Machine Learning with Applications 10 (2022) 100406

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

# Increasing trust and fairness in machine learning applications within the mortgage industry

W. van Zetten *, G.J. Ramackers, H.H. Hoos

*LIACS, Leiden University, Niels Bohrweg 1, Leiden, 2333 CA, The Netherlands*

ARTICLE INFO

*Keywords:*
Explainable artificial intelligence
Interpretable machine learning
Fair artificial intelligence
Bias detection
Fraud detection

ABSTRACT

The integration of machine learning in applications provides opportunities for increased efficiency in many organisations. However, the deployment of such systems is often hampered by the lack of insight into how their decisions are reached, resulting in concerns about trust and fairness. In this article, we investigate to what extent the addition of explainable AI components to ML applications can contribute to alleviating these issues. As part of this research, explainable AI functionality was developed for an existing ML model used for mortgage fraud detection at a large international financial institution based in The Netherlands A system implementing local explanation techniques was deployed to support the day-to-day work of fraud detection experts working with the model. In addition, a second system implementing global explanation techniques was developed to support the model management processes involving data-scientists, legal experts and compliance officers. A controlled experiment using actual mortgage applications was carried out to measure the effectiveness of these two systems, using both quantitative and qualitative assessment methods. Our results show that the addition of explainable AI functionality results in a statistically significant improvement in the levels of trust and usability by its daily users. The explainable AI system implementing global interpretability was found to considerably increase confidence in the ability to perform the processes focused on compliance and fairness. In particular, bias detection towards demographic groups successfully aided in the identification and removal of bias towards applicants with a migration background.

## 1. Introduction

Artificial Intelligence systems are being deployed in a broad range of application domains, increasing the scope of AI applications and the degree to which they affect our daily lives. These systems play important roles across various sectors, including healthcare, education, entertainment and finance. AI systems, or more specifically machine learning (ML) software, can achieve high precision in difficult prediction tasks where humans are unable to detect the patterns required to solve a given problem.

### 1.1. Problem statement

A trade-off in the deployment of these high-performance ML systems is that often they are so complicated that the exact decision making process followed by the system is unclear, making it very difficult for humans to understand why a specific result was obtained. Several problems arise from this inability to examine the decision making process. Firstly, it is difficult to validate that outcome of the process is defensible, in that it follows a compelling train of thought. Secondly, it is difficult to determine whether the decisions thus made are unfairly biased towards certain demographics. Methods that can help gain insight into such opaque decision making processes and ensure that systems make fair decisions are subject to extensive research. Learning from the decision making process could enhance knowledge of the target domain, and ensuring systems are unbiased and follow sensible decision processes is a requirement for the ethical application of machine learning, and AI systems in general.

One frequently used domain for application of machine learning is in estimating the risk of fraud. A landmark case is the Dutch system *Systeem Risico Indicatie* (SyRI), which linked personal information from a large number of governmental databases in an effort to identify potential fraudulent individuals. The precise requirements for a person to be marked as potentially fraudulent, combined with the fact that being marked as a potential fraudster had great impact on the individual, caused the system to be prohibited in 2020, after having been in use for six years (ANP, 2020).

Insurers and banks use ML to estimate the fraud risk associated with loan applicants, for both personal loans as well as mortgages. Organisations applying these systems want to ensure that the systems do not suffer from unintended bias, helping them adhere to anti-discrimination

* Corresponding author.
*E-mail addresses:* wessel@vanzetten.eu (W. van Zetten), g.j.ramackers@liacs.leidenuniv.nl (G.J. Ramackers), hh@liacs.nl (H.H. Hoos).

https://doi.org/10.1016/j.mlwa.2022.100406
Received 20 May 2022; Received in revised form 25 July 2022; Accepted 29 August 2022
Available online 5 September 2022

laws. Furthermore, these organisations must refrain from making automated decisions based on predictions obtained from the ML system. This guideline, as well as others, stem from the Ethical Framework for Insurers, a guidance framework setup by the Dutch Association of Insurers based on recommendations by the High-Level Expert Group on Artificial Intelligence advising the European Commission (Verbond van Verzekeraars, 2021).

### 1.2. Hypotheses

In this article, we present an extensive implementation and evaluation case study in cooperation with a large Dutch insurer, henceforth referred to as "the organisation". The organisation offers life and non-life insurance as well as mortgages, amongst many other products and services. It uses ML fraud risk systems in different areas, including mortgage and insurance. The organisation is looking for ways to further improve its fair use of AI, in order to adhere to the obligations set forward in the Ethical Framework. In order to achieve this, it would like to have a set of tools that help data scientists to better understand and explain the models they use, to validate the decision making process, and investigate and demonstrate the fairness of a given model. Furthermore, the organisation requires methods to explain the individual outcomes of a classification model, so that the employees working with those models on a daily basis are able scrutinise the particular decision making process. Allowing employees to gain insight into the decision making process in the case they are handling enables them to draw knowledge from the model and thus supports their own investigation. Also, as the organisation is ultimately responsible for any decision made using ML systems, it is crucial that employees and managers have the ability to understand the predictions made by ML systems.

In order to validate that the method of developing prototype systems and work instructions proposed in this research helps organisations implement interpretable and unbiased ML systems, we set up the following hypotheses.

1. Techniques that allow for individual predictions to be interpretable and transparent improve trust, satisfaction and usability of ML tools with their daily users.
2. Techniques that allow for ML tools to be globally interpretable and demonstrably free of discriminatory bias enable organisations to streamline internal processes concerning fair and balanced AI.

### 1.3. Methodology

Our study presented here applies research by design, by validating two prototype AI systems developed jointly with the organisation. By researching existing techniques in expanding domain of xAI research, we were able to select a set of techniques and combine them into prototype systems and work instructions that enabled the organisation to achieve the goals described above. We investigate how the possible techniques identified fit into existing AI processes at the organisation, how effective these techniques are, and what should be considered for their implementation.

We developed two prototype systems. The first of these focuses on implementing operational explainability, granting daily users insight into the individual predictions obtained from a model. In literature, this is known as local interpretability. This first system will henceforth be referred to as MLX1. It does not offer bias detection, merely interpretability. Our second system focuses on bias detection, granting insight into factors contributing to the overall behaviour of an ML model, and testing a given model for fairness. This allows for a model to be audited after development, before it is put into production. This is known as global interpretability and bias detection in the XAI literature. This system will be called MLX2. Together, these systems allow an organisation to gain insight into the behaviour of their ML models on a micro- and macro-level, as well as to ensure that an ML system is demonstrably free of bias.

MLX1 was validated in close cooperation with a group of 11 mortgage application reviewers, who work with an ML model assessing mortgage applications on a daily basis. A short survey measuring trust, usability and perceived performance was performed. MLX2 was validated with a second survey, in cooperation with data scientists, specifically focusing on the use of the system in their development process, and also with officers from Legal, Compliance and Risk departments, in order to evaluate how the system and insights gained from using it can help streamline the acceptance processes involved in every new and existing ML model in use within the organisation.

The result of our study presented here is an evaluation of applicable techniques for the given research problem, and implementation considerations in a practical scenario.

## 2. Background

The background for this work is determined in part by the business context, namely rules and regulations to which the organisation must adhere, as well as ethical guidelines to be met. Furthermore, research in the field of interpretable machine learning has produced a range of techniques that can in principle be used by our prototype systems.

### 2.1. Research in XAI

In recent years, research in the field of interpretable machine learning and bias detection has seen enormous growth. Some of this research focuses on the effect of interpretable AI in its application areas and the interactions with its users, studying the impact on user trust and model adaptation (Doran et al., 2018; Hoffman et al., 2018). Another sub-field of research focuses on developing techniques to gain insights into machine learning models (Adadi & Berrada, 2018; Mehrabi et al., 2021). It is further subdivided into work on developing machine learning models that are interpretable by design, or creating methods to develop a separate layer that enables the interpretability of black-box machine learning models. Our work focuses on this last area, identifying model-agnostic techniques for interpreting black-box models and for detecting bias. The reason for investigating model-agnostic techniques, as opposed to model-specific techniques, is that we aim to develop systems that are applicable to all kinds of models rather than being limited to a certain type of ML models.

Interpretable machine learning is a subfield of explainable AI, with many different surveys aiming merely to give an overview of inherently interpretable models *vs* black-box models (Rai, 2020), while others dive more deeply into the different views and perspectives associated with explainable AI (Doran et al., 2018). Yet other publications investigate the impact explainable models have their users and people impacted by the decisions produced by them (Shin, 2021). These review papers offer several taxonomies of applications, but lack explanations of existing relevant techniques. Roscher et al. (2020) specifically highlight the applications of interpretable machine learning for scientific research, categorising existing scientific research and applications of interpretable ML, covering mostly model-specific examples.

