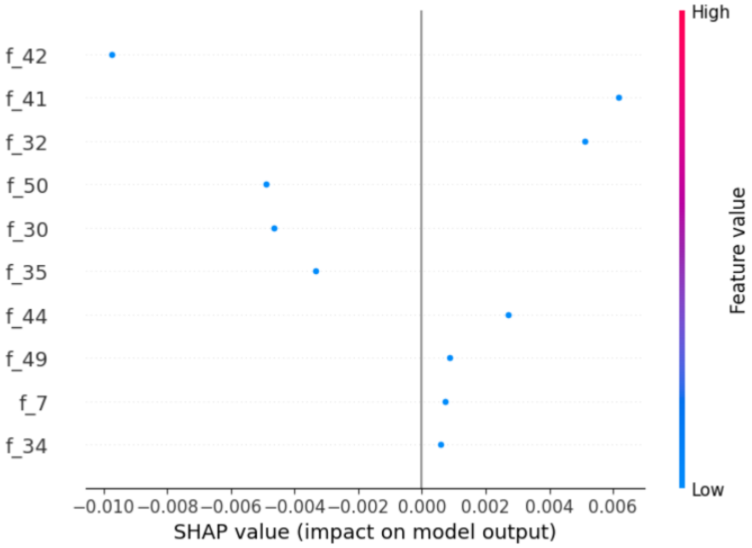


Fig. D16: Global Explanations: Beeswarm Plot with One Transaction

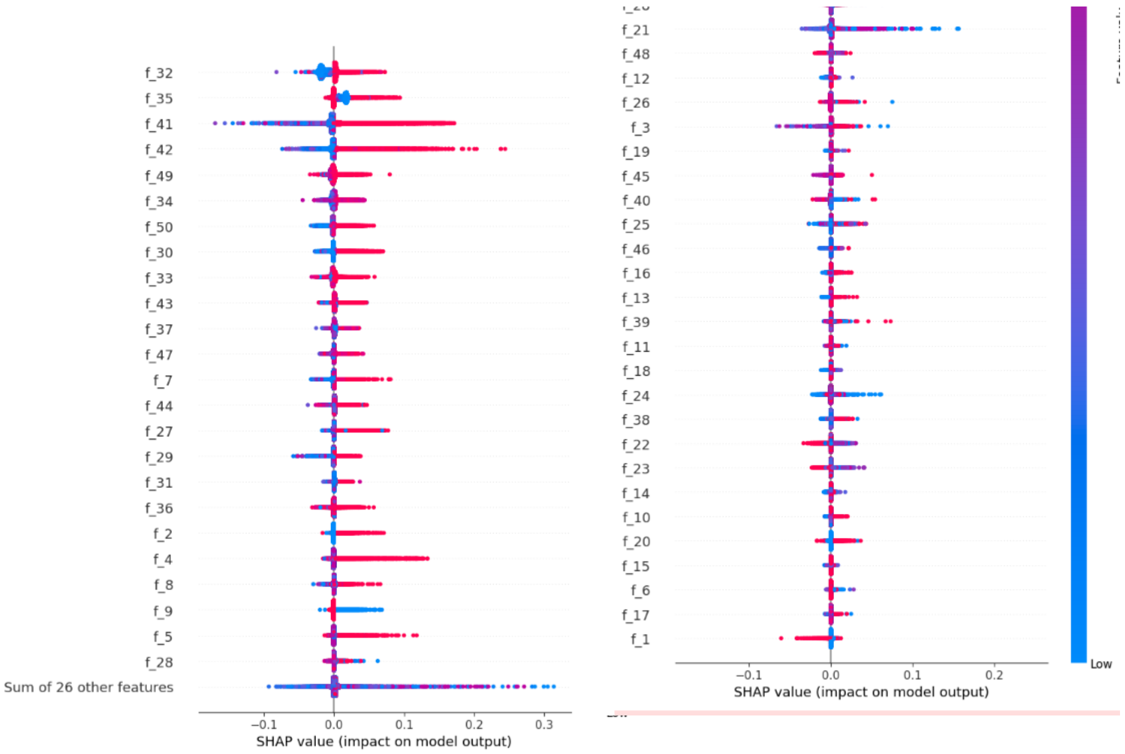


Fig. D17: 1:100 Beeswarm Plot With all 50 Features

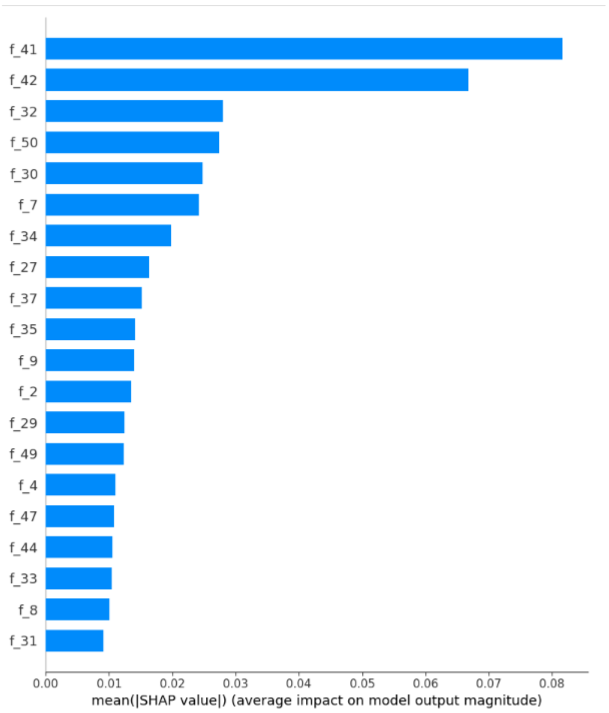


Fig. D18: False Positives SHAP Bar Plot

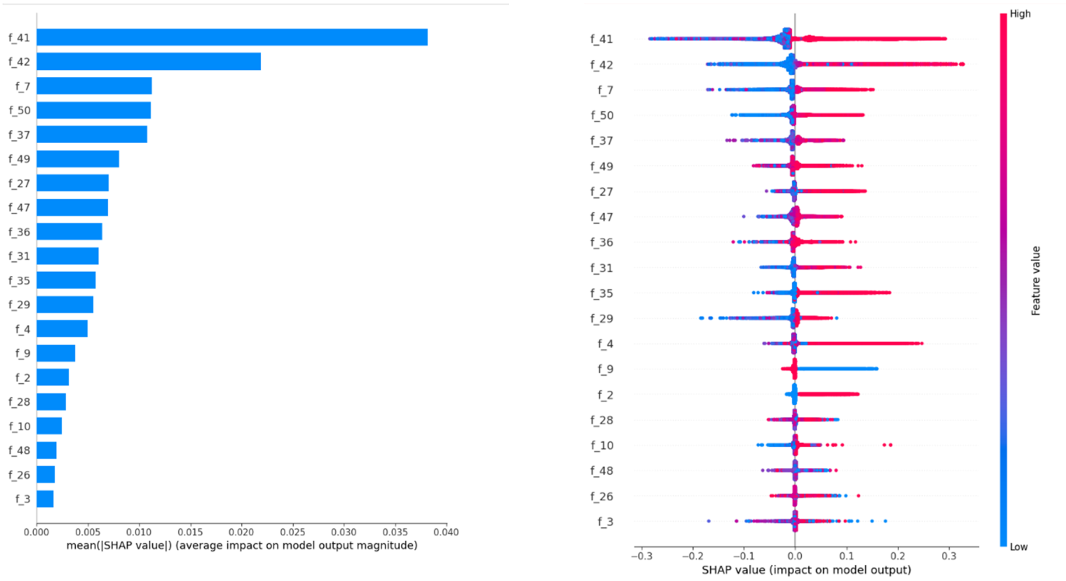


Fig. D19: 1 : 1000 No Correlation Features Only

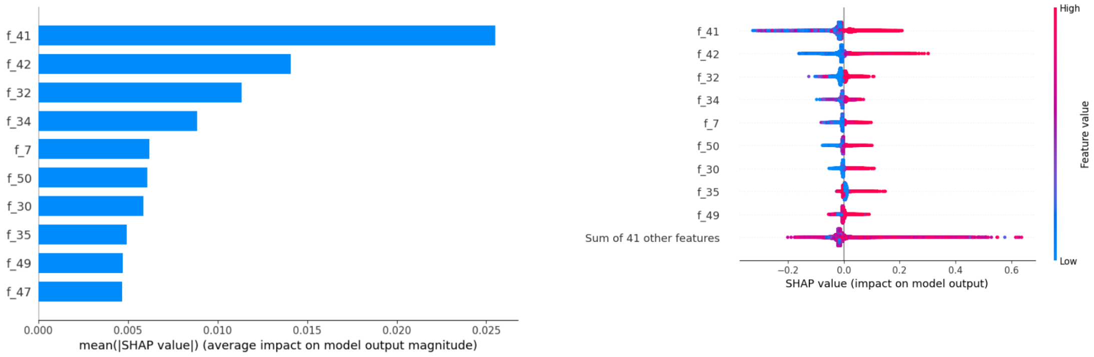


Fig. D20: SHAP Results 1:100 Test Dataset and 1:1000 Train Dataset

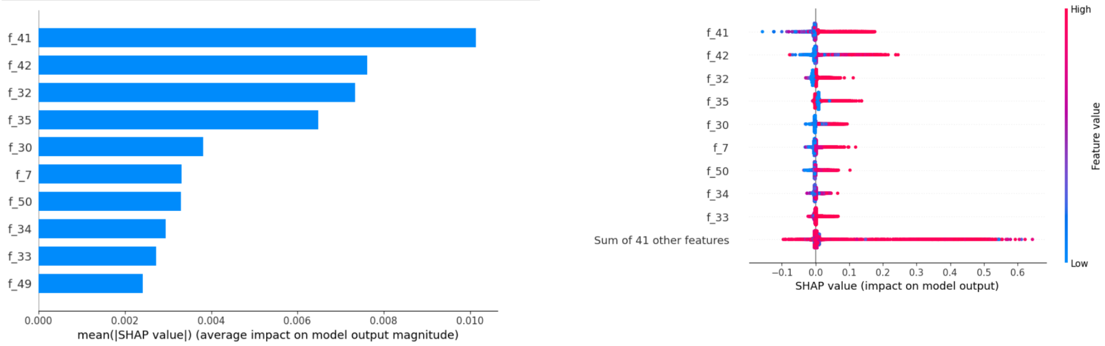


Fig. D21: SHAP Results 1:1000 Test Dataset and 1:100 Train Dataset

E XAI Model Results

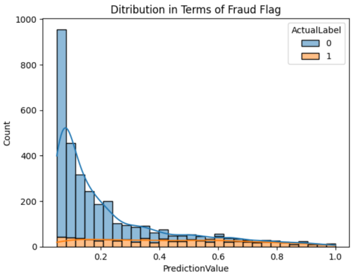


Fig. E22: 1:100 Top 10 Features Only Model Results for a Threshold of 0.05

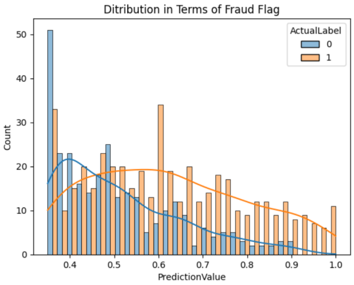


