Universiteit Leiden

Master ICT in Business and the Public Sector

Interpretable Machine Learning and Bias Detection for Machine Learning Applications in the Financial Industry: mortgage fraud detection at a Dutch insurer

Name:	Wessel van Zetten
Student ID:	s1836765
Date:	December 21, 2021
1st supervisor:	Dr. G.J. Ramackers
2nd supervisor:	Prof. dr. H.H. Hoos

Master's Thesis

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Machine learning provides a promising set of new technologies for increasing efficiency in many organisations. However, their 'black-box' nature often creates challenges to their acceptance due to the lack of insight into their internal decision process, and potential issues with regards to fairness. This thesis identifies and implements a number of post-hoc, model-agnostic techniques that make individual model predictions interpretable, granting insight into the decision making process of the models. In addition, it develops a method to test models for direct and indirect bias. We investigate the effects of these techniques by performing a case study in cooperation with a large Dutch financial organisation. Two prototypes were developed and applied to their mortgage application fraud risk model.

The first prototype focuses on local interpretability techniques, and was validated using a survey of a group of expert mortgage reviewers. The effects of the prototype for making models interpretable were measured by testing the trust, satisfaction and perceived performance of employees working with the model decisions on a daily basis. Furthermore, the effects of such techniques on internal processes covering model acceptance were investigated.

The second prototype addresses global interpretability and bias detection, including indirect racial bias detection by extending an existing technique with zip code aggregated data on migration background. Its effectiveness in both the development of machine learning applications as well as the potential to streamline internal processes was validated by a group of data scientists and senior managers in legal, compliance and risk departments.

Our results demonstrate that post-hoc, model-agnostic techniques aimed at making black-box models locally interpretable statistically significantly improve the trust, satisfaction and usability of those tools with its daily users. Furthermore, techniques that cover global interpretability and bias detection towards demographic groups were found to streamline internal model management processes concerning the application of fair and balanced AI.

Acknowledgements

I would like to thank my primary supervisor, Dr. Guus Ramackers, for his invaluable guidance throughout the project. Your help in conducting the research while balancing organisation and academic interests in this case study was crucial.

Secondly, I want to extend my thanks to my second supervisor, Prof. dr. Holger Hoos, for offering much more help than required for a second supervisor. Your extra feedback during and after the research has brought this work to a higher level.

Furthermore, I want to acknowledge my two supervisors at the organisation, as well as my other colleagues. You provided me with all the necessary tools to conduct this research.

Finally, I would like to thank my parents and my partner, who helped me find distractions and supported me while writing this thesis.

Contents

1 Introduction

Artificial Intelligence systems are being deployed in a multitude of different application domains, increasing the scope and scale with which AI affects our daily lives. These systems have secured a vital position in a number of different industries, such as healthcare, education, entertainment and finance. AI systems, or more specifically Machine Learning (ML), can achieve high precision in difficult prediction tasks where humans are unable to see the patterns required to understand the 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 certain decision was made. Several problems arise as a result of this inability to examine the decision making process. Firstly, it is difficult to validate that the decision making process is defensible in that it follows a logical train of thought. Secondly, it is difficult to determine whether the decisions thus made are unfairly biased towards certain demographic groups. Methods that could help gain insight into this obscured decision making process and verify that systems make fair decisions are an extensive research area. Learning from the decision making process could enhance knowledge of the target domain, and ensuring systems are bias free and follow sensible decision processes is a requirement for the ethical application of machine learning.

One frequently used domain for application of Machine Learning is in estimating the risk of fraud. The Dutch government developed the *Systeem Risico Indicatie* (SyRI), a system which linked personal information from a large number of governmental databases in an effort to identify potential fraudulent individuals. Individuals who were judged to have a higher risk were placed on a list, after which tax authorities could make further investigations into these individuals. As the SyRI system was a black-box ML system, the precise requirements to mark an individual as a potential fraud risk were unclear and the decision making process was inexplicable. Furthermore, it was impossible to prove whether SyRI suffered from (unintended) bias. These reasons, combined with the fact that being marked as a potential fraudster had great impact on the individual, causing the system to be prohibited in 2020, after being used for six years [ANP20].

The potential for ML applications to estimate fraud risks extends beyond governmental organisations. For example, 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, in order to avoid unfair treatment. This helps the organisations adhere to anti-discrimination laws. Furthermore,

the 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 [Ver21].

1.2 Hypotheses

This thesis consists of 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 multiple different areas, for example in mortgage and insurance. The organisation is looking for ways to further improve fair AI, in order to adhere to the obligations set forward in the Ethical Framework. Fairness in the context of AI decision making is defined as follows: *"The absence of any prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics"* [Meh+21]. In order to achieve this, the organisation would like to have a set of tools that help data scientists to better understand and explain their models, to validate the decision making process, and investigate and demonstrate the fairness of a model. Furthermore, it requires methods to explain the individual outcomes of a classification model, so that the employees working with said models on a daily basis are able follow the particular decision making process. Allowing employees to gain insight in to the decision making process in the case they are handling enables them to draw knowledge from the model. This knowledge may help them in their own research into the case, or give them a starting direction for their 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 models, and whether to agree or disagree with it.

Techniques developed to gain insights in to the workings of black-box models fall into the areas of Explainable Artificial Intelligence, defined as follows: *"Explainable AI (XAI) is the class of systems that provide visibility into how an AI system makes decisions and predictions and executes its actions. XAI explains the rationale for the decision-making process, surfaces the strengths and weaknesses of the process, and provides a sense of how the system will behave in the future*" [Rai20].

1.3 Methodology

This thesis will apply research by design, and develop several different prototypes which will be validated at the organisation. By researching existing techniques in this expanding domain, we will be able to select a set of techniques and combine them into prototypes and work instructions that enable the organisation to achieve the goals described above. The prototypes that use

these techniques will make it possible to examine the overall global decision making process of a model (so-called global interpretability), as well as be able to explain individual predictions in detail (local interpretability). The prototypes will also be able test models for bias on grounds of gender or race. The first prototype, the local prototype, will focus on generating detailed explanations for individual predictions. The second prototype, the global prototype, focuses on gaining insights into the overall decision making process of the given ML system, as well as ensuring that it is bias free. More concretely, this thesis investigates which techniques for explainable AI can be used in an organisational context, and what are the implementation considerations. Furthermore, we investigate how the possible techniques fit in to existing AI processes at the organisation, and how effective these techniques are. In order to verify that the method of developing prototypes and work instructions helps organisations achieve interpretable and unbiased, and to answer the research question, 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.

After development of the two prototypes, they were validated with their respective target groups. The local prototype was validated with a group of mortgage application reviewers, who work with a model assessing mortgage applications on a daily basis. The global prototype was validated once with data scientists, specifically focusing on the use of the tools in their development process, and once with several people from Legal, Compliance and Risk departments, in order to investigate how the prototype and insights gained can help streamline the acceptance processes involved in every new and existing ML model in use within the organisation.