Based on surveys by Adadi and Berrada (2018) as well as Guidotti et al. (2019) combined with the reviews listed above, we established that in order to fully satisfy the interpretable machine learning aspect of our two prototype systems described later, these would have to provide both local and global interpretability. Local interpretability includes techniques that help grant insight into the considerations made by the model in a specific case, often listing the most important features that lead to a certain model output (e.g., classification result). Global interpretability techniques focus on building an overall understanding of the decision making process of the model, examining which features are most important and most often used when judging new cases.

Mehrabi et al. (2021) review different types of bias and fairness definitions. They consider two aspects of detecting bias in machine learning: bias in data and bias in the model. Training any model with an unbalanced or biased data set will produce a model that is also biased. Therefore, it is important to put into place techniques to explore the data set and identify bias. Secondly, in instances where the data set does not contain bias, the trained model can still display bias towards certain features, for example gender or race. Caton and Haas (2020) provide an overview of the different approaches to detecting and mitigating bias, as well as increasing fairness; we used this as a starting point to explore possible techniques to detect bias in both data and model.

### 2.2. Relevant techniques

Based on review papers mentioned in the previous section, we compiled a short-list of promising techniques fitting the business context and requirements.

#### 2.2.1. Techniques for local interpretability

Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) was one of the earlier local techniques. LIME learns an interpretable model around each specific prediction. LIME is a linear approach, determining the importance score of each individual feature and whether the feature has a positive or negative impact on the final prediction. Non-linear (or rule-based) approaches can take into account the combined effects of features on individual predictions. It is possible that one feature alone does not impact the prediction enough to flip it in case of a binary prediction task, but a group of features together might. An example of such a technique is Anchor (Ribeiro et al., 2018). An 'Anchor' is an explanation that anchors a local prediction, so that changes to the other feature values do not influence the outcome. Because the technique uses rules, it is able to combine features and is more faithful compared to LIME, as risk often stems from the combination of features as opposed to a single feature.

Another technique which uses rules is Local Rule-based Explanations (LORE) (Guidotti et al., 2018). Like Anchor, it uses rules combining features to explain local predictions. As an improvement over Anchor, it also introduces counterfactual rules, which are rules that would result in the opposite classification. LORE typically achieves higher accuracy and coverage than LIME and Anchor (Guidotti et al., 2018).

One of the better known techniques for explaining machine learning models is SHAP (Lundberg & Lee, 2017), which uses a key concept from cooperative game theory called the Shapley value. SHAP attempts to explain the prediction of a specific instance by computing how much each feature contributed to that prediction, in the presence of combinations of other features. Computing Shapley values for all predictions results in a matrix of values, which can be combined to explain the entire model, thus providing global interpretability (Lundberg & Lee, 2017; Molnar, 2020).

#### 2.2.2. Techniques for global interpretability

SHAP (Lundberg & Lee, 2017) can be considered the best known technique for model explainability, and can also be used for to estimate global feature importance. As with local interpretability, it is possible that some features have an impact on the global model that only becomes apparent when combined with a different feature. An interesting technique to determine both the overall importance of features, as well as whether they impact other features indirectly, proposes iteratively obscuring features in the training data set to certain degrees to assess the impact on the test set. This technique, called BlackBox-Auditing (Adler et al., 2018), can be used for global interpretability to gain an overview of the importance ranking of the features in the data set, as well as determine whether a protected feature is still impacting the model through its effect on secondary features, through monitoring the drop in error rate achieved by the model.

Besides techniques for global interpretability that produce feature importance weights or rankings, or deliver rules determining the models behaviour, there is also the approach of using interactive visualisation to capture the model. The *What if*? tool allows the user to probe, visualise and analyse systems without too much coding (Wexler et al., 2020). This way, users can inspect model behaviour in different scenarios, and build their own understanding of global model behaviour.

#### 2.2.3. Techniques for bias detection

Besides exploring techniques for local and global interpretability, we also investigate existing techniques to determine bias in a model. For this, we must look at both bias in the data and bias in the resulting model, since training a model with biased data usually also yields a biased model.

Aequitas (Saleiro et al., 2018) focuses on data exploration as well as model output, and might be applicable as its visualisation properties make it attractive to both data scientists and officers from Legal, Risk and Compliance departments. Given the nature of the system and its role in facilitating fair and unbiased decision making, simple visualisations are an attractive property. Aequitas mostly focuses on bias detection and visualisation based on specific demographic groups. The library does not directly interface with the model, but uses cases predicted by the model together with the ground truth (the known, correct classification of given cases). Fairness measures, for example false positive rate parity, are calculated and visualised in small diagrams, displaying the groups in categories with their respective fairness measure parity relative to the reference group.

It is possible that, even though a feature is not used in the training of the model, it still impacts the eventual bias. In the case that 'protected' features, such as sex or race, are present in the training data, the resulting model could be biased for these groups. BlackBoxAuditing (Adler et al., 2018), which has been briefly discussed in the subsection on global interpretability, might also be used to ensure there is no bias with regards to 'protected' features. By obscuring protected features and evaluating model performance on the test set, it is possible to determine whether the protected features are used in the model.

### 2.3. Ethical considerations

There are several ethical aspects involved in this research, stemming from both the industry in which the organisation operates as well as the nature of deploying a machine learning model in any industry. The deployment of machine learning models brings with it ethical issues, specifically concerning human oversight and the potential impact of an automated action based on the outcome of a machine learning model.

#### 2.3.1. Automated machine learning decisions

The organisation handles many applications a day. This causes the organisation to strive for efficiency in handling these applications, which may cause an individual mortgage application to receive less attention than ideal. This in turn may cause a valid, non-fraudulent application to be denied, because the reviewer simply does not have the required time to sift through all documents. In their search for efficiency, many organisations and industries are deploying machine learning models for classification tasks. In this paper, the classification model in use at the organisation judges an incoming mortgage application on its fraud risk. The implications of an incorrect decision either way are vast. In the case of a fraudulent application slipping through the machine learning model, a mortgage offer may be extended to the applicant, potentially resulting in financial losses for the organisation. However, in the event of a valid application being flagged as fraudulent by the machine learning model, the applicant will not receive a mortgage offer. The results of a wrong evaluation by the machine learning model have a bigger impact for the applicant than for the organisation.

In the situation above, a machine learning model is left unchecked by human oversight, meaning it can both cause financial damage to the organisation or deny a valid mortgage to an applicant. From an ethical aspect, the event of a valid mortgage being denied is unacceptable. Thus, we can conclude that simply deploying an automated machine learning model to evaluate all incoming mortgage application is ethically unacceptable, and that this task should be done by a qualified mortgage reviewer.

However, a balance may be struck between machine learning model and mortgage reviewer. A machine learning model for which individual decisions for applications can be transparent, and the decisions of the model are understandable, can advise a mortgage reviewer in their task. The reviewer has the final say in the evaluation of the application, but the information granted by the transparent decision of the machine learning model may assist the reviewer in making their own, informed decision. Care must be taken that this balance between advice from the model and the reviewers domain knowledge remains in favour of the reviewer, so that they keep the final say. This process of making the decisions of existing machine learning models understandable, and guarding the balance between model and reviewer is one of the tasks of MLX1.

#### 2.3.2. Ethical framework for insurers

The organisation is bound to the Ethical Framework, set up by the Dutch Association for Insurers, an organisation with which all large Dutch insurers are associated. The association represents the interests of all members, and aims to connect the insurance sector with societal developments. The Ethical Framework was set up in response to developments in European and Dutch legislation governing the use of AI applications, and its guidelines were based on existing guidelines by the High-Level Expert Group on AI advising the European Commission, who set up a document containing seven key requirements for trustworthy AI. These requirements for models also aim to address ethical concerns noted in Section 2.3.1, by ensuring meaningful human control.

The Ethical Framework sets forward 30 guidelines, grouped by seven key requirements set up by the High-Level Expert Group. The guidelines that are relevant for our work presented here are listed below, with their original number in parentheses, so that they can be easily found in the original Dutch version of the Ethical Framework (Verbond van Verzekeraars, 2021).