Fig. E23: 1:100 Top 10 Features Only Model Results for a Threshold of 0.35

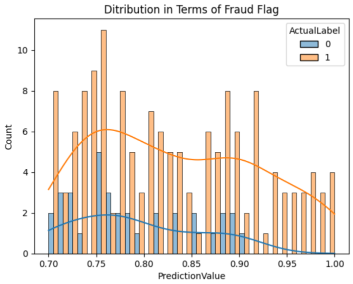


Fig. E24: 1:100 Top 10 Features Only Model Results for a Threshold of 0.7

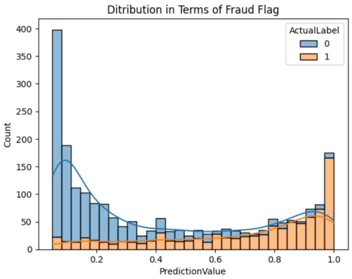


Fig. E25: 1 : 1000 Top 10 Features Only Model Results for a Threshold of 0.05

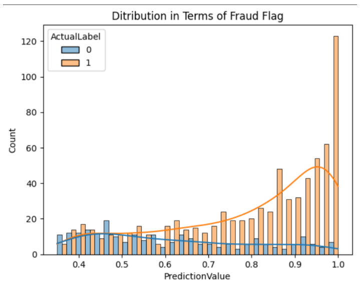


Fig. E26: 1 : 1000 Top 10 Features Only Model Results for a Threshold of 0.35

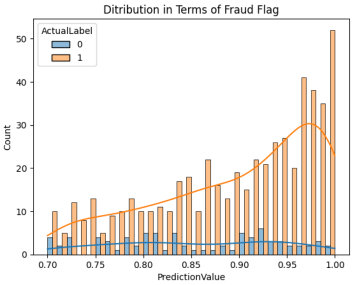


Fig. E27: 1 : 1000 Top 10 Features Only Model Results for a Threshold of 0.7

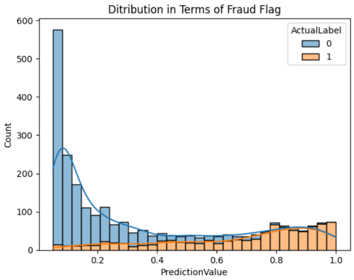


Fig. E28: 1 : 1000 Model Results for a Threshold of 0.05

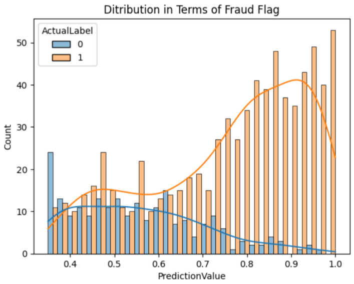


Fig. E29: 1 : 1000 Model Results for a Threshold of 0.35

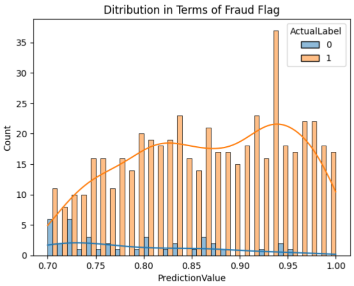