The result of this thesis will be an evaluation of applicable techniques for the given research problem, and implementation considerations in a practical scenario. Furthermore, the thesis covers the effectiveness of the techniques, and serve as a implementation example for similar techniques in domains besides the financial domain.

1.4 Structure

In the following, we first explore related work in Section 2, in order to identify interesting and relevant prior research and techniques that could be used in building the prototypes. Section 3 describes the business requirements mandated by the organisation, stemming from the business context. Section 4 covers

system design, in which we discuss the logical design and related technical design of both prototypes. The validation of the developed prototypes is discussed in Section 5 by first explaining the survey design and execution which were used to validate the first prototype, followed by the demonstration session used in the validation of the second prototype. Section 6 covers the results of the validation of both prototypes, followed by the discussion of those results. Finally, in Section 7, we present some concluding remarks and suggestions for future research.

2 Related work

In recent years, research in the field of interpretable machine learning and bias detection has grown exponentially. This research is multi-faceted: some research focuses on the effect of interpretable AI in its application fields and the interactions with its users, studying the impact on user trust and model adaptation [DSB18; Hof+18]. Another sub-field of research focuses on developing techniques to gain insights in machine learning models [AB18; Meh+21]. This research is further subdivided into developing machine learning models which are interpretable by design, or creating methods to develop a separate layer that enables interpretable black-box machine learning models.

This literature review focuses on the last area, identifying model-agnostic techniques to interpret black-box models. The review also focuses on investigating techniques to discover bias in data as well as model-agnostic techniques to discover bias in a black-box models. The reason for investigating model-agnostic techniques, as opposed to model-specific techniques, is that this thesis aims to develop prototypes that are applicable to all kinds of models, and not be limited to a certain kind of model architecture.

2.1 Explorative literature

Firstly, review papers covering many aspects of explainable and interpretable machine learning were studied, followed by survey papers that cover state-of-the-art-techniques used in different sub-fields of machine learning. After completing this process for interpretable machine learning, it was repeated to investigate different types of bias occurring in machine learning applications, followed by gathering different techniques that might be applied.

2.1.1 Interpretable machine learning

Interpretable machine learning is a subfield of explainable AI, with many different surveys aiming to give just a quick overview of inherently interpretable models vs. black-box models (e.g. [Rai20]), while others dive more deeply into the different views and perspectives associated with explainable AI (e.g. [DSB18]). Yet other papers investigate the impact explainable models have on daily users and people impacted by its decisions (e.g. [Shi21]). These review papers offer several taxonomies of applications, but lack explanations of existing relevant techniques. Roscher et al. [Ros+20] 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 [AB18] as well as Guidotti et al. [Gui+19] combined with the reviews listed above, we established that in order to fully satisfy the interpretable machine learning aspect of the prototypes,

these have to be able to both provide local interpretability, as well as 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 classification. Insight into the reason(s) why the model has scored a specific case as high-risk might help the reviewer in its validation process. 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. It is necessary to provide insight into the overall decision making process of the model, to ensure that the process is correct and reliable. This can also help prove the models reliability to stakeholders and regulators.

2.1.2 Bias detection

Mehrabi et al. [Meh+21] provides a review of different types of bias and fairness definitions. It is established that there are two aspects to 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 select 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. This bias does not stem from the training set, but can creep into the model regardless. Caton and Haas [CH20] provide an overview of the different approaches to detect and mitigate bias as well as increase fairness and was used as a starting point to explore possible techniques to detect bias in both data and model.

2.2 Techniques

Based on the review and survey papers described in previous sections, this section provides an overview of the possible techniques identified. These techniques cover model-agnostic, local and global techniques for interpreting black-box machine learning models, or are meant to detect bias in data or a trained black-box model. This subsection covers the benefits and drawbacks of identified techniques. Based on the techniques selected in this subsection, two prototypes will be built.

2.2.1 Interpretable machine learning

Starting of with techniques for explaining local predictions, Local Interpretable Model-agnostic Explanations (LIME) [RSG16], was one of the earlier local techniques. LIME learns an interpretable model around each specific prediction. The authors also developed a method using local explanations to present a representative interpretation of the full model. LIME is a linear approach, showing the importance score of each individual feature and whether the feature has a positive or negative impact on the final prediction. Non-linear approaches can

take into account the combined effects of features on the individual prediction. 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.

Rule-based approaches instead of weight-based approaches such as LIME can combine features into rules that together explain the individual predictions. An example of such a technique is Anchor [RSG18], developed by the same group of authors as LIME. The paper describes an 'Anchor' as an explanation that anchors a local prediction, so that changes to the other feature values do not influence the outcome. It defines a rule as a set of feature combinations, such that the rule returns 1 if all feature combinations are true for a specific case. The rule is an 'Anchor' if it returns 1 for an instance, and if a similar instance (for which the rule would also apply) is likely to be classified the same as the original instance. The technique attempts to choose the Anchor with the highest coverage, meaning the one that includes the largest part of the input space. Because the technique uses rules, it is able to combine features and is more faithful compared to LIME.

Another technique which uses rules is Local Rule-based Explanations (LORE) [Gui+18]. Like Anchor, it uses rules combining features to explain a local prediction. As an improvement over Anchor, it also introduces counterfactual rules, which are rules that would result in the opposite classification. LORE achieves a higher accuracy and coverage than both LIME and Anchor [Gui+18]. A small improvement over LORE is introduced as LoRMIkA [RBB20], adding hypothetical supporting and contradicting rules. Hypothetical supporting rules are rules that, if true, would further cement the current classification of the case. Hypothetical contradicting rules do the opposite, and if true would change the prediction to the opposite classification. However, the authors of the LoRMIkA paper do not provide a completed code package, making the technique unsuitable for implementation in a prototype.

One of the better known techniques for explaining machine learning models is SHAP [LL17], which uses a game theory approach and Shapley values. SHAP attempts to explain the prediction of a specific instance by computing how much each feature contributed to that prediction. For this, it uses a surrogate model, which is a model that mimics the underlying, complex model that must be explained. This surrogate model is a linear model, and the method is an additive feature attribution method, like LIME. The Shapley values for all features are computed by simulating that some features are present, while others are not. The method above gives us Shapley values for individual predictions, helping us with local interpretability. Computing Shapley values for all predictions results in a matrix of values, which can be combined to explain the entire model, together forming global interpretability [LL17; Mol20].

Our literature research also identified a large number of promising techniques that had not yet been developed far enough to be considered for a prototype.

An example of such a technique is the Confident Itemsets Explanation (CIE) [MS21], which has no need to perturb the training set to generate explanations. However, the code was not fully developed and messy, and the paper has not (yet) produced major impact. Another example of an interesting technique is Interpretability via Model Extraction [BKB17], which proposes an approach to extract a decision tree from a black-box model, which could be used to then extract rules. However, it provides no code. Other techniques were promising but were not written for Python, which is crucial in order to fit into the organisations infrastructure and codebase, or included no code [Den19; PB19].