- **Technical robustness and security**
  - (7) The insurer ensures adequate quality of (training) data used for data-driven applications.
- **Privacy and governance**
  - (14) The insurer ensures that employees working with data-driven applications have received adequate training, specifically to avoid confirmation bias and to ensure human autonomy.
  - (15) The use of data-driven applications in production will always be subject to adequate human oversight.
- **Transparency**
  - (18) When employing data-driven applications, human intervention will always be possible, and explanations can be obtained by customers regarding the results of an application.
- **Diversity, non-discrimination and fairness**
  - (19) When infringement on fundamental rights, including unfair discriminatory bias in data-driven applications cannot be avoided, the insurer will not deploy the application.
  - (20) In deciding to use data-driven applications, the insurer considers diversity and inclusivity, especially regarding groups who are at risk of exclusion or disadvantage as a result of special needs.
- **Social well-being**
  - (21) The insurer will monitor the impact of employing data-driven decision making on groups of clients.
- **Accountability**
  - (23) The insurer will set up an internal control and accountability system for the use of AI applications and data sources.
  - (24) The insurer improves the knowledge of executives and internal auditors with regards to data-driven applications.
  - (25) The insurer ensures adequate internal communication on the use of data-driven applications.
  - (26) The insurer performs a risk and effect assessment with regards to the immediate stakeholders for each data-driven application.

The organisation has introduced processes and artefacts to control and evaluate systems before deploying them, as well as provided training to increase knowledge with employees working with models. This is further detailed in Section 2.5. The processes are the first of several significant steps envisioned by the organisation, the next step being the introduction of tools to improve the interpretability of existing models, and to provide the appropriate roles with understandable and clear metrics on the degree of bias in systems, so that they can be understood and audited before deployment. Following the development and evaluation of the systems, in the following, we also investigate to what degree the systems have enabled the organisation to improve the adherence to the guidelines presented earlier in this section.

### 2.4. Mortgage fraud at the organisation

The organisation uses a black-box model in the form of a tree ensemble to estimate the fraud risk associated with an incoming mortgage application. The model uses several basic properties of a mortgage application, as well as a number of specific indicators designed in cooperation with mortgage experts. In total, the version of the model used in this research uses 75 features per case. Cases are assigned a fraud risk score, and the top five cases are passed on to mortgage application reviewers for validation. Due to capacity issues, only five cases a day are passed on, from a pool of cases with scores exceeding an undisclosed threshold.

Currently, a rule-driven system generates an explanation for the five cases with the highest risk, before they are forwarded to the reviewers. This explanation is generated using 22 expert-constructed fixed rules, which trigger on specific values of individual figures, meaning the coverage of these rules does not match the 75 features used by the model. Consequently, it is possible that the model identifies a high-risk case in which the risk predominantly comes from a feature that is not covered by the 22 rules. The reviewers who are assigned a case usually spend a maximum of five minutes reviewing the generated explanation, and are obliged to comment on the different reasons and features which, according to the current explanation method, are the reason the particular case was assigned a high risk of fraud. In contrast, our MLX1 system allows for all 75 features used by the model to be present in the explanation passed for reviewers, and allows for the combining of features that jointly indicate a high risk of fraud.

### 2.5. Acceptance of ML models

The introduction of the requirement of a Project Initiation Document (PID) at the organisation is one of the reasons for the organisation's push for interpretable models. The PID covers technical and organisational aspects of every new machine learning model. It also includes sections on bias and discrimination, explainability and the ethical concerns involved in the application. It must be accepted by the Data Privacy Officer, Legal, Compliance and Risk departments before any steps to developing a new model or altering an existing model can be made. In the current situation, questions covering discrimination and explainability are answered based on a Record of Processing Activity (ROPA) on all data used by the model that is reviewed by the Compliance department. Specific internal guidelines and development methods are followed to ensure that the data used is free of bias. Furthermore, the PID must include a section on the desired decision making capabilities of the model. The data scientists currently use informal assessment and intuitive language to describe these decision making capabilities.

Therefore, the organisation expressed an interest in using automated systems to quantify possible bias and discrimination, as well as to gain insight into the decision making process. Furthermore, the organisation was interested in developing a way to systematically determine whether a given model is biased towards certain vulnerable demographic groups. Our MLX2 system was designed to support a generalised approach to including aspects of interpretability and bias in the Project Initiation Document. This enables the organisation to streamline internal processes concerning these aspects in the acceptance of new models.

## 3. Methods

This section covers the logical design of the two systems, MLX1 and MLX2, a description of how they fit into the business context and the model compliance processes at the organisation as described in the previous section. Furthermore, it describes the development of the systems, as well as a demonstration of their capabilities. Lastly, we outline the methods for validation of both systems.

### 3.1. Logical design

The system for local interpretability involves generating explanations for mortgage applications that have been assigned a higher fraud risk by the model. In Fig. 1, we visualise the two different paths for an application after it has been classified by the model. If an application is classified as a possible fraud risk, it is passed on to the method for generating an explanation, after which that explanation is passed on to a reviewer. The reviewer first interprets the explanation before assessing the mortgage application like they would normally, to ensure the model decision is verified independently. This ensures that the model decision has not been blindly copied, but taken into consideration by the reviewer after which they have made their own, informed decision. The other path shows the normal assessment by a reviewer, if the application has not been judged as a fraud risk. Based on their own independent research and assessment, the reviewer decides whether or not the application constitutes a fraud risk. If a fraud risk is identified by a reviewer, the mortgage application is passed onto the anti-fraud department, who do more in depth research into the application and the documents provided by the applicant. Fraud is rarely conclusively proven, as a result of the selective information an applicant can provide to the organisation. However, reasonable suspicion of fraud by the anti-fraud officer is a valid reason to reject the application at first, giving the applicant to appeal the rejection. Often times, applicants forgo this appeal, supporting the fraud suspicion but not conclusively proving it.

The MLX1 system simply replaces the current method for generating the explanation, which in the current situation is done using simple decision rules. It aims to give better insight into the decision made by the model, identifying the indicators that support the model's risk (or no risk) indication. Given that the reviewers are expected to spend roughly 5 min evaluating the model explanation, care must be taken to structure the explanation in such a way that this time limit can still be met.

Internal processes at the organisation concerning the approval needed to start new AI projects benefit from a system providing global interpretability and bias detection. MLX2 was designed to aid the responsible data scientists in demonstrating that their model is free of bias, and to provide an overview of the decision making process.

Currently, the results of the ROPA on all data used for the model are used to ensure that the developed model is bias free, but the organisation lacks tools for quantifying the degree of bias. Once a PID has been filled in by the responsible data scientist, it has to be approved by Compliance, Legal and Risk departments before the model can be developed. Any elements of the PID being unclear could lead to disapproval from one of the involved parties. Providing data scientists with tools to measure bias and including the results of this check in the PID could streamline the acceptance process for AI projects for all parties involved.

The set of MLX1 and MLX2 together offers the organisation two things. Firstly, MLX1 allows for local interpretability, giving the organisation the opportunity to grant mortgage reviewers insights into the decision making of the model for individual mortgage applications that they review. MLX2 allows for global interpretability and bias detection, enabling the organisation to test new and existing models for bias and gain overall insights into their behaviour.

### 3.2. Development of MLX1

For local interpretability, both Anchor (Ribeiro et al., 2018) and LORE (Guidotti et al., 2018) were tested, since both are non-linear and are able to capture the combined influence of multiple features on the local prediction. However, it was found that both techniques had a very long running times on individual test cases (approximately 4 min for Anchor, 11 min for LORE), making them unsuitable for use in the organisation's workflows.

In light of this, it was decided to use SHAP (Lundberg & Lee, 2017) for local interpretability instead, because being able to identify individual features that made a contribution to the result of a single local case is beneficial to the human reviewers working with the outcome of the model. The way SHAP can be used to see which features had a big or small influence on the outcome of a case grants these reviewers a starting point for their research. SHAP offers two possibilities, KernelSHAP and TreeSHAP. KernelSHAP is a model-agnostic, kernel-based estimation approach for Shapley values, whereas TreeSHAP is a more efficient, model-specific estimation approach for tree-based models. In the case of a tree-based model, where $T$ is the number of trees, $L$ is the number of leaves and $D$ is the maximum depth KernelSHAP has time complexity $\mathcal{O}(T \cdot L \cdot 2^M)$, whereas TreeSHAP has time complexity $\mathcal{O}(T \cdot L \cdot D^2)$ (Molnar, 2020).

In our experiments, we found that the running time for KernelSHAP to be able to produce local explanations for all cases in our test set was about 90 s, where TreeSHAP was able to produce explanations in just a few seconds. This difference in running time is not overly important, as SHAP only has to run once per batch, or once per day in this case. We wanted to ensure that KernelSHAP and TreeSHAP did not produce wildly differing importance rankings. To do this, we evaluated a number of cases with both methods, observing the features ranked in the top 5, as these are the features that would be used in the explanation of our MLX1 system. We found that, on average, the top 5 features produced by both methods contained 4 of the same indicators. In practice, this functional difference has no impact on the use of the system, since reviewers only act on the first few features.