Fig. E30: 1 : 1000 Model Results for a Threshold of 0.7

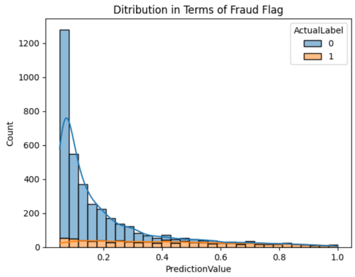


Fig. E31: 1 : 100 No Correlating Features Model Results for a Threshold of 0.05

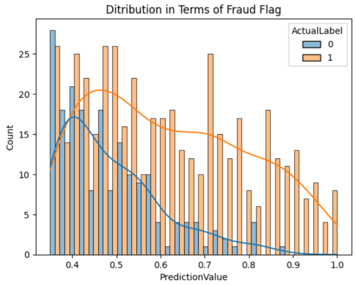


Fig. E32: 1 : 100 No Correlating Features Model Results for a Threshold of 0.35

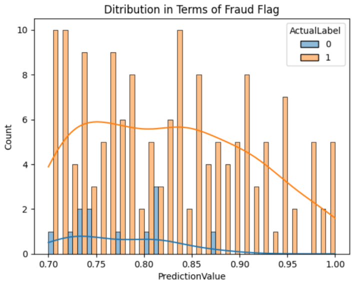


Fig. E33: 1 : 100 No Correlating Features Model Results for a Threshold of 0.7

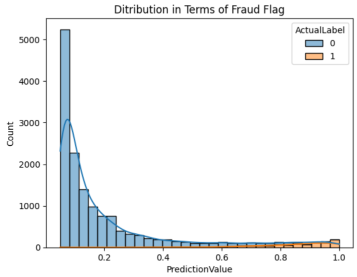


Fig. E34: 1:1000 Train Dataset and 1:100 Test Dataset Results for a Threshold of 0.05

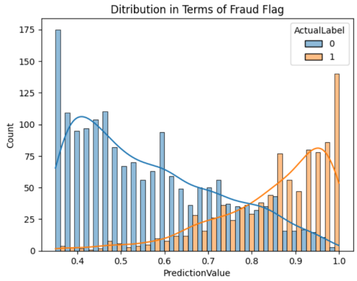


Fig. E35: 1:1000 Train Dataset and 1:100 Test Dataset Results for a Threshold of 0.35

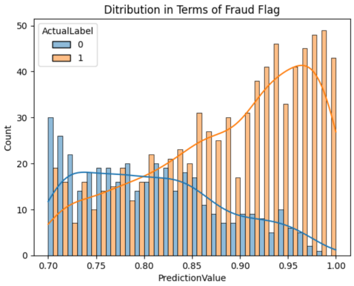


Fig. E36: 1:1000 Train Dataset and 1:100 Test Dataset Results for a Threshold of 0.7