The search for promising work identified a number of techniques aimed at providing global interpretability. As mentioned before, the most well-known technique for global explainability is SHAP [LL17], which can be used to estimate global feature importance, not taking into account the combined effect that features may have. As with local interpretability, it is possible that some features have an impact on the global model which only becomes apparent when combined with a different feature. Therefore, we explored techniques which employ a non-linear approach. One of these methods is Measure of Feature Importance (MFI) [Vid+16], which aims to score individual features on their global importance. However, this technique is only theoretical and no code is provided. Model Understanding through Subspace Explanations (MUSE) [Lak+19] proposes a framework and approach to build a rule-based explanations for subspaces in the data indicated by the user. This approach is interesting to characterise areas of interest, but not applicable in the case of the organisation since a global overview is required.

An interesting technique to determine both the overall importance of features, as well as whether they impact other features indirectly, proposes obscuring features in the training data set to certain degrees to see what the impact on the test set is. This technique, called BlackBoxAuditing [Adl+18], 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. The technique introduces gradient feature auditing, computing the influence of a feature by obscuring it from the data set, and measuring the difference in error rate achieved by the model. It is based on the idea of iteratively obscuring individual features, until they can no longer be predicted using the other features. By doing this feature by feature for all features in the data set, the technique is able to generate a list of features ordered by their influence on the model prediction [Adl+18].

Besides techniques for global interpretability that produce feature importance weights or rankings, or deliver rules determining the models workings, there is also the approach of using interactive visualisation to capture the model. An example of this is presented in Visualising the Feature Importance for Black Box Models [CMB18], which proposes an R package to visualise expected (conditional) feature importance for both global and local predictions of a model.

Another interactive tool is the What if? tool [Wex+20]. This tool allows the user to probe, visualise and analyze systems without too much coding. This way, users can inspect model behaviour in different scenarios, and build their own understanding of global model behaviour. Another interesting visualisation tool is RuleMatrix [MQB18], which uses rule-based induction to build model understanding and visualises this.

2.2.2 Bias detection

Besides exploring techniques for local and global interpretability, we also explore 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 will also yield a biased model. For both aspects, we explored existing surveys and identified techniques, working through citations to discover related and new techniques to cover.

Starting with bias in data, one of the techniques to uncover possible bias is to use basic data exploration, to explore class imbalance and distributions in the data. In the case of our prototype, this general approach can also be applied to establish a general interpretation of the data set.

Several toolkits designed to identify bias in machine learning applications were identified. Of the techniques designed to identify bias, only Aequitas [Sal+18] focused on data exploration as well as model output, and might be applicable as its visualisation properties make it attractive to both data scientists and auditors. Given the nature of the prototype and the responsibility it carries with regards to fair and unbiased decision making, simple visualisations are an attractive property. Aequitas mostly focuses on bias detection and visualisation based on chosen demographic groups. The tool does not directly interface with the model, but uses cases predicted by the model together with the ground truth (the known, correct classification of the case). 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. Aequitas was originally designed to be used in risk assessment tools, fitting the mortgage fraud risk model the prototype is built for.

There are several different techniques to identify bias in a trained model. 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. The technique also mentioned in the subsection on global interpretability, BlackBoxAuditing [Adl+18], 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 could be determined whether the protected features were not used in the model.

There are multiple toolkits focused on identifying bias and discrimination in the predictions of a model. The Unwarranted Associations framework (UA) [Tra+17] is a principled methodology to discover unfair or discriminatory treatment. However, the code provided is written for Python 2.7, and unmaintained for 4 years. Another framework is AuditAI [pym], focused on discrimination and bias in hiring applications. It uses simple statistical tests on the output of a model to identify possible discrimination.

This literature review identified a large number of techniques that could help in creating prototypes for local interpretability and global interpretability and bias detection. These techniques range from fully supported and widely used methods to more conceptual and less maintained ideas. The next sections explore the business context within the organisation, which will give rise to a number of rigid requirements any selected technique for the prototypes must meet. These requirements will be used in the system design, creating a logical design fitting the techniques into processes within the organisation, and a technical design covering the purely technological part of the prototypes.

3 Business requirements

This section explains how the Ethical Framework for Insurers [Ver21] influences the business context, and maps the different processes at the organisation in which the two prototypes would fit, subsequently identifying certain requirements the selected techniques have to meet. Firstly, we discuss the Ethical Framework for Insurers, which plays a large part in determining requirements and aims that ML applications must adhere to. We investigate the relevant areas of the framework, and how they can influence our selection of relevant techniques. Secondly, we discuss the mortgage fraud application detection model in use at the organisation, and how the reviewers who handle the every day requests interact with this system. This allows us to identify requirements and properties the techniques for the prototype for local interpretability must meet. Lastly, we follow the same method to determine requirements and properties that the prototype for global interpretability and bias detection must adhere to. Based on the requirements determined in this section, Section 4 covers the development of the prototypes.

3.1 Ethical Framework for Insurers

The Ethical Framework was 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.

The Ethical Framework sets forward 30 guidelines, grouped by the seven key requirements set up by the High-Level Expert Group. The guidelines that are relevant for this thesis are listed below, with their original number in parentheses so that they can be found in the original Dutch version of the Ethical Framework [Ver21].

- **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 the infringement on fundamental rights, including the 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 3.3. 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 proper roles with understandable and clear metrics on the degree of bias in systems, so that they can be understood and audited before deployment. The techniques selected to be used in the prototypes must therefore offer the possibilities to fit into and enhance the current situation and processes. Following the development and evaluation of the prototypes, this thesis investigates to what degree the prototypes have enabled the organisation to improve the adherence to the guidelines explained above.

3.2 Local interpretability: mortgage fraud

This thesis focuses on the model in use within the organisation that examines mortgage applications. This system uses a black-box model, particularly a tree ensemble. The model uses several basic properties of a mortgage application, as well as a number of specific indicators designed in cooperation with mortgage experts. These basic properties include information on the age and profession of the applicant, conditions of the loan and the collateral involved. In total, the version of the model used in this research uses 75 features per case. The model runs daily and assesses all new applications received. Cases are given a certain fraud risk score, and the top five cases are passed on to mortgage application reviewers for validation.

In the current situation, an explanation is generated for the five cases with the highest risk, before they are forwarded to the reviewers. This explanation is generated using 22 fixed rules, set up by mortgage experts to flag properties of a case that potentially signal fraud. The 22 rules are triggered on specific values of individual or combinations of figures, which means 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 one feature, say, Feature X. The set of rules used for explanation may not contain Feature X, which means Feature X will not be mentioned in the explanation for this high-risk case, even though it is predominantly responsible for the high fraud risk. This happens, because the method does not use a technique for locally interpretable ML, simply a set of predefined 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.

This current situation helps us to identify potential areas of improvement in the approach, which the prototype might alleviate. Firstly, the current explanation method only uses 22 rules, some of which also cover the same feature. The version of the model used in this research uses 75 features, meaning a large number of features are used by the model but are unknown to the reviewers, since they are never used in explanations. Secondly, the current method can only report on individual features that signify a potential high risk of fraud, whereas often several features combined signify a high risk of fraud.