Three variants of the MLX1 system were developed and presented to data scientists and mortgage application reviewers working for

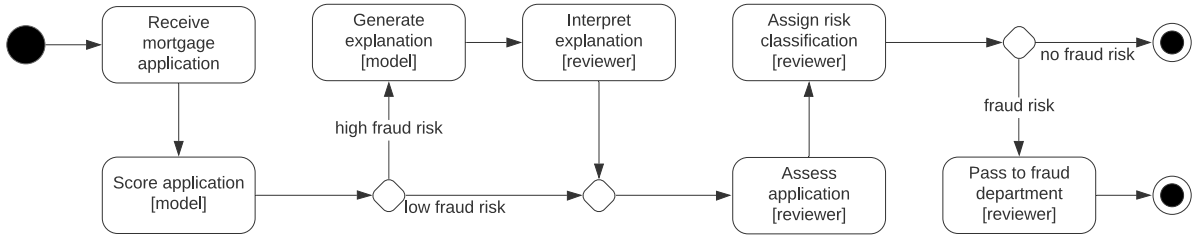


**Fig. 1.** Activity diagram showing how MLX1 fits into the process of determining the risk of an incoming mortgage application, by replacing the current method for generating an explanation of a high fraud risk prediction.

- High risk indicator(s):
  - The main applicant has been working for four months.
- Medium risk indicator(s):
  - The main applicant is 29 years old.
  - The requested loan amount is 298000 euros.
  - There are 70 days until the date of sale.
  - The *proprietary indicator* is *xxx*.

- High risk indicator(s):
  - The main applicant has **only** been working for four months.
- Medium risk indicator(s):
  - The main applicant is **only** 29 years old.
  - The requested loan amount is **above average, namely** 298000 euros.
  - There are **still** 70 days until the date of sale.
  - The *proprietary indicator* is **above average, namely** *xxx*.

**Fig. 2.** Output produced by variant 2 (left) and variant 3 (right) of MLX1. It shows the top five indicators explained and classified in two different impact groups. The risk score is withheld in these two variants. The subjectivity added by variant three is highlighted in bold.

the organisation. Based on their feedback, one variant was developed further.

We started by transforming the SHAP values into impact on the probabilities of our model, allowing us to rank features based on impact on the specific case. The first variant we proposed used this ranking of features and displayed the top 5 features with the highest impact on the case outcome, as well as the precise impact of each feature.

The second and third variant used understandable, textual explanations for each feature, in the form of written sentences describing the feature. The second variant used straightforward, objective explanations, while the third variant made use of formulations such as 'below average' and 'above average' thresholds set for each individual feature. These variants used three different impact groups to rank the features. The 'high impact' group contained features with an impact on the fraud probability that was 0.10 or higher, while 'medium impact' features had an impact between 0.10 and 0.05. Finally, 'low impact' features were those with an impact below the 0.05 threshold. The variants presented the top 5 highest-impact features, dividing them into the 'high impact' and 'medium impact' categories. The output of these two variants is shown in Fig. 2.

After presenting these three variants to the department manager of the reviewers and discussing them with data scientists, the organisation decided to continue with variant two, using the objective textual explanations. The department manager argued that variants one and three were too suggestive, and therefore undermined the validation task carried out by the human reviewers, according to the Ethical Framework (Verbond van Verzekeraars, 2021). The reviewers are tasked with objectively reviewing a case, and it was feared that showing them the precise impact on the probability of a fraud classification would influence them too much, and move them to copy the classification given by the model instead of performing their own objective, independent research. Variant 3 was rejected because of the thresholds used, and the fear that those thresholds might confuse the reviewers.

In our view, the subjective variant 3 offered more information than the eventually chosen, objective, variant 2. The search for possible fraudulent applications depends on detecting outlying cases, which have properties that differ from the average application. Therefore, when judging an application, it is important to have knowledge of the 'average' application.

Using a truncated data set, we were able to employ Anchor to generate explanations using combinations of rules. Using SHAP, we find the 15 globally most important features for the model. From discussions with reviewers, we concluded that using the top 15 features would be acceptable in the given business context, since in practice, the reviewers rarely use more than these top features. Then, we train a model (identical to the original model) on this data set. Finally, we use Anchor with this truncated model and data set to generate explanations for the cases. Using this approach, the Anchor running time per case is reduced to several seconds, versus four minutes per case using the full data set.

In another attempt to add grouping of features to MLX1, we manually created groups of features that are closely related. For example, there is a group containing ratio features describing the relations between house price, mortgage amount, renovation price, etc. If the added impact on the outcome of an individual case can be classified as a 'high impact' or 'medium impact', we present this grouping with a textual explanation as described earlier.

To summarise, MLX1 uses SHAP to provide the reviewers with textual explanation for features that have a certain impact on the classification probability of a case. This is further extended by adding groups of features that are closely related, to offer the reviewer a better overview on where to start if they decide to investigate a feature identified as risky by the MLX1 system. Lastly, with the addition of Anchor using a truncated model and data set, we offer the reviewers some insights into what combination of features impact the outcome of an individual case. Fig. 3 shows the explanation for a case used in the validation of the MLX1 system that was identified by the trained model as having a higher risk of fraud.

### 3.3. Validation of MLX1

For the validation of MLX1, we recruited the help of mortgage application reviewers within the organisation. We were able to source 11 reviewers for this task, due to an overall shortage of reviewers within the organisation. As a result, this was the maximum of reviewers that could be available us at the time of the survey. We presented the participating reviewers with two of the selected cases that had comparable explanations (in terms of length and detail), of which one was a fraud case and the other was not. The first case was accompanied by an explanation generated by the hand-made rules, while the second case had an explanation generated by the proposed approach. The order of the fraud and non-fraud cases for different reviewers was changed for each new reviewer. We asked each reviewer to consider the explanation for both cases as they usually do, and afterwards fill

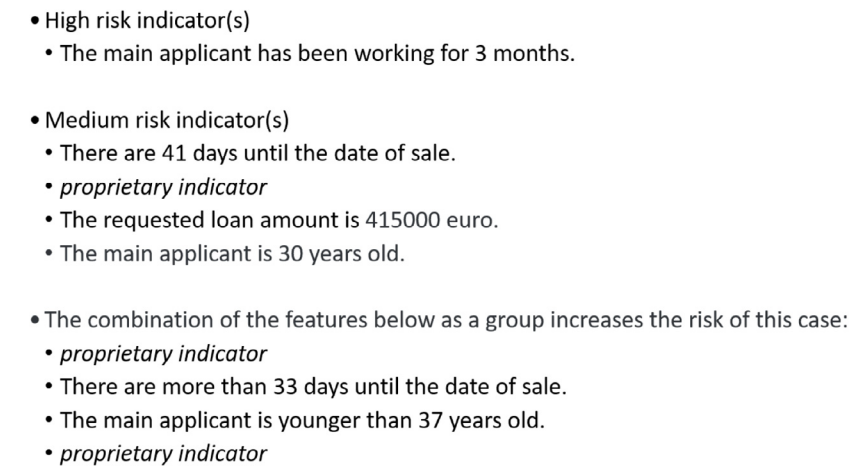


**Fig. 3.** Explanation generated by MLX1 for a mortgage application with elevated fraud risk, but no determined fraud. The SHAP component of MLX1 results in the indicators grouped by high and medium risk, whilst the Anchor component is responsible for the combination of features.

in a short survey consisting of eleven questions, followed by a short (5 min) oral interview for additional feedback.

As mentioned in Section 3.1, fraud is rarely conclusively proven. As a result, validating MLX1 by investigating whether the explainable AI system enables reviewers to more frequently make correct decision is not a possibility. Furthermore, the mortgage reviewers using MLX1 are the first line in fraud detection, and as mentioned in Section 3.1 pass on an application to the anti-fraud department when there is a fraud suspicion. However, it is possible to qualitatively measure whether the new MLX1 system is an improvement over the existing rule-based explanation.

Based on research by Hoffman et al. (2018) concerning metrics for explainable AI, we designed questions to measure explanation satisfaction and trust, as well as the reviewers' opinions on their own perceived speed and accuracy. We also added two specific questions covering the grouping of features and the Anchor component we developed. The relevance of these metrics is explained below.

- **Trust**
  Trust in the model provided by the model was measured using a Trust Scale distilled from existing research (Hoffman et al., 2018) and uses three questions to assess the confidence of the users in the model, and whether they feel the model is predictable, reliable, efficient and believable.
- **Explanation Satisfaction**
  This is defined as the 'degree to which users feel that they understand the AI system or process being explained to them.' (Hoffman et al., 2018). For this, research by Hoffman et al. (2018) also provides a list of questions, of which three were used.
- **Performance**
  The performance of the human reviewers is also important. Ideally, working with an interpretable system will improve the performance of the reviewers. To validate this, we used two performance metrics specific to the fraud use case.
  - **Speed** – Does the explanation help the reviewer identify the important aspects of a case quicker? The sooner the reviewer identifies potentially fraudulent aspects of a case, the quicker the case is handled.
  - **Accuracy** – Does the reviewer feel more confident in making a decision on a case thanks to the explanation?