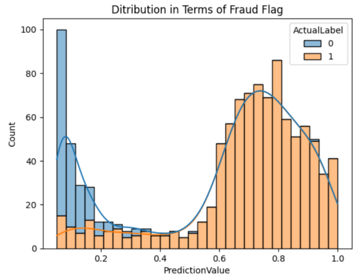


Fig. E37: 1:100 Train Dataset and 1:1000 Test Dataset Results for a Threshold of 0.05

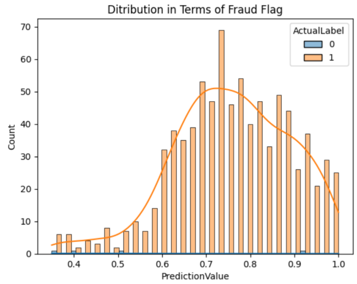


Fig. E38: 1:100 Train Dataset and 1:1000 Test Dataset Results for a Threshold of 0.35

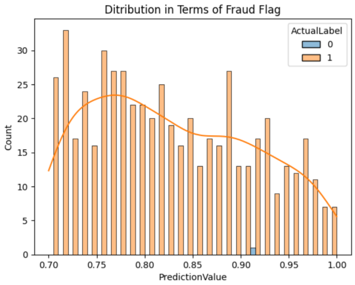


Fig. E39: 1:100 Train Dataset and 1:1000 Test Dataset Results for a Threshold of 0.7

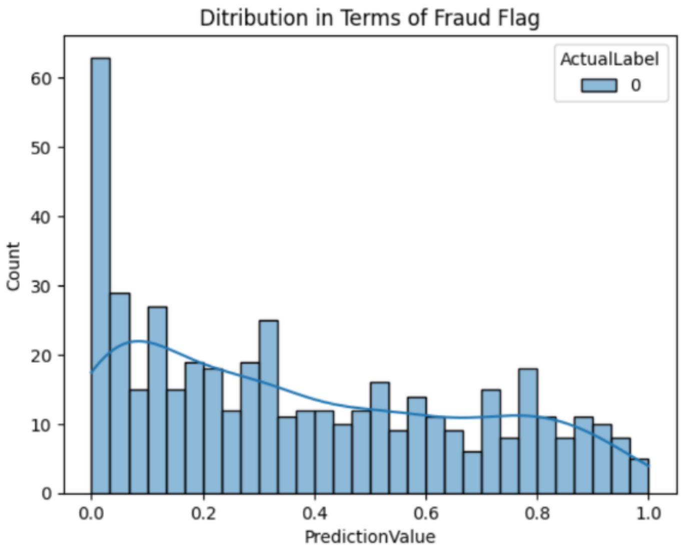


Fig. E40: False Positives XAI Model Results

F Robustness Tests

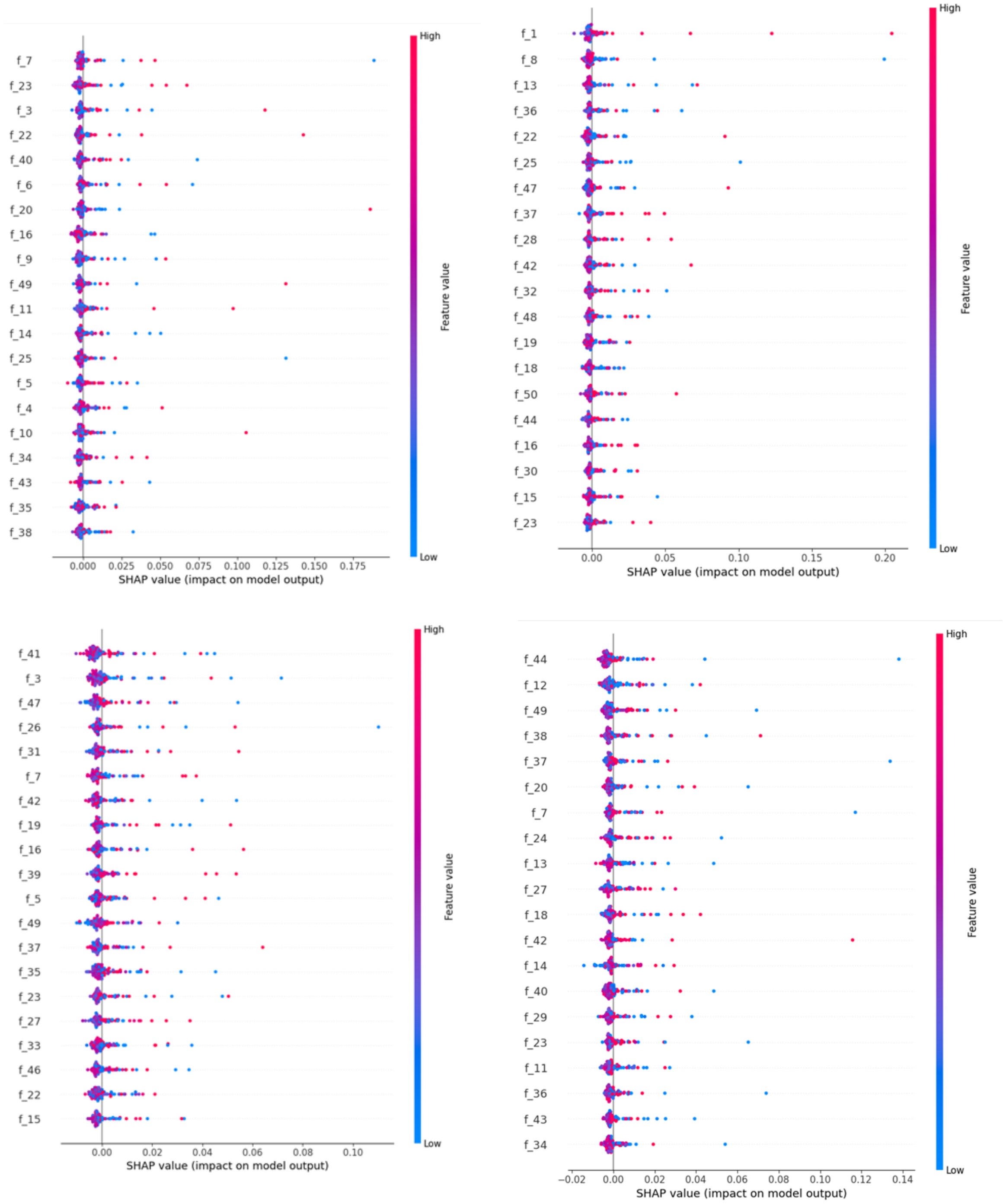


Fig. F41: Four different runs to test the models robustness

References

Abdul, A., Vermeulen, J., Wang, D. (2018). *Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda.* Department of Computer Science.