3.3 Global interpretability and bias detection

The introduction 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 Compliance reviewed Record of Processing Activity (ROPA) on all data used by the model. Certain guidelines and development methods are followed to ensure that the data used is bias free and that decision making of the model is clear. For example, the lead data scientist is required to thoroughly investigate the training set they want to use, and document this investigation in the PID. Furthermore, the PID must include a section on the supposed decision making of the model. However, the organisation would like to have tools to quantify possible bias and discrimination, as well as tools to gain insight into the decision making process. Furthermore, the organisation would be interested in developing a way to systematically determine whether a given model is biased towards certain vulnerable demographic groups, possibly based on data containing information on the percentages of demographic groups in certain areas. The tools which will be developed in the form of a prototype would help with a generalised approach to including aspects of interpretability and bias in the Project Initiation Document, allowing the organisation to streamline internal processes concerning the acceptance of new models.

This context allows us to draw up several more requirements. Firstly, techniques selected for this prototype must be able to give a broad overview of the decision making process, as well as be able to zoom in on certain cases and their particular decision making process. Secondly, techniques must be able to determine overall feature importance rankings, while also being able to take into account features that might indirectly influence each other. Thirdly, techniques must be able to detect bias on data sets enriched with demographic information.

Now that we have explored the business context and identified specific requirements stemming from it, we can move on to the system design. The requirements distilled from the business context are defined below.

Requirements for the local prototype

- Must be able to work with a large number (roughly 75) of individual features.
- Must be able to identify individual features that indicate a high fraud risk.
- Must be able to identify groups of features that together form a high fraud risk.
- Must be able to work with such a performance as to not unduly impact the typical time set for the reviewer's task.
- Must be able to be understood by reviewers possessing only the most fundamental understanding of machine learning.

Requirements for the global prototype

- Must be able to give a global overview of the decision making process, as well as focus on subgroups of cases.
- Must be able to identify features that indirectly influence others.
- Must be able to detect bias in models using data sets enriched with demographic information.

In the next section, these requirements as well as several general requirements are used to select the techniques employed in the two prototypes.

4 System design

This section covers how the prototypes would ideally fit in the corresponding workflows, but also how the prototypes were developed and what techniques were chosen. Logical design covers the use of the prototype for local interpretability and its integration into the workflows of the mortgage reviewers, as well as the internal processes concerning the development and deployment of new and adapted ML models by data scientists. Technical design discusses the techniques chosen for both prototypes, as well as how the prototypes were developed. We use UML diagrams to capture the situation at the organisation, which allows for the approach to be generalised to other domains.

4.1 Logical design

As mentioned before, the prototype for local interpretability involves generating explanations for mortgage applications that have been assigned a higher fraud risk by the model. In Figure 1, we visualise the two different paths for an application after it has been classified by the model. If the 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 the reviewers, who first assess the explanation before assessing the application like they would normally, to ensure the model decision is verified independently. 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.

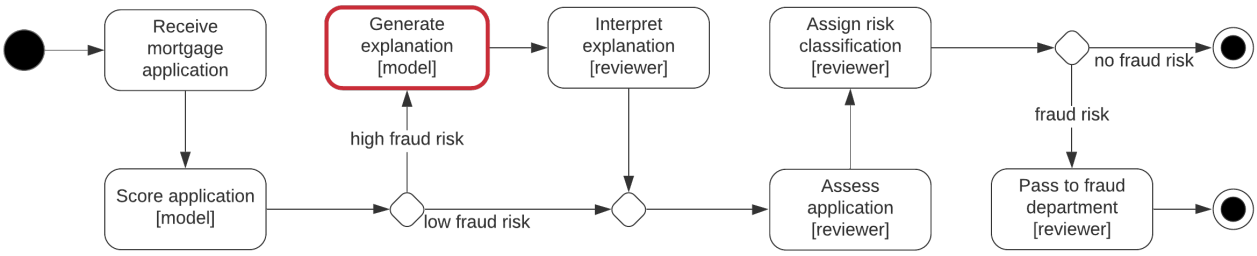

Figure 1: Activity diagram showing how the process of mortgage reviewers uses the ML tool and explanations. THe application of the prototype is shown in red.

The prototype would simply replace the current method for generating the explanation, which in the current situation is done using simple decision rules. Possibly, the interpretation of the explanation must be changed in this process, since the reviewer is currently expected to verify each part of the explanation. The current method highlights one to three aspects of the application, while the

prototype method will be able to generate a much more extensive and detailed explanation. Given that the reviewers are expected to spend roughly 5 minutes 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 tool providing global interpretability and bias detection. The main document, the Project Initiation Document (PID), details the type of data used, the goal of the project, the techniques used, and whether the model is bias free and fair. This PID must be filled in by the responsible data scientist, and must also gain approval from the business stakeholder, the data privacy officer, as well as legal, risk and compliance contacts. Figure 2 is based on internal documents detailing the progress of PIDs, and shows the parties that must give their approval before a project can start.