These three groups of metrics were further developed leveraging research by Hoffman et al. (2018) and statements with which the participant could fully disagree or fully agree on a five-point Likert scale (Likert, 1932) were set up. Immediately after a reviewer completed the survey, we conducted the short interview by asking four qualitative questions, the first three of which focused on opinions that they may not have felt able to express in the survey questions. The fourth question focused on the reviewers' experience with the Anchor explanation, and how they felt about combining different features that together indicate a risk of fraud, for example a group containing ratio features describing the relations between house price, mortgage amount, renovation price, etc.

### 3.4. Development of MLX2

Given the requirements for global explanation and bias detection within the organisation, several libraries were chosen to be included in MLX2. The *What If?* tool was selected in order to give data scientists a visual overview of the model performance on a given test set. BlackBoxAuditing was included to discover potential bias with regards to protected features, such as nationality. Finally, Aequitas was selected to be able to test a model for (demographical) bias.

We set out to apply all three libraries to the model for mortgage fraud risk. The *What If?* tool could be applied and executed without any issues. For BlackBoxAuditing, we were forced to correct some mistakes in the source code, which used depreciated Python 2.7 techniques that were incompatible with the model code, which was in Python 3. Finally, getting Aequitas to work on our data also posed no significant challenges.

Below, we briefly describe our use of each of the three libraries selected for the MLX2 system.

#### 3.4.1. BlackBoxAuditing

The output generated by BlackBoxAuditing has two main parts. Firstly, the library generates a simple ranking of features based on their importance, measured by how much the Balanced Classification Rate (BCR), which denotes the balanced accuracy score, changes when each feature is removed from the model. This outputs a simple list of features. Secondly, BlackBoxAuditing allows the data scientist to list features that they consider to be 'protected', meaning features that might contain sensitive personal information. BlackBoxAuditing is then used to determine whether the Balanced Classification Rate (BCR) changes less than a certain threshold when omitting the protected features from the model. This threshold is set by the organisation. If the observed change is below the threshold, this indicates the information contained in the protected feature is also present in other features used by the model, effectively meaning the information in the protected feature is actively used by the model to make predictions.


**Fig. 4.** The first pane of the *What if?* tool, visualising a predicted data set. The data scientist sees a scatter plot containing all predicted data points, as well as their ground truth. They can slice or zoom in to certain areas or categories, and use the left pane to alter the predicted case and rerun the classification.

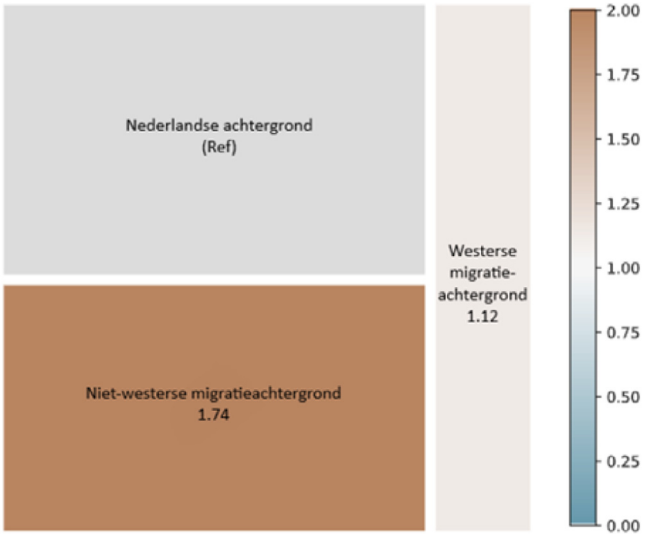


**Fig. 5.** A hypothetical Aequitas output of a test comparing FPR for Dutch, non-western, and western migration backgrounds.

### 3.4.2. What If? tool

The *What If?* tool is an interactive library that allows the data scientist to visualise the results of their model, and to slice, aggregate and zoom in on different groups within features. The data scientist can use this library to visualise the distribution of features in their data set. Using the library, it is also possible to view confusion matrices for specific features, or the overall confusion matrix. Lastly, it offers the ability to adjust properties of an individual case, and run it through the model again. The visualisation pane of the library is shown in Fig. 4.

The use of this library in relation to the MLX2 system is threefold. Firstly, the data scientist can use it to visualise the performance of their model, and observe whether it is behaving as expected, also for certain groups in the data set. Secondly, the library might be of use when investigating specific cases, for example when a client requests the model decision of their application (fulfilling guideline 18 of the Ethical Framework). Thirdly, the library can simply be of use to the data scientist in obtaining an overview of the features in the data set, along with their distributions.

### 3.4.3. Aequitas

Aequitas is the statistical library included in MLX2. We have chosen to enrich the data set with aggregated demographic data gathered by the Dutch Central Bureau of Statistics (Centraal Bureau voor de Statistiek, or CBS), which contains information on the ratios of inhabitants of The Netherlands who have a migration background per area of 100 by 100 m. MLX2 allows the user to select both the feature they want to investigate, as well as the fairness measure, for example false positive rate (FPR) disparity, which shows how the FPR of different groups within a feature compare to the largest group. The addition of data on migration backgrounds means we can investigate whether the tested model is fair with regards to this background, or whether it is unfairly inclined to indicate applications from non-western areas as fraudulent. A mock-up visualisation was created for internal evaluation purposes, and is displayed in Fig. 5.

Aequitas can be used to test a model for bias with regards to migration background, as a result of the CBS data. This allows data scientists as well as executives to gain insights into the fairness of the model they are evaluating for production, helping fulfil guidelines 19 and 20 of the Ethical Framework, as well as providing extra information for the PID.

## 3.5. Validation of MLX2

In order to validate the three libraries contained in MLX2, we demonstrated their use by applying them to the mortgage fraud model, and presenting their results and workings to several different functional user groups. In the first session, we presented the technical and application side of the libraries, mainly the *What if?* tool and BlackBoxAuditing, and how these libraries could help data scientists understand and prove their models. We collected several points of feedback from this session, which we then processed into a short questionnaire that was answered by the participating data scientists. In the second demonstration session of the MLX2 system, we demonstrated the use of Aequitas to executives from the Legal, Risk and Compliance departments, including the Data Privacy Officer connected to the team. We focused on the combination of Aequitas with CBS data on migration backgrounds, allowing for the model to be tested on false positive rate (FPR) disparity based on migration background through indirect bias.

**Table 1**
Median degree of agreement for all statements covering trust in the model, explanation satisfaction and whether the proposed explanation would help the reviewer complete their task.

| Statement | Median degree of agreement |
|---|---|
| Trust in the model | |
| 1 - I trust the explanation. I feel like the model is working well. | Agree |
| 2 - I like using the explanations to make decisions. | Agree |
| 3 - I feel like I will make the correct decision only using this explanation. | Neutral |
| Explanation satisfaction | |
| 4 - The explanations make me understand how the model reaches its judgement. | Agree |
| 5 - I am satisfied with the explanation. | Agree |
| 6 - The explanation is sufficiently detailed. | Agree |
| 7 - The explanation on how the model works seems sufficient. | Neutral |
| 8 - The explanation of the result tells me how accurate the model is. | Agree |
| Performance | |
| 9 - I reach a decision quicker because the case has an explanation. | Agree |
| 10 - I am able to make a better informed decision because the case has an explanation. | Agree |
| 11 - I find the addition of combinations of risk indicators to the explanation important. | Agree |

## 4. Results

In this section, we present the results of the survey held with mortgage application reviewers, followed by a discussion and evaluation of the first hypothesis. Then, we summarise the findings gathered from the two demonstration sessions with data scientists and Legal representatives, leading to an evaluation of the second hypothesis. Finally, we discuss how the two systems we have developed help the organisation improve their adherence to the relevant guidelines from the Ethical Framework.

### 4.1. Survey results for MLX1

MLX1 implements SHAP and Anchor, two approaches to improve local interpretability for mortgage reviewers working with the model estimating fraud risk for mortgage applications on a daily basis. Previously, a set of 22 fixed rules was used to generate an explanation for a given case and to model risk prediction. In order to evaluate the effectiveness of MLX1, 11 reviewers were presented with a short survey measuring the trust, satisfaction and perceived performance they achieved using MLX1. In Table 1 we present the median degree of agreement for all statements, divided in the three groups as discussed before (trust, explanation satisfaction, impact on performance). Fig. 6 shows the underlying responses to all statements.