Achary, R., & Shelke, C. (2023). Fraud Detection in Banking Transactions Using Machine Learning. *2023 International Conference on Intelligent and Innovative Technologies in Computing.*

A.C.M. (2024). *Advancing Computing as a Science & Profession.* Retrieved from https://www.acm.org/

Acun, C., & Nasraoui, O. (2023). In-Training Explainability Frameworks: A Method to Make Black-Box Machine Learning Models More Explainable. *2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT.*

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box. *A Survey on Explainable Artificial Intelligence (XAI.* Computer and Interdisciplinary Physics Laboratory.

Ai, Q., & Narayanan.R, L. (2021). Model-agnostic vs. Model-intrinsic Interpretability for Explainable Product Search. *ACM*, ,

Aldughayfiq, B., Ashfaq, F., Jhanjhi, N.Z., Humayun, M. (2023, June). Explainable AI for Retinoblastoma Diagnosis: Interpreting Deep Learning Models with LIME and SHAP. *Diagnostics*, *13*(11), 1932, https://doi.org/10.3390/diagnostics13111932 Retrieved 2024-03-22, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10253103/

Arrieta, A., & Dıaz-Rodrıguez, N. (2019). *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies.* TECNALIA.

Balagopalan, A., & Zhang, H. (2022). *The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations.* Massachusetts Institute of Technology.

Barceló, P., Monet, M., Pérez, J., Subercaseaux, B. (2020). *Model Interpretability through the Lens of.* Institute for Mathematical and Computational Engineering.

Barrionuevo, G., Ramos-Grez, J., Walczak, M., Betancourt, C. (2021). Comparative evaluation of supervised machine learning algorithms in the prediction of the relative density of 316L stainless steel fabricated by selective laser melting. *The International Journal of Advanced Manufacturing Technology*, ,

Bau, D., & Gilpin, L. (2019). *Explaining Explanations: An Overview of Interpretability of Machine Learning.* Computer Science and Artificial Intelligence Laboratory.

Belharet, A., Bharathan, U., Dzingina, B. (2020). Report on the Impact of Artificial Intelligence on Project Management. *Management of Technology – Information System*, ,

Belotti, V. (2004). *Intelligibility and Accountability Human Considerations in Context Aware Systems.* Human-Computer Interaction.

Benrimoh, D., Israel, S., Fratila, R. (2021). *Editorial: ML and AI Safety, Effectiveness and Explainability in Healthcare.* Medicine and Public Health.

Bertossi, L., Kimelfeld, B., Livshits, E., Monet, M. (2023, August). The Shapley Value in Database Management. *ACM SIGMOD Record*, *52*(2), 6–17, https://doi.org/10.1145/3615952

.3615954 Retrieved 2024-03-21, from http://arxiv.org/abs/2401.06234 (arXiv:2401.06234 [cs])

Boiarskaia, E., Albert, N., Lee, D. (2019, May). *Detecting Financial Fraud at Scale with Decision Trees and MLflow on Databricks.* Retrieved from https://www.databricks.com/blog/2019/05/02/detecting-financial-fraud-at-scale-with-decision-trees-and-mlflow-on-databricks.html

Bruijn, H., Warnier, M., Janssen, M. (2021). *The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making.* Delft University of Technology.

Bukhori, H., & Munir, R. (2023). Inductive Link Prediction Banking Fraud Detection System Using Homogeneous Graph-Based Machine Learning Model. *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC).*

Bunt, A., Lount, M., Lauzon, C. (2012). *Are Explanations Always Important? A Study of Deployed, Low-Cost Intelligent Interactive Systems.* Computer Science Department.

Butin, A., & Markova, M. (2021). *The Safety oncept and a New Approach to.* Lipetsk State Technical University.

Carneiro, B., & Junior, R. (2010). *Identifying Bank Frauds Using CRISP-DM and Decision Trees.* International Journal of Computer Science and Information Technology.

Chari, S., Seneviratne, O., Gruen, D. (2020). *Explanation Ontology: A Model of Explanations for User-Centered AI.* Cambrdige: IBM Research.

Chen, J. (2024). *What Is Fraud? Investopedia.*

Chou, Y.-L., & Moreira, C. (2021). *Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms and Applications.* Information Fusion.

Chromik, M., Eiband, M., Buchner, F., Krüger, A., Butz, A. (2021, April). I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. *26th International Conference on Intelligent User Interfaces*, 307–317, https://doi.org/10.1145/3397481.3450644 Retrieved 2024-03-21, from https://dl.acm.org/doi/10.1145/3397481.3450644 (Conference Name: IUI '21: 26th International Conference on Intelligent User Interfaces ISBN: 9781450380171 Place: College Station TX USA Publisher: ACM)

Chung, E., & Zhang, L. (2024). *Improving Classification Performance With Human Feedback: Label a few, we label the rest.* Cornell University.

Cirqueira, D., Nedbal, D., Helfert, M., Bezbradica, M. (2020). Scenario-Based Requirements Elicitation for User-Centric Explainable AI. A. Holzinger, P. Kieseberg, A.M. Tjoa, & E. Weippl (Eds.), *Machine Learning and Knowledge Extraction* (pp. 321–341). Cham: Springer International Publishing.