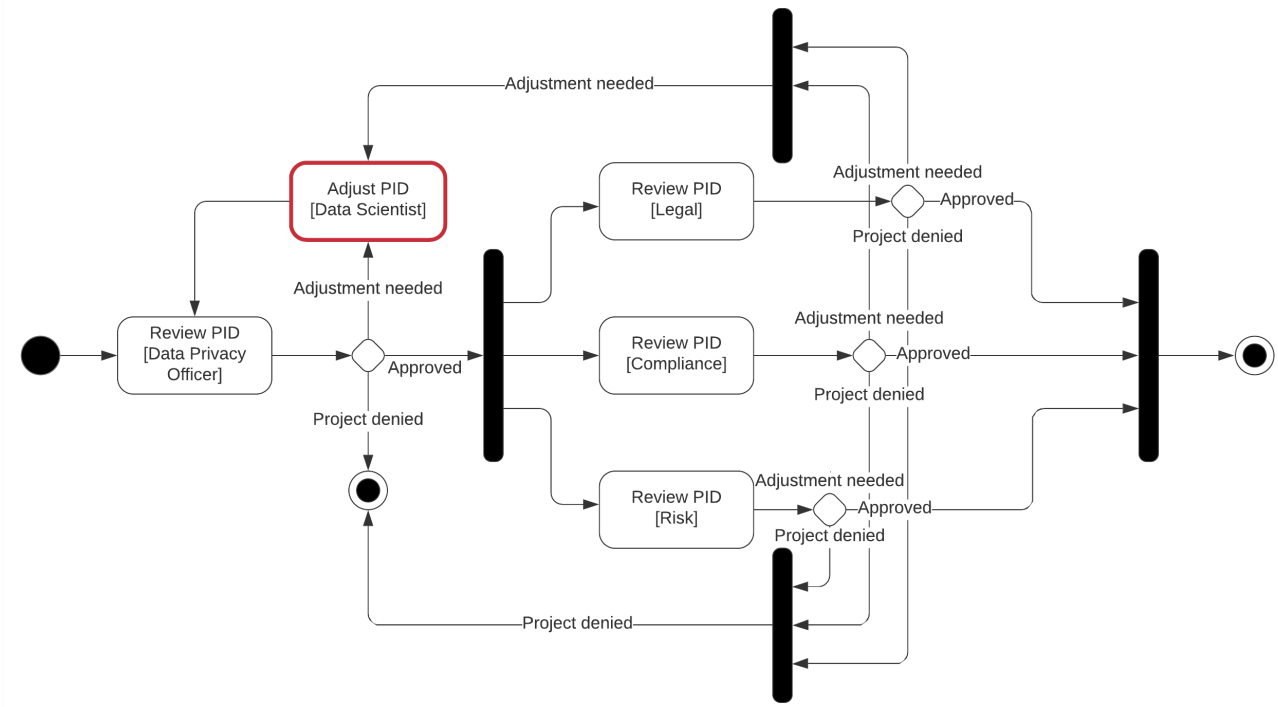

Figure 2: Activity diagram showing the process for approving Project Initiation Documents. The application of the prototype tools is shown in red.

Currently, the results of the Record of Processing Activities (ROPA) on all data used for the model are used to ensure that the developed model is bias free, but

the organisation lacks the tools to quantify the degree of bias. Any elements of the PID being unclear could lead to disapproval from one of the involved parties, meaning the document has to be updated and possibly run through the acceptance pipeline again. 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 clearer the bias and overall decision making process of the tool, the smaller the chance that any involved party will disapprove a PID, sending it back to the data scientists.

4.2 Technical design

In addition to the specific requirements identified previously (Section 3.3), pertaining to the two different prototypes, there are several requirements that must be met by all techniques. Firstly, since all existing infrastructure within the organisation involving machine learning and data is based on Python, including the models the prototypes will run on, the techniques must be written in a recent version of Python. Secondly, the selected techniques must be more than a concept presented in a paper, and have completed code (on GitHub), as well as be actively used and supported by its developers. Thirdly, the selected techniques must must allow for commercial use. Concretely, this means that the Python package is released under an open-source license which allows for use in proprietary products without the obligation to release the code, such as the Apache 2.0 or BSD license. Finally, while the prototypes will be developed for and validated on the mortgage fraud model, they must be generalisable to all Python models in use at the organisation. This means that the selected techniques must be able to work with more than one ML architecture, so-called model-agnostic techniques. These requirements are listed below.

- Must be written in or compatible with Python 3.7
- Must be released as a package and have active support
- Must be released under a license that allows for commercial use
- Must be model-agnostic.

4.2.1 Selected techniques

Finally, this section presents the techniques which were used in the development of the prototype. These techniques have been selected using the criteria stemming from the business context, as well as the general requirements listed above. These requirements were combined with input from the organisation concerning the desired explanations and insights into the decision making of the model.

For local interpretability, both Anchor [RSG18] and LORE [Gui+18] were tested, since both are non-linear and are able to capture the combined influence of multiple features on the local prediction. They deliver clear, interpretable rules,

which is important for the mortgage reviewers that are tasked with reviewing flagged cases. However, it was found that both techniques had a very long running time for individual cases (approximately 4 minutes for Anchor, 11 minutes for LORE) when evaluating on test cases, making them unsuitable for use in the organisation's workflows.

In light of this, it was decided to use SHAP [LL17] 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 a complexity of $\mathcal{O}(TL2^M)$, whereas TreeSHAP has a complexity of $\mathcal{O}(TLD^2)$ [Mol20].

In our testing, we found that the running time for KernelSHAP to be able to produce local explanations for all cases in the test set was about 90 seconds, where TreeSHAP was able to produce explanations in just a few seconds. This running time is not too important, as SHAP only has to run once per batch, or once per day in this case. However, we did want to ensure that both KernelSHAP and TreeSHAP do 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 prototype. 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 prototype, since reviewers only act on the first few features. Beside using SHAP for individual features, it was desirable to identify combinations of features that contributed to the outcome of a single case, which necessitates the use of a non-linear technique. Therefore, it was decided to employ Anchor on a truncated data set as an addition to the SHAP prototype. This process is further explained in subsection 4.2.2.

For global interpretability, it is important to choose a technique that can produce a general overview of the feature importance in the model, as well as offer the possibility to really explore the model and zoom in on individual outcomes. This way, the model can be demonstrated to conform to all the requirements set forward in the Ethical Framework [Ver21] as described in Section 3.1, and also grant insights into the decision making process of the model. Not a single identified technique could offer this, so several techniques had to be combined. The adherence to the Ethical Framework is explained in detail in Section 6.2.2. To generate a general overview of feature importance, BlackBoxAuditing [Adl+18]

was used. It obscures certain features to establish their overall importance to the performance of the model. For interactive visualisation, the What if? tool [Wex+20] was used, which allows the user to follow the decision making process of the model and manipulate hypothetical cases to see what would change in the outcome. Both BlackBoxAuditing and the What if? tool have supported Python packages which could be integrated into the prototype.

To identify bias in the data set, it was proposed to establish certain standards for data exploration, to gain insight into possible imbalance and distributions in the data. Furthermore, Aequitas [Sal+18] will be used to further explore possible disparity between protected groups in the data set. Aequitas has a Python package which is well supported and a license which allows for its reuse.

Finally, to identify bias in the model two methods used earlier were applied, BlackBoxAuditing [Adl+18] and Aequitas [Sal+18]. BlackBoxAuditing can be used to investigate whether protected attributes which cannot be used in the model are not indirectly influencing the outcome of the model through other attributes. Aequitas was used to identify possible discrepancies in selected metrics between demographic groups such as false positive rate disparity or false discovery rate disparity, to ensure that the model does not discriminate.

4.2.2 Prototype development: local interpretability

Three variations of the local prototype were developed, which we presented to data scientists and mortgage application reviewers working for the organisation. Based on their feedback, one variation was developed further.

We started by transforming the SHAP values into impact on the probabilities of our model, so that the individual impact of features add up to 1.00. This way, we can rank the features based on impact on the specific case. The first variation we proposed used this ranking of features and displayed the top 5 features (with understandable names) with the highest impact on the case outcome, as well as the precise impact each feature had. The end result of this variant, illustrated using a real case also used in testing but with one proprietary indicator redacted, is shown in Figure 3.

Risk score: 0.711

Indicator	Value	Risk impact
Job length of main applicant (in months)	4	0.197
Age of main applicant (in years)	29	0.069
Loan amount	298000	0.062
Days until date of sale	70	-0.052
Proprietary indicator	proprietary	0.051

Figure 3: Output produced by variation 1 of the local prototype. It shows the top five indicators explained, with the value of the indicator and the impact on the overall risk score. Indicators can either increase or decrease the risk score.