Besides the statements presented to all reviewers, shown in Table 1, we also presented two randomly selected reviewers with a case in which we manually grouped several related factors and requested their opinion in the same way as above, in terms of agreement on a five-point scale. One reviewer answered that this grouping of features did not help her at all (*Completely disagree*), while the second reviewer answered that this was one of the more helpful features of the MLX1 explanation (*Completely agree*). This difference in agreement could indicate an interesting direction for future research.

After presenting the participating reviewers with the survey, we conducted a quick interview to enable them to provide any feedback they could not give us in the survey. Below, we list the most common feedback received in these sessions.

- **Insights** (Trust)
  Reviewers reported that the extra insights into the decision process of the model increased the trust they had in the model.
- **Clarity** (Explanation satisfaction)
  Several reviewers noted that some of the fraud indicators that express a ratio (e.g., the ratio between the house price and estimated renovation costs) were explained poorly, and that they did not know what these features meant exactly. This could be addressed by rewording the explanation in cooperation with reviewers, and to include domain terms in the text.
- **Missing information** (Explanation satisfaction)
  Some features might be unknown to the organisation. However, reviewers reported that they did not know how to handle the 'unknown' indicators. From domain knowledge, sometimes missing information can be a good fraud indicator, however reviewers were confused as to the meaning of 'unknown'. 'Unknown' indicators are simply indicators that have no value, because background data needed to form the indicators is not available, for example if there is no co-applicant for the mortgage.
- **Overview** (Performance)
  Ten out of eleven reviewers were positive about the additional information provided by the explanations generated by MLX1. While stressing the importance of their own independent assessment, they indicated that MLX1 provides them with much better guidance as to what to investigate first, compared to the current explanation method, thus improving the quality of their assessment and saving time.

*H1: Techniques that allow for individual predictions to be interpretable and transparent improve trust, satisfaction and usability of ML models with their daily users.*

Fig. 6 shows the degree of agreement on a Likert scale, therefore making the questions Likert-type data. It is common practice to analyse the statements by observing the median degree of agreement indicated (Jamieson, 2004; Sullivan & Artino, 2013). We separate this investigation by the three metrics that are part of the hypothesis: trust, explanation satisfaction and perceived performance.

*Trust.* Following the overall answers to the first three statements, we conclude that the trust in the model increased with the use of the explanation method employed by MLX1. This could simply be due to the fact that MLX1 offers far more detail than the existing approach, which leads reviewers to believe it is more knowledgeable than the current explanation. This might mean the reviewers are more eager to put their trust in the model decision. Hence, care should be taken to ensure that reviewers do not take the model decision for granted solely based on the level of detail that MLX1 provides, falsely believing that this extra detail means the model decision is the right one. Reviewers might be more inclined to blindly follow the classification by the model, undermining the human oversight that is required by the Ethical Guidelines.

*Explanation satisfaction.* Overall, the responses to the statements covering explanation satisfaction can be viewed as a positive outcome, as user satisfaction was increased when compared to the current method

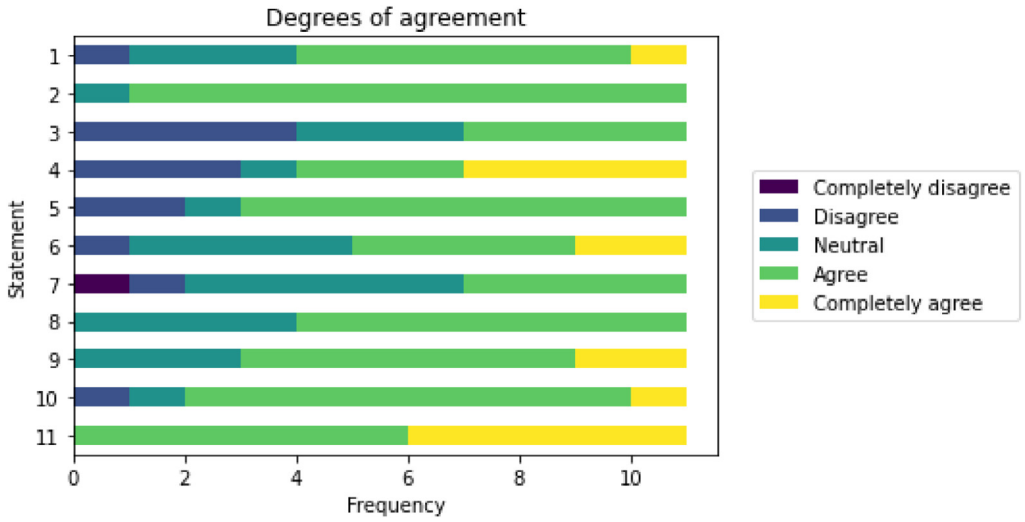


**Fig. 6.** Figure showing the frequency of reviewers indicating certain degrees of agreement to statements on a five-point Likert scale.

for explanations. However, there is room for improvement of this metric by addressing the points identified during the interviews. This room for improvement is also evident from Fig. 6, which shows us that the responses cover a wide range of agreement. The increase in satisfaction might be explained by the fact that the existing solution was lacking in depth and detail of the explanation it offered. This possibility was also reinforced by the feedback gathered from the interviews, in which reviewers indicated explanations for certain features and the overall goal of the explanation should be improved.

*Performance.* Finally, we measured the perceived change in performance using the current explanation method versus MLX1. Three statements queried reviewers as to their experiences with MLX1 and finally the addition of the Anchor component, which allowed for combinations of fraud indicators that together signalled a fraud risk. According to the survey and interview results, we can conclude that MLX1 has a positive impact on perceived user performance, compared to the current method for generating explanations.

*Statistical significance.* In order to test whether MLX1 overall improved user experience, performed a parametric test to investigate whether the results are statistically significant. According to Norman (Norman, 2010), parametric tests are a valid option in our situation, in which we have Likert-type statements and a small sample size. We performed a simple t-test, as suggested by Boone and Boone (2012). First, we combined all 11 Likert-type statements in our questionnaire into a Likert scale. The purpose of the questionnaire was to measure whether MLX1 positively improves the user experience compared to the current situation. We set $H_0$ to be no improvement over the current situation due to MLX1, and $H_1$ being an improvement over the current situation due to MLX1. Interpreting our Likert-scale as an interval scale, we obtained $\mu = 3.67$ and $\sigma = 0.86$, leading to a $t$-value of 2.5839 and a resulting $p$-value of 0.013617. Using a standard significance level $\alpha = 0.05$, we therefore concluded that there is a statistically significant improvement in user experience when using the explanations generated by MLX1.

*Evaluation of H1.*

By evaluating the three separate parts of the hypothesis covering user trust, explanation satisfaction and perceived performance, we can now assess our first hypothesis. The response to MLX1 tested on the mortgage fraud model was overwhelmingly positive. Several statements had a high spread of answers, as seen in Fig. 6. Interviews also indicated there is room for improvement, specifically involving explanation satisfaction. This could lower the spread of answers, leading to even stronger overall agreement.

Nonetheless, we can state that the MLX1 system, which uses techniques that enable individual predictions to be interpretable and transparent, achieved a statistically significant improvement in trust, satisfaction and usability of ML with their daily users. Even the reviewers who indicated slightly differing levels of agreement stated that they experienced an improvement by using MLX1 compared to the current approach.

### 4.2. Demonstration results

Two demo sessions were held for MLX2. In one session, we demonstrated the use of the Aequitas component of MLX2 in combination with CBS data to vet an ML system for potential demographic bias. The CBS data contains aggregated data on a 100 meter by 100 meter geographical basis, providing (amongst other data) information on the distribution of migration history of inhabitants within the area. All participants reported that the Aequitas component of MLX2 would be highly useful, and stressed that if further developed so that it could more easily be used by different departments, it could be useful throughout the organisation. This further development would cover the ease-of-use of the library, using a web interface, and better understandable presentation of the results of the analysis. It was decided that Legal would take the lead on the further development of the tool, starting by investigating the possibilities of the output reports.

The other session consisted of a 30 min demonstration of the *What if?* tool, BlackBoxAuditing and Aequitas, which are incorporated in the MLX2 system. We then opened the floor to feedback and questions concerning the presentation, as well as a group discussion. From this discussion, the following four concrete points of feedback were distilled:

- **Example applications**
  Even though the use of all three libraries was demonstrated using the mortgage fraud model and example cases, the data scientists requested additional concrete examples from different ML models. This reflects the need for stakeholders to be presented with ‘evidence’ from a wider range of ML application domains.
- **Concrete thresholds**
  MLX2 gives the data scientist certain indications of indirect influence and fairness measures, but feedback suggested the need for

**Table 2**
Average importance ranking given to the four points gathered during the demonstration to data scientists, as well as standard deviation.

| Feedback point | Average importance (1 to 5) | Standard deviation |
|---|---|---|
| Example applications | 4.25 | 0.50 |
| Concrete thresholds | 4 | 0.81 |
| Clear work instructions | 4.5 | 1.00 |
| Choosing fairness measures | 3.75 | 0.50 |

concrete thresholds for these indications, to enable easier reporting in the PID. Having set thresholds for these indications, which are relative for each model and therefore comparable, would eliminate unclear and fuzzy questions in the PID, optimising the process for both data scientists and personnel from Legal, Risk and Compliance departments.