Coma-Puig, B., & Carmona, J. (2021). *A Human-in-the-Loop Approach based on Explainability to Improve NTL Detection.* Universitat Politecnica de Catalunya.

Commission, E. (2019). *Ethics Guidelines for Trustworthy AI.* High-Level Expert Group on Artificial Intelligence.

Cortez, P., & Embrechts, M. (2011). Opening Black Box Data Mining Models Using Sensitivity Analysis. *IEEE Symposium on Computational Intelligence and Data Mining (CIDM.*

Crook, B., & Schluter, M. (2023). *Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI.* University of Bayreuth.

Cylynx (2020, December). *Machine Learning for Fraud Detection.* Retrieved from https://www.cylynx.io/blog/machine-learning-for-fraud-detection/

D.A.R.P.A. (2020). *Explainable Artificial Intelligence (XAI) (Archived.* Retrieved from https://www.darpa.mil/program/explainable-artificial-intelligence

Deloitte (2022). *A review of Explainable AI (XAI) concepts, techniques, and challenges.* Deloitte.

Demajo, L.M., Vella, V., Dingli, A. (2020, November). Explainable AI for Interpretable Credit Scoring. *Computer Science & Information Technology (CS & IT)* (pp. 185–203). AIRCC Publishing Corporation. Retrieved 2024-03-22, from https://aircconline.com/csit/papers/vol10/csit101516.pdf

Dong, C. (2017, August). *The evolution of machine learning.* Retrieved from https://techcrunch.com/2017/08/08/the-evolution-of-machine-learning

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning.*

Dreyer, M., & Achtiba, R. (2023). *Understanding the (Extra-)Ordinary: Validating Deep Model Decisions with Prototypical Concept-based Explanations.* Fraunhofer Heinrich-Hertz-Institute.

Fawcett, T. (2006, June). Introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874, https://doi.org/10.1016/j.patrec.2005.10.010

Ferdib-Al-Islam, Saha, A., Bristy, E.J., Rahatul Islam, M., Afzal, R., Ridita, S.A. (2023, July). LIME-based Explainable AI Models for Predicting Disease from Patient's Symptoms. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1–6). Retrieved 2024-03-22, from https://ieeexplore.ieee.org/document/10307223 (ISSN: 2473-7674)

Freitas, A. (2014). *Comprehensible Classification Models – a position paper.* School of Computing.

Ghorbani, A., & Wexler, J. (2019). Towards Automatic Concept-based Explanations. *NeurIPS*, ,

Ghorbani, Z., & Kazemi, A. (2022). *The effect of probiotic and synbiotic supplementation on lipid parameters among patients with cardiometabolic risk factors: a systematic review and meta-analysis of clinical trials.* Oxford University Press.

Gopavaram, S., & Vinothiyalakshmi, P. (2023). Cloud Based Credit Card Fraud Detection System in Banking Using Machine Learning and Deep Learning algorithms. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT).*

Gunning, D., & Aha, D. (2019). *DARPA's Explainable Artificial Intelligence Program. AI magazine.*

Hanae, A., Youssef, G., Saida, E. (2023). Analysis of Banking Fraud Detection Methods through Machine Learning Strategies in the Era of Digital Transactions. *2023 7th IEEE Congress on Information Science and Technology (CiSt.*

Hase, P., & Bansal, M. (2020). *Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?* UNC Chapel Hill.

Hashemi, S., Mirtaheri, S., Greco, S. (2023). *Fraud Detection in Banking Data by Machine Learning Techniques.* IEEE Access.

Hastie, T., Tibshirani, R., Friedman, J. (2008). *The Elements of Statististical Learning.* Springer.

Hayashi, Y., & Takano, N. (2020). *One-Dimensional Convolutional Neural Networks with Feature Selection for Highly Concise Rule Extraction fro Credit Scoring Datasets with Heterogeneous Attributes.* Department of Computer Science.

Hildebrandt, M., & Castillo, C. (2020).
FAT '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* Association for Computing Machinery.

Ho, W., Tang, B.-S., Wong, S. (2020). *Predicting property prices with machine learning algorithms.* Journal of Property Research.

Hoffman, R., Johnson, M., Bradshaw, J. (2013). *Trust in Automation.* Florida Institute for Human and Machine Cognition.

Hulsen, T. (2023). *Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare.* Department of Hospital Services & Informatics, Philips Research.

Hung, L.-P., Xu, C.-H., Wang, C.-S., Chen, C.-L. (2023). Applying the Shapley Value Method to Predict Mortality in Liver Cancer Based on Explainable AI. D.-J. Deng, H.-C. Chao, & J.-C. Chen (Eds.), *Smart Grid and Internet of Things* (pp. 133–143). Cham: Springer Nature Switzerland.

Ignatiev, A. (2020, July). Towards Trustable Explainable AI. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (pp. 5154–5158). Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization. Retrieved 2024-03-22, from https://www.ijcai.org/proceedings/2020/726

I.N.G. (2023a). *History.* Retrieved from https://www.ing.com/About-us/History.htm

I.N.G. (2023b). *ING at a glance.* Retrieved from https://www.ing.com/About-us/ING-at-a-glance.htm

Inspector General, O. (2021). *What is Fraud?* Retrieved from https://www.mass.gov/info-details/what-is-fraud

Janiesch, C., Zschech, P., Heinrich, K. (2021). Machine Learning and deep learning. *Faculty of Business Management & Economics*, ,

Johnson, J., & Khoshgoftaar, T. (2023). *Data-Centric AI for Healthcare Fraud Detection.* Florida Atlantic Universitiy.