The second and third variation used understandable, textual explanations for each feature, in the form of written sentences describing the feature. The second variation used straightforward, objective explanations, while the third variation used words such as 'lower than' and 'higher than' thresholds set for each individual feature. These variations used three different impact groups to rank the features. The 'high impact' group contained features with an impact on the probability that was 0.10 or higher, while 'medium impact' features had an impact between 0.10 and 0.05. Finally, 'low impact' features all features below the 0.05 threshold. The variations presented the top 5 highest impact features, dividing them into the 'high impact' and 'medium impact' categories. If there were only features in the 'low impact' group, we showed the top three. The output of these two variations is shown in Figure 4.

- 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*.

Figure 4: Output produced by variation 2 (left) and variation 3 (right) of the local prototype. 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 variation three is highlighted in bold.

After presenting these three variations to the head of the reviewers, and discussing with data scientists, the organisation decided to continue with variation 2, using the objective textual explanations. The head of reviewers argued that variants 1 and 3 were too subjective, and therefore undermined the validation task that the human reviewers have, according to the Ethical Framework [Ver21]. 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 (like in the first variation) would influence them, stemming from insufficient ML knowledge to understand that a fraud classification according to the ML model does not indicate there is a definite case of fraud. Variation 3 was declined because of the thresholds used, and the fear that said thresholds might confuse the reviewers. After all, cases are supposed to be judged in isolation, and telling a reviewer that the fact that a certain feature does not meet a threshold is reason for a fraud classification negates this isolation.

In our view, the subjective variation 3 offered more information than the eventually chosen, objective, variation 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. These reviewers have this domain knowledge, and are trained to recognise cases that differ from the norm. As such, it makes sense to allow the explanation method to also present this information, to help the reviewer notice a case that might differ from the norm. The argument that this might endanger the objective decision of the reviewers seems implausible, as the reviewers already have this knowledge through their training and domain experience. However, as the organisation prefers erring on the side of caution, variation 2 was eventually selected.

The possibility of presenting groups of features that together may impact the outcome of a case was also explored. Before starting on the eventual prototype, we investigated the possibility of using Anchor or LORE, two non-linear techniques that identify combinations of features that influence the case. Using the full data set available in the mortgage fraud applications, containing 75 features, resulted in performance implications which were unacceptable in the context of the day-to-day tasks executed by the reviewers. However, we found that the running time of Anchor is much lower when using a truncated data set, containing 15 features. Therefore, we extended the prototype. Using SHAP, we find the 15 globally most important features for the model, and then truncate our data set so that we only keep said 15 features. From discussions with reviewers, we concluded that using the top 15 features has no implications on business context, since in practice the reviewers rarely use more than these top features. Then, we train a model on this data set. This truncated model is identical to the original model, so it uses all the same hyperparameters. 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 back to several seconds, versus four minutes per case using the full data set. This aspect of the prototype is shown in Figure 5.

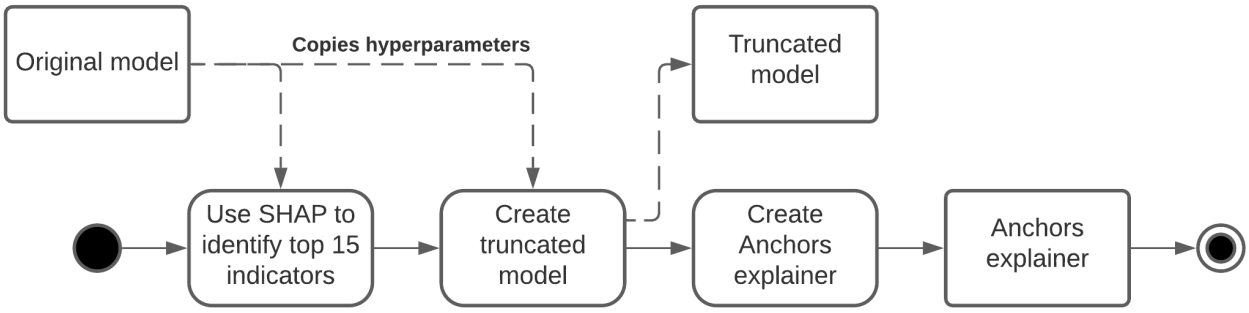

Figure 5: Diagram showing how SHAP is used to find the 15 globally most important features, which are then used to build an Anchor model.

In another attempt to add grouping of features to our prototype, 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 clarify the components and workings of the local prototype, we include a diagram showing how the datasource, SHAP aspect, grouping and Anchor implementation together generate a textual output. The diagram is shown in Figure 6. It illustrates how the pre-trained SHAP and Anchor explainers are used to find the most important indicators and find an Anchor. Then, a dictionary containing template textual explanations for all indicators is queried to generate a textual explanation. Finally, all three explanations are appended, forming the textual explanation shown to the user.

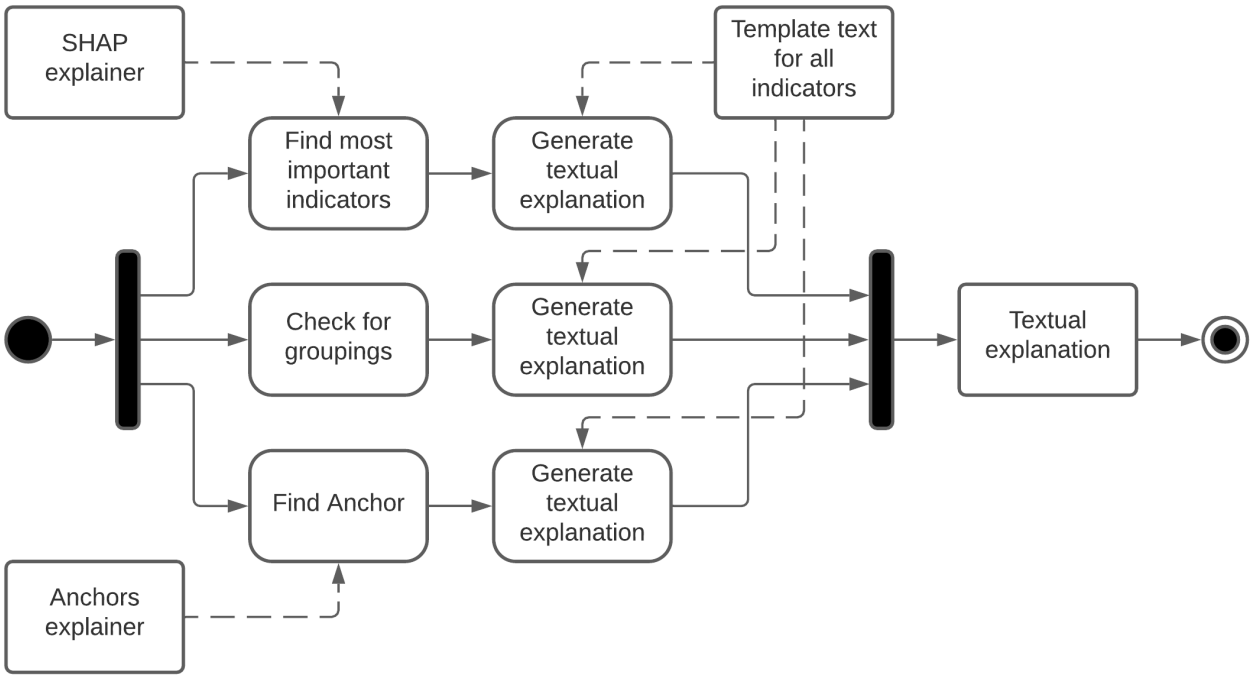

Figure 6: Diagram illustrating the SHAP, grouping and Anchor aspects that together make up the textual explanation for specific cases.