- **Clear working instructions**
  Besides the demonstrations, the data scientists indicated that clear working instructions, accompanied by examples, would be essential in applying the ML2 system to their own models. Data scientists can be considered the stakeholders most likely to be able to comprehend the global explanation techniques embedded in MLX2. Therefore, their response emphasises an overall need for thorough training programmes to enable effective deployment of xAI within organisations.
- **Choosing fairness measures**
  Aequitas offers seven different fairness measures, and the data scientists requested additional explanations and background information regarding these, so that they could choose the fairness measure (or measures) that best fits their application.

These four main points of feedback were compiled into a short survey, which was filled in by the data scientists who also attended the demonstration session.

Table 2 lists the average importance on a Likert scale from 1 (not important) to 5 (very important).

The data scientists indicated that, when the points discussed above could be addressed, MLX2 could be a great asset to support them in their mode of working and in meeting their responsibilities. By using MLX2, they felt that big steps in quantifying global interpretability and bias could be made, besides granting them additional insight into the models for which they are responsible. The data scientists discussed the possibilities of requiring output of the MLX systems to be included in every PID, as they acknowledged the added benefit of having quantified results.

*H2: Techniques that allow for ML tools to be globally interpretable and demonstrably free of discriminatory bias enable organisations to streamline internal processes concerning fair and balanced AI.*

After gathering results from the two demonstration sessions, we can begin to evaluate the second hypothesis. The demonstration session given to the group of data scientist gathered feedback in the form of a group discussion session, which resulted in four main feedback points. These four points were processed into a short survey, which was distributed among the participants of the demonstration session in order to gauge their importance. These results are shown in Table 2.

Looking at the means of the importance scale displayed in the table, we can infer that providing clear working instructions is the most crucial point, followed by sharing example applications, and helping data scientists set concrete thresholds and choosing fairness measures. The table also shows that the highest-ranked point regarding work instructions also has the highest standard deviation. The ranking and standard deviations together show that much of the uncertainties of using the libraries could be addressed by developing clear working instructions for the data scientists to use, as well as supplying example applications to inspire the data scientists to use the techniques. Based on this feedback, we developed and evaluated work instructions, which were validated and accepted by the organisation. The third most important point, addressing the determination of concrete thresholds for fairness measures and indirect influence, also has a high standard deviation. In the group discussion, data scientists indicated that concrete thresholds and measures could greatly help them in filling in project documents and streamline the paperwork process that is usually involved with initiating or adapting ML models. The least important point addresses choosing fairness measures, which also has the lowest standard deviation of 0.5, implying data scientists largely agree on this ranking.

*Evaluation of H2.*

From the feedback gathered from the demonstration session to data scientists, as well as the data gathered in the session where Aequitas was presented to Legal, Risk and Compliance officers, we can conclude that the inclusion of the *What if?* tool, BlackBoxAuditing and Aequitas libraries into MLX2, and subsequently the integration of MLX2 in the data scientist's workflows would improve their efficiency and facilitate easier documentation. It would also facilitate the processes involved in bringing a new or adapted model to production. The points covering work instructions and the development of proper thresholds can be tackled, as also indicated by the participants in the management demonstration session. These participants showed great interest in MLX2 and decided to tackle the step of setting proper thresholds for different sources of possible bias. They concluded that deployment of a working Aequitas tool could help improve and streamline the processes surrounding the deployment of ML tools. This provides clearly evidence that techniques that allow for ML tools to be globally interpretable and demonstrably unbiased enable organisations to streamline internal processes concerning fair and balanced AI, specifically through the additions to the PID as a result of the Aequitas thresholds and fairness measures.

### 4.3. Adherence to the ethical framework

As mentioned in Section 2.3, there are a number of ethical guidelines to which the organisation must adhere. These guidelines stem from requirements set forward by the High Level Expert Group on AI advising the European Commission, and aim to establish meaningful human control over AI models, also addressing ethical concerns around deploying AI in the mortgage sector, discussed in Section 2.3.1. The results from the survey held with mortgage application reviewers, as well as the demonstration sessions held with data scientists and personnel from Legal, Compliance and Risk, clearly demonstrated that both MLX1 as well as MLX2 help the organisation adhere to the guidelines set forward in the Ethical Framework for insurers. MLX2, specifically its Aequitas component, is particularly useful, as it is relevant for a large number of the guidelines. The demonstration session held for Risk, Legal and Compliance managers resulted in much discussion, and participants indicated they better understood the possibility of bias in a system and how to test for it, allowing them to better adhere to the relevant guidelines. Below, we list all relevant guidelines and whether MLX1 and MLX2 helped the organisation's adherence.

**(7) The insurer ensures adequate quality (including integrity, correctness, representativeness) of (training) data used for data-driven applications.**
MLX2, specifically the Aequitas component, allows data scientists and management to evaluate whether the distribution of different groups in the data, for example male/female, is adequate. This helps fulfil the representativeness part of this guideline. Other techniques or processes must be used to ensure integrity and correctness.

**(14) The insurer ensures that employees working with data-driven applications have received adequate training, specifically to avoid confirmation bias and to ensure human autonomy.**

MLX1 does not aid in the training of employees, but it does assist in ensuring human autonomy and reducing confirmation bias. Because reviewers are confronted with two explanations, they are less susceptible to confirmation bias. The reviewer will have their own opinion on a case, using their domain knowledge and potentially suffering from confirmation bias. MLX1 offers an explanation which is potentially different from the reviewer's, which may cause them to reconsider. In the end, the assessment of the case remains the reviewer's responsibility, ensuring human autonomy.

**(15) The use of data-driven applications in production will always be subject to adequate human oversight.**

The organisation uses a long and extensive process involving the Project Initiation Document before deploying a predictive model in production. The libraries used in MLX2, specifically the *What If?* tool and Aequitas, help this process, as demonstrated in Section 4.2. Furthermore, the PID process can be improved by forcing the development of thresholds for the Aequitas fairness measures, as well as reasoning for the chosen fairness measure and the identification of possible indirectly biased features.

**(18) When employing data-driven applications, human intervention will always be possible and explanations can be obtained by customers regarding the results of an application.**

Human intervention is possible as a result of the process, in which the organisation refrains from making automated decisions. MLX1 and partly MLX2 offer the ability to provide a client with more information on the results of an application pertaining to their case. On request, MLX1 can be used to generate the same explanation as offered to the reviewers.

**(19) When the infringement on fundamental rights, including the unfair discriminatory bias in data-driven applications cannot be avoided, the insurer will not employ the application.**

All three libraries included in MLX2 are used to complete the Project Initiation Document, and evaluation indicated that certain thresholds involving the output of the libraries should be developed. Therefore, MLX2 helps the organisation adhere to this guideline.

**(20) In deciding to use data-driven applications, the insurer considers diversity and inclusivity, especially regarding groups who are at risk of exclusion or disadvantage as a result of special needs.**

Aequitas, included in MLX2, helps both the data scientist as well as Legal, Compliance and Risk officers evaluate false positive rate disparity for different demographic groups, provided that the information is available and is allowed to be used, considering GDPR regulations. Evaluation indicated that Aequitas will be used in this manner wherever possible.

**(21) The insurer will monitor the impact of employing data-driven decision making on groups of clients.**

Similar to Guideline 20, Aequitas (included in MLX2) will be employed to evaluate the fairness of data-driven applications, by comparing false positive rate (FPR) disparity for demographic groups based on migration background, which is achieved through enriching the data with aggregated CBS data based on zip code. In the future, the tool can be used to evaluate different fairness measures besides FPR disparity.

**(23) The insurer will set up an internal control and accountability system for the use of AI applications and data sources.**

This accountability system was set up prior to the development of both MLX systems. However, the tools included in MLX2 will be included in the main artefact of this accountability system, the Project Initiation Document. The outputs of the libraries in MLX2 must meet certain thresholds, to enforce internal control.

**(24) The insurer improves the knowledge of executives and internal auditors with regards to data-driven applications.**

The use of Aequitas (included in MLX2) is suitable for interpretation by executives and auditors, as proven in the demo session to senior Legal, Risk and Compliance officers. This is partly the result of simple visualisations.

**(25) The insurer ensures adequate internal communication on the use of data-driven applications.**

Following the use of the Project Initiation Document and the acceptance process of this document, the organisation adheres to this guideline. The Aequitas component of MLX2 will be used in this document, furthering the knowledge of possible bias with the involved stakeholders.