Jolliffe, I.T., & Cadima, J. (2016, April). Principal component analysis: a review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *374*(2065), 20150202, https://doi.org/10.1098/rsta.2015.0202

Joseph, V.R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *15*(4), 531–538, https://doi.org/10.1002/sam.11583 Retrieved 2024-07-26, from https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11583 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sam.11583)

Koh, P., & Nguyen, T. (2020). Concept Bottleneck Models. *Proceedings of the 37th International Conference on Machine Learning.*

Korelc, J. (2020). Explainable taxonomy of AI, courtesy of DataScienceCentra. *DaZajn DaSeln k. d*, ,

Kortz, M., & Doshi-Velez, F. (2016). *Accountability of AI Under the Law: The Role of Explanation.* Science, Fiction and Philosopy.

Kubat, M. (2017). *An Introduction to Machine Learning.* Springer.

Kumar, A. (2022, May). *Model Complexity & Overfitting in Machine Learning.* Retrieved from https://vitalflux.com/model-complexity-overfitting-in-machine-learning/

Le, T.-T.-H., Prihatno, A.T., Oktian, Y.E., Kang, H., Kim, H. (2023, May). Exploring Local Explanation of Practical Industrial AI Applications: A Systematic Literature Review. *Applied Sciences*, *13*(9), 5809, https://doi.org/10.3390/app13095809 Retrieved 2024-03-21, from https://www.mdpi.com/2076-3417/13/9/5809

Lent, M., Fisher, W., Mancuso, M. (2004). *An Explainable Artificial Intelligence System for Small-unit Tactical Behavior.* IAAI EMERGING APPLICATIONS.

Lepri, B., & Oliver, N. (2017). *Fair, transparent and accountable algorithmic decision-making processes.* MIT Open Access Articles.

Lia, Z., & Huang, H. (2023). *Combining Reinforcement Learning and Barrier Functions for Adaptive Risk Management in Portfolio Optimization.* Department of Electrical and Electronic Engineering.

Liao, Q., Gruen, D., Miller, S. (2020). Questioning the AI: informing design practices for explainable AI user experiences. *Proceedings of the 2020Conference on Human Factors in Computing Systems (CHI.*

Lim, B. (2011). *Improving understanding, trust and control with intelligibility in context-aware applications.* Carnegie Mellon University.

Linardatos, P., & Papastefanopoulos, V. (2020). *Explainable AI: A Review of Machine Learning Interpretability Methods.* Entropy.

Lipton, Z. (2016). *The mythos of model interpretability.* Cornell University.

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017.*

Menant, L., & Gilibert, D. (2021). *The Application of Acceptance Models to Human Resource Information Systems: A Literature Review.* Frontiers in Psychology.

Merrick, L., & Taly, A. (2020, June). *The Explanation Game: Explaining Machine Learning Models Using Shapley Values.* arXiv. Retrieved 2024-03-21, from http://arxiv.org/abs/1909.08128 (arXiv:1909.08128 [cs, stat])

Miller, T. (2017). *Explanation in artificial intelligence: Insights from the social sciences.* School of Computing and Information Systems.

Minh, D., Wang, X., Nguyen, T. (2021). *Explainable artificial intelligence: a comprehensive review.* Springer.

Mohseni, S., & Pitale, M. (2019). *Practical Solutions for Machine Learning Safety in Autonomous Vehicles.* Texas A&M University.

Molnar, C. (2023). *Interpretable Machine Learning.* Retrieved 2024-03-22, from https://christophm.github.io/interpretable-ml-book/

Mueller, S., & Hoffman, R. (2018). *Metrics for Explainable AI: Challenges and prospects.*

Nakayama, Y., & Sawaki, T. (2023). *Causal Inference on Investment Constraints and Non-stationarity in Dynamic Portfolio Optimization through Reinforcement Learning.* Cornell University.

Narahari, Y. (2012, October). Game Theory. *Department of Computer Science and Automation*, ,

Pascua, J.A.A., Prado, A.J.A., Solis, B.R.B., Cid-Andres, A.P., Cambiador, C.J.B. (2019, September). Trends in fabrication, data gathering, validation, and application of molecular fluorometer and spectrofluorometer. *Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy*, *220*, 116837, https://doi.org/10.1016/j.saa.2019.02.061

Poeta, E., & Ciravegna, G. (2023). *Concept-based Explainable Artificial Intelligence: A Survey.*

Polydoros, A., & Nalpantidis, L. (2017). Survey of Model-Based Reinforcement Learning: Applications on Robots. *Journal of Intelligent and Robotic Systems*, ,

Porayska-Pomsta, K., & Rajendran, G. (2019). *Accountability in Human and Artificial Intelligence Decision-Making as the Basis for Diversity and Educational Inclusion.* Artificial Intelligence and Inclusive Education.

Powers, D. (2008, January). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Mach. Learn. Technol.*, *2*, ,

Psychoula, I., Gutmann, A., Mainali, P., Lee, S.H., Dunphy, P., Petitcolas, F.A.P. (2021, May). *Explainable Machine Learning for Fraud Detection.* arXiv. Retrieved 2024-03-20, from http://arxiv.org/abs/2105.06314 (arXiv:2105.06314 [cs])

Purohit, K., & Vats, S. (2023). Improvement in K-Means Clustering for Information Retrieval. *Proceedings of the Fourth International Conference on Electronics and Sustainable Communication Systems (ICESC-2023.*

Qin, Z., & Liu, Y. (2022). *Explainable Graph-based Fraud Detection via Neural Meta-graph Search. Chinese academy of sciences.*

Ramos, F. (2023, November). Legal technology: A comprehensive analysis of their impact on legal industry and enhancing entrepreneurial opportunities. *Derecho Global. Estudios sobre Derecho y Justicia*, *9*(25), 367–385, https://doi.org/10.32870/dgedj.v9i25.701 Retrieved 2024-03-14, from http://www.derechoglobal.cucsh.udg.mx/index.php/DG/article/view/701

Rao, S.X., Zhang, S., Han, Z., Zhang, Z., Min, W., Chen, Z., ... Zhang, C. (2021, November). xFraud: Explainable Fraud Transaction Detection. *Proceedings of the VLDB Endowment*, *15*(3), 427–436, https://doi.org/10.14778/3494124.3494128 Retrieved 2024-03-20, from http://arxiv.org/abs/2011.12193 (arXiv:2011.12193 [cs])

Rawal, A., & McCoy, J. (2022). Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives. *IEEE TRANSACTIONS ON ARTIFICIAL INLIGENCE.*

Ribeiro, M., & Singh, S. (2016). *"Why Should I Trust You?" Explaining the Predictions of Any Classifier.* University of Washington.