To summarise, the prototype for local interpretability 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 said feature. 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.

4.2.3 Prototype development: global interpretability and bias detection

The three tools selected for global interpretability and bias detection, namely the What If? tool, BlackBoxAuditing and Aequitas, were developed using a different method than the local interpretability prototype. We set out to apply all three tools 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

techniques that were incompatible. Finally, Aequitas also posed no challenges, and was quick to get up and running on our data.

For all three tools, one of the more difficult challenges was properly preprocessing our data. This included making sure the data set was in the correct format, and all feature columns were of the correct type for the tool to be used. However, we opted to create work instructions and tutorial notebooks that should enable data scientists to easily apply the preprocessing techniques on their specific data sets. These work instructions also show data scientists how to apply the tools on any Python classifier in use with the organisation. These work instructions were then used by a data scientist to apply all three tools to a different model and data set, namely one to predict whether a customer is at risk of missing a mortgage payment. Based on their feedback, the work instructions were completed and accepted by the organisation.

4.3 Prototype demonstration

In this subsection, we demonstrate the workings of both the local and global prototype. We illustrate the workings using real life data, using the same mortgage applications which were used for internal presentation and validation of the prototype in the organisation. Where necessary, proprietary information such as fraud indicators designed by the company has been redacted. First, we demonstrate the different possible outcomes of the local prototype, followed by short descriptions and demonstrations for the three tools included in the global prototype.

4.3.1 Prototype demonstration: local prototype

The prototype for local interpretability uses SHAP to obtain Shapley values for each feature in the case of the specific, individual case. We then transform these Shapley values to decimals which indicate the impact on the classification. After obtaining the individual impact for each feature, we can select the features with the highest impact and display them in simple categories as explained in Section 4.2.2. Secondly, the prototype uses SHAP to select the 15 most important features globally, which are then used to create a truncated data set and train an identical model. Next, Anchor is used to generate explanations involving multiple features, to add a non-linear method to the overall explanation delivered to the user. The explanation shown in Figure 7 is for a case used in the validation of the prototype, and was indicated by the trained model to have a higher risk of fraud. However, no fraud was identified in this case.

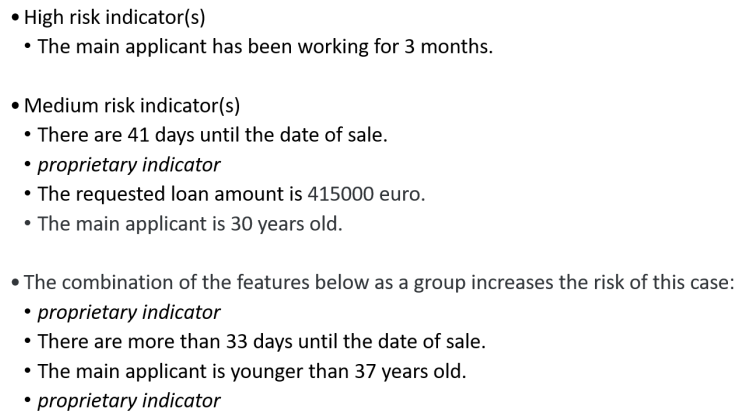

Figure 7: Explanation generated by the local prototype for a mortgage application with elevated fraud risk, but no determined fraud.

The SHAP application, providing textual explanations for five features in two risk groups, gives the user a clear idea of where to start. It makes sense to start at the top, as that aspect of the application supposedly constitutes the biggest fraud risk. This presentation allows the reviewer to simply work down the list, investigating aspects of the case connected to the specific features. However, the combinations of indicators can also pose a risk of fraud, which gives the reviewer extra information in the case these indicators did not pose a risk great enough by themselves. The last paragraph of Figure 7 shows the Anchor output, describing the combinations of features which together pose a risk. The four indicators shown together constitute a fraud risk, according to Anchor. We have redacted two of the four features, since they were proprietary and developed by the organisation. However, we can hypothesise that, if these two redacted indicators said something about the salary of the applicant, this combination of features would help the reviewer investigate aspects of the application specifically regarding the age of the applicant in combination with their salary.

Figure 8 shows another application similar to the first example, classified as a similar risk. However, this case was a proven fraud case. For both cases, the local prototype is able to generate satisfying explanations. Both explanations include features in high- and medium-risk categories, and both have an Anchor explanation. Therefore, since the prototype is able to generate satisfying explanations in both fraud and non-fraud cases, demonstrating it is capable in both situations, it can be concluded that the prototype is ready for testing.

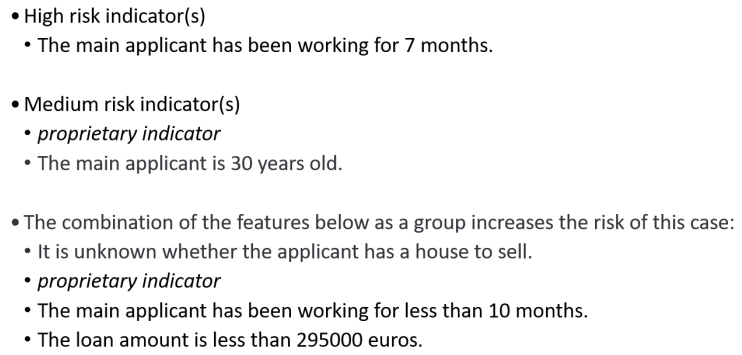

Figure 8: Explanation generated by the local prototype for a mortgage application with elevated fraud risk according to the model. The case was later proven fraudulent.

The local prototype as demonstrated and explained above will be used to validate the approach, and investigate whether the selected techniques improve user trust, satisfaction and efficiency.

4.3.2 Prototype demonstration: global prototype

As mentioned earlier, the global prototype contains three different tools. They are meant to identify bias in model and data, as well as give data scientists better insights into the global decision making of their models. First, we show how BlackBoxAuditing was used to identify features that potentially leak information from protected features. Then, the What If? tool for global insights is briefly shown. Finally, Aequitas extended with CBS data is demonstrated, to illustrate how it would be helpful to identify possible demographic bias in a model.