**(26) The insurer performs a risk and effect assessment with regards to the immediate stakeholders for each data-driven application.**

The Project Initiation Document was set up to adhere to this guideline, among others. With the use of Aequitas (included in MLX2), the document will be extended with insights on the impact of potentially vulnerable demographic groups, based on migration background and the thresholds for disparity between the groups, aiding in the risk assessment necessary for every application.

The demonstration session held for Risk, Legal and Compliance managers resulted in much discussion, and participants indicated they better understood the possibility of bias in a system and how to test for it. Therefore, we can conclude that not only do the techniques employed by MLX1 and MLX2 improve user experience and allow for the streamlining of internal processes, they also aided the organisation in tailoring their processes to further adhere to the Ethical Guidelines.

## 5. Conclusions and future work

In this work, we have explored the effectiveness of post-hoc, model agnostic techniques for ensuring the interpretability of machine learning models. We considered local as well as global explanation techniques and reported two main findings:

1. Techniques that allow for individual predictions to be interpretable and transparent can improve trust, satisfaction and usability of ML tools with their daily users.
2. Techniques that allow for ML tools to be globally interpretable and demonstrably free of discriminatory bias enable organisations to streamline their internal processes concerning fair and balanced AI.

The results of our research were validated in cooperation with a large international financial organisation based in the Netherlands. A case study using a model for estimating fraud risk for mortgage applications was performed. This model produces daily reports of potentially fraudulent cases, which are enriched with an explanation as to why there might be a risk. High-risk cases are then reviewed, with a focus on the factors mentioned in the explanation. Prior to our work, the organisation used a purely rule-based system to generate explanation, addressing only a subset of possible features. In our work, we replaced this with an automated method using a combination of SHAP and Anchor to provide detailed explanations containing all possible features and high-risk combinations of features in the model.

The explainable AI system implementing these local explainability techniques, dubbed MLX1, was validated using an in-depth survey and short interviews with eleven mortgage fraud experts. The survey captured their assessment of changes in trust, explanation satisfaction, usability and perceived performance. These three measures allowed for a qualitative measurement of whether MLX1 improved the reviewers way of working with the existing AI model for mortgage fraud

detection. It was found that MLX1 achieved a statistically significant improvement in all three measures.

A second explainable AI system (MLX2) focused on techniques for global explanation of the mortgage risk model, based on the *What if?* tool, BlackBoxAuditing and Aequitas. It aimed to provide insights into the overall decision making of the model, as well as offer ways of quantitatively measuring (demographical) bias. In addition, work instructions and examples were provided to the data scientists working with the libraries. In this system, the Aequitas techniques were extended with aggregated data concerning migration background based on zip-code, allowing for the possibility to test models on bias with regards to this background. This enabled the organisation to quantify the bias present in its ML systems, and allowed data scientists to properly verify their models.

The potential of this collection of global explanation techniques was demonstrated to data scientists and representatives from risk, legal and compliance departments, and it was found that this enabled the organisation to streamline processes concerned with the construction or adaptation of models.

Furthermore, we conclude that the explainable AI systems we have developed in the context of this work have aided the organisation in adhering to guidelines concerning the ethical use of machine learning in the insurance domain, particularly with respect to transparency, accountability and technical quality as expressed in the Ethical Guidelines for the Insurance Industry in the Netherlands. This enables the organisation to establish human control over its AI systems, addressing ethical concerns around deploying AI in the mortgage sector as raised in Section 2.3.1.

Overall, the findings from our research described in this article highlight the benefits of using interpretable AI techniques in practice. The opportunity to implement and validate our systems in a business environment, with a wide range of different stakeholders, some familiar with machine learning and others not, clearly demonstrates that these libraries can indeed render ML systems easier to interpret and critically assess, both on a local and global level.

Finally, a number of practical observations were made with respect to deploying explainable AI in a real-world setting. We found that deploying an explainable AI system in a business environment requires trading off performance and extensiveness, as demonstrated by the need to truncate the data set to the top 15 globally most important features for use in the Anchor model in order to achieve acceptable run time. This gain in execution time is essential, especially in this case where, in practice, reviewers rarely use any feature outside the top 15.

Furthermore, human experts do not always want all the possible information offered by an explainable ML technique; in our use case, detailed explanations were feared to bias the reviewers of mortgage applications and to negatively impact proper human oversight. As a result, the organisation rejected a prototype system that provided, as part of the explanations it generated, comparisons to averages in the data for certain features, in order to provide human experts with additional context.

Our evaluation of MLX1 showed that users experience a higher level of trust using AI systems implementing MLX1 than those without. However, we recommend implementing business processes that guard against users blindly accepting the explained results, disregarding their own research. This endangers the human oversight required by the Ethical Guidelines. In the mortgage application context, reviewers could be required to verify and comment on every single risk indicator offered by MLX1, to ensure that they have viewed and considered the model decision and insights, but not blindly copy it.

There are several limitations to our research that should be addressed in future work. The fact that it was unknown whether suspected fraud cases ended up confirmed as such, as well as limited access to reviewers, meant that the true gain in performance, as well as the potential of a more subjective implementation variant that adds more statistically based explanations, remained untested.

Other limitations could arise when applying the methodology introduced in our work broadly in practice, for example the added computation time of KernelSHAP in the case of a model architecture other than a tree ensemble, or privacy and data protection laws making it impossible to gather aggregated data with regards to migration background.

Further areas for future work include comparing MLX1 for local explanations against a baseline providing a limited textual explanation. Evaluating the Anchor component of MLX1 against manual grouping of related features could also give insight into the difference between the two, as the reviewers presented with the manual grouping had widely differing reactions. Furthermore, developing a method to quantify the increase in trust and gains would be interesting, and allow for a comparison between an objective variant as built in this research, and a more subjective variant of MLX1 using context.

The research presented in this paper has allowed the organisation to improve the acceptance and effectiveness of their ML models, by implementing a new explanation method that allows daily users to better understand model decisions. Furthermore, the paper presented an approach to test models for unwanted bias and discrimination, in a bid to streamline internal processes regarding model acceptance and deployment. Both methods presented are designed to be applicable to classification systems beyond the fraud risk domain, hopefully improving model explainability and bias testing with all ML models used in high impact applications.

## CRediT authorship contribution statement

**W. van Zetten:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **G.J. Ramackers:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration. **H.H. Hoos:** Methodology, Writing – review & editing, Supervision.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Wessel van Zetten reports financial support was provided by NN Group NV. Wessel van Zetten reports a relationship with NN Group NV that includes: employment.

## References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. http://dx.doi.org/10.1109/ACCESS.2018.2870052.

Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., & Venkatasubramanian, S. (2018). Auditing black-box models for indirect influence. *Knowledge and Information Systems*, *54*(1), 95–122.

ANP (2020). Overheid stopt met gebruik SyRI na uitspraak rechter. *Het Parool*, URL https://www.parool.nl/nederland/overheid-stopt-met-gebruik-syri-na-uitspraak-rechter~bbf3993a/, Accessed: 2021-02-04.

Boone, H. N., & Boone, D. A. (2012). Analyzing likert data. *Journal of Extension*, *50*(2), 1–5.

Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. *CoRR*, arXiv:2010.04053.

Doran, D., Schulz, S., & Besold, T. R. (2018). What does explainable AI really mean? A new conceptualization of perspectives. In *CEUR workshop proceedings*. CEUR, Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017..

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local rule-based explanations of black box decision systems. *CoRR*, arXiv:1805.10820.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42. http://dx.doi.org/10.1145/3236009.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: challenges and prospects. *CoRR*, arXiv:1812.04608.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education*, *38*(12), 1217–1218. http://dx.doi.org/10.1111/j.1365-2929.2004.02012.x.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777). Red Hook, NY, USA: Curran Associates Inc..

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), http://dx.doi.org/10.1145/3457607.

Molnar, C. (2020). SHAP (shapley additive explanations). In *Interpretable machine learning: A guide for making black box models explainable*. Leanpub.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632. http://dx.doi.org/10.1007/s10459-010-9222-y.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/2939672.2939778.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 32*.

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, *8*, 42200–42216. http://dx.doi.org/10.1109/ACCESS.2020.2976199.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. Published as a background and development guide, available at https://arxiv.org/abs/1811.05577.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, Article 102551.

Sullivan, G. M., & Artino, A. R. (2013). Analyzing and interpreting data from likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541–542. http://dx.doi.org/10.4300/jgme-5-4-18.

Verbond van Verzekeraars (2021). Ethisch kader. URL https://www.verzekeraars.nl/media/7541/ethisch-kader.pdf.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., & Wilson, J. (2020). The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 56–65. http://dx.doi.org/10.1109/TVCG.2019.2934619.