Ribeiro, M.T., Singh, S., Guestrin, C. (2018, April). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), , https://doi.org/10.1609/aaai.v32i1.11491 Retrieved 2024-03-21, from https://ojs.aaai.org/index.php/AAAI/article/view/11491 (Number: 1)

Rosenfeld, A., & Richardson, A. (2019). *Explainability in human-agent systems,"Autonomous Agents and Multi-Agent Systems".*

Rudin, C. (2019). *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.* Duke University.

Saeed, W., & Omlin, C. (2022). *Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities.* School of Computer Science and Informatics.

Sah, S. (2020). *Machine Learning: A Review of Learning Types.* Creative Commons CC BY.

Saiful Islam, M. (2023). *A rule-based machine learning model for financial fraud.* International Journal of Electrical and Computer Engineering (IJECE.

Saraswat, D., & Bhattacharya, P. (2022). *Explainable AI for Healthcare 5.0: Opportunities and Challenges.* Department of Computer Science and Engineering.

Sarker, I. (2021). *Machine Learning: Algorithms, Real-World Applications and Research Directions.* SN Comput Sci.

Sawada, Y., & Nakamura, K. (2022). *Concept Bottleneck Model with Additional Unsupervised Concepts.* Tokyo Research Center.

Schröer, C., & Kruse, F. (2021). *A Systematic Literature Review on Applying CRISP-DM Process Model.* (Place: Centeris)

SHAP (2024, July). *bar plot — SHAP latest documentation.* Retrieved 2024-07-25, from https://shap.readthedocs.io/en/latest/example$_notebooks/api_examples/plots/bar.htmlGlobal-bar-plot$

Shokri, R., Stronati, M., Song, C. (2017). *Membership Inference Attacks Against Machine Learning Models.* Cornell University.

Skibińska-Fabrowska, I. (2023, May). Demand for Cash and its Determinants - a Post-Crisis Approach [1]. *Journal of Central Banking Theory and Practice*, *12*(2), 103–131, https://doi.org/10.2478/jcbtp-2023-0016 Retrieved 2024-03-14, from https://www.sciendo.com/article/10.2478/jcbtp-2023-0016

Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H. (2020, February). *Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods.* arXiv. Retrieved 2024-03-22, from http://arxiv.org/abs/1911.02508 (arXiv:1911.02508 [cs, stat])

Speith, T. (2022). *A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods.* Association for Computing Machinery.

Strumbelj, E., & Kononenko, I. (2010). An Efficient Explanation of Individual Classifications using Game Theory. *Journal of Machine Learning Research*, *11*, 1–18,

Suriyanarayanan, C. (2020). Anomaly detection using machine learning techniques. *Malaya Journal of Matematik*, ,

Swartout, W., & Moore, J. (1993). Explanation in Second Generation Expert Systems. *Conference paper*, ,

Synetics (2017, December). *The Evolution of Machine Learning.* Retrieved from https://smdi.com/the-evolution-of-machine-learning

Testart, C., Fruchter, N., Gilpin, L. (2018). Explaining explanations to society. *NIPS Workshop on Ethical, Social and Governance Issues in AI.*

Tibshirani, S., & Friedman, H. (2008). The Elements of Statitical Learning. *Springer Series in Statistics*, ,

Tritscher, J., & Wolf, M. (2023). *Evaluating Feature Relevance XAI in Network Intrusion Detection.* (Series: Communications in Computer and Information Science book series)

van der Pol, S. (2023, December). *ING introduction interview.*

van der Pol, S. (2024, March). *XAI Taxonomy Master Thesis.* Retrieved 2024-03-21, from https://app.diagrams.net/G1Y7mxGSGWbKW3qcH$_{n}Ywlk4huzVkgFQkj$

Villaronga, E., & Kieseberg, P. (2018). Humans forget machines renember: Artificial Intelligence and the right to be forgotten. Scholarly Commons at Boston University School of Law.

Wang, J., & Biljecki, F. (2022). *Unsupervised machine learning in urban studies: A systematic review of applications. Cities: the international journal of urban policy and planning.*

Wang, Y., Cheng, D., Liu, X. (2019, February). Matrix expression of Shapley values and its application to distributed resource allocation. *Science China Information Sciences*, *62*(2), 22201, https://doi.org/10.1007/s11432-018-9414-5 Retrieved 2024-03-21, from http://link.springer.com/10.1007/s11432-018-9414-5

Weerts, H. (2019). *Interpretable Machine Learning as Decision Support for Processing Fraud Alerts.* Department of Mathematics and Computer Science Data Mining Research Group.

Weld, D., & Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Communications of the ACM*, ,

Y, S., & Challa, M. (2023, June). A Comparative Analysis of Explainable AI Techniques for Enhanced Model Interpretability. *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)* (pp. 229–234). Retrieved 2024-03-22, from https://ieeexplore.ieee.org/abstract/document/10266190

Yang, W., Wei, Y., Wei, H. (2023). *Survey on Explainable AI: From Approaches, Limitations and Applications Aspects.* Human-Centric Intelligent Systems.

Zhao, Y., & Wang, Y. (2022). *Fairness and Explainability: Bridging the Gap Towards Fair Model Explanations.* Vanderbilt University.

Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A. (2021, March). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, *10*(5), 593, https://doi.org/10.3390/electronics10050593 Retrieved 2024-07-24, from https://www.mdpi.com/2079-9292/10/5/593