BlackBoxAuditing
The output generated by BlackBoxAuditing has two main parts. Firstly, the tool generates a 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, the prototype 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 BCR changes less than a certain threshold when omitting the protected features from the model. For example, it will alert the data scientist if the BCR drops less than 0.05 when omitting gender of the applicant. One would logically imagine that the gender of the applicant is quite important for the model. In the case the BCR then does not drop, the data scientist is alerted that the influence of this feature might leak information through other features, for example a correlated feature, such as profession. The change in BCR is also

used to identify features in the model that have only a slight impact on the performance. It lists these features, together with the impact they have on overall model performance. For example, this method applied to the mortgage model illustrated that several Boolean features which are rarely 1 but often 0, have no influence whatsoever on model performance according to BlackBoxAuditing.

Overall, the BlackBoxAuditing tool allows the data scientist to generate feature importance, identify 'protected' features that leak information to correlated features, as well as identify features that contribute very little to the model.

What If? tool

The What If? tool is an interactive visualisation tool that allows the data scientist to visualise the results of their model, and slice, aggregate and zoom in on different groups within features. Furthermore, it offers the ability to adjust properties of an individual case, and run it through the model again. Using the tool, it is also possible to see confusion matrices for specific features, or the overall confusion matrix. Lastly, the data scientist can use the tool to visualise the distribution of features in their data set. The visualisation pane is shown in Figure 9.

Figure 9: The first pane of the What if? tool, visualising a predicted data set. The user sees a scatterplot 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.

The workings and functionality of the tool have not been adjusted in the prototype. The use of this tool in relation to the global prototype 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. This can be done using the overall visualisation, as well as the confusion matrices. Secondly, the tool 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, per Section 3.1). In that case, the specific application can be visualised, and aspects of it tweaked, to see in what way they would have influenced the model decision. Thirdly, the tool can simply be of use to the data scientist in obtaining an overview of the features in the data set, and their distributions.

The What If? tool therefore allows the data scientist to test the global behaviour of the model they are developing, as well as zoom in on specific applications in the case of a client request.

Aequitas

Aequitas is the statistical tool included in the global prototype. It requires no connection to the model, but rather works with a data set enriched with the model decisions as well as the truth label. In our prototype, we have chosen to enrich the data set with aggregated demographic data gathered by the Central Bureau of Statistics (CBS), which contains information on the ratio's of migration backgrounds in areas of 100 by 100 meters. Aequitas can be used to test whether different groups within a feature, for example gender, are treated equally by the model. Our prototype 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.

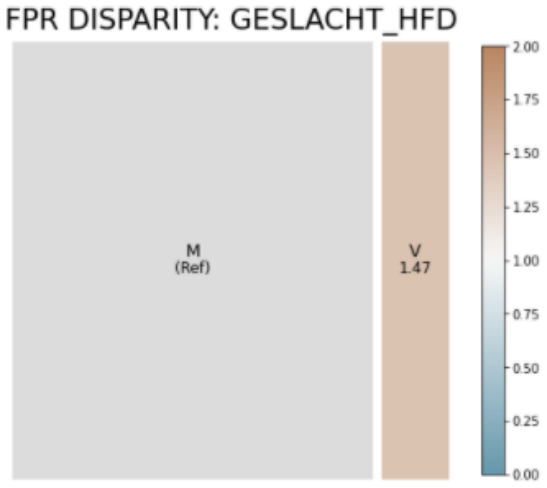

Figure 10: False positive rate disparity by gender, in the mortgage application fraud model. Male (M) is the reference group, with female (V) having a FPR almost 1.5 times higher.

For example, we deployed the tool on the mortgage model, and used it to visualise the FPR disparity for the gender feature, using male (M) as the reference group. This result is shown in Figure 10. Male (M) is the reference group, as it is also the larger group (as apparent by the difference in size of the two areas). Females (V) are shown to have a false positive rate 47% higher than the reference group, implying that their applications are falsely identified as fraud almost 1.5 times as often as applications by males.

However, discrimination based on gender is not legally forbidden. We extended our prototype to use CBS data on migration backgrounds in a given 100 meter by 100 meter area, which allows us to compare false positive rates for applications in which the applicant might live in a neighbourhood with a predominantly non-western migration background. This means we can investigate whether the tested model is fair with regards to migration 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 Figure 11. The hypothetical output shows us that applications which come from a neighbourhood with a predominantly non-western population are falsely classified as fraud 74% more often.

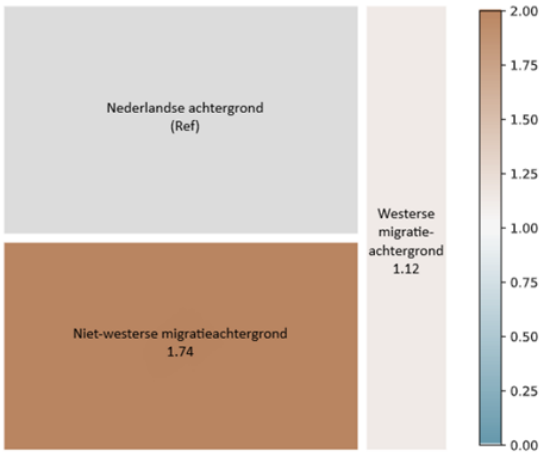

Figure 11: A hypothetical Aequitas output of a test comparing FPR for Dutch, non-western, and western migration backgrounds.

The visualisations which are generated using Aequitas allow for a model to be tested on gender bias, as well as bias stemming from migration background, thanks to 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 fulfill guidelines 19 and 20 of the Ethical Framework (per Section 3.1). Furthermore, the concrete fairness measure resulting from Aequitas can be included in the Project Initiation Document. For example, a threshold can be

defined for bias with regards to migration background. It is impossible for the FPR for two groups to be the same, as there might be underlying, valid reasons for the model to flag a certain group more often than others, for example a generally worse financial position. We can set this example threshold to 5%, meaning any demographical group can only have a FPR that differs a maximum of 5% compared to the group of applicants with a Dutch background. The information in Figure 11 could then be included in the PID, which would show Legal executives that the proposed model exceeds the migration bias threshold devised. This could be a reason to deny the PID, since the model is determined to be unfairly biased with regards to migration background, which is both illegal and ethically undesirable.

Overall, the three tools included in the global prototype allow data scientists to gain more insights into the workings of the model they are building, for example through visualisation, feature importance or possible indirect influence. Furthermore, executives can gain insights into the fairness and possible discriminatory bias of a model they are evaluating. Using these tools, several guidelines from the Ethical Framework can be adhered to.

5 Validation

This section describes the validation of the two prototypes. Since this thesis has two main focuses, local interpretability versus global interpretability and bias detection, the validation of the two prototypes was also split up. This section first covers the validation of the local prototype, followed by the process of validating the tools included in the global prototype.

5.1 Local interpretability

For the first prototype, we opted to validate it using real-life cases presented to reviewers, measuring their experiences with the current explanation method as well as the prototype method. We recruited the help of 11 mortgage application reviewers within the organisation. We evaluated and categorised the different kinds of explanations which could be generated using the prototype and current approach, and from this selected a total of 8 cases, which we found to represent a good cross-section of the possible explanations. For example, some cases had very short explanations using the hand-made rules, but longer explanations with the proposed method (SHAP + Anchor). Also, half of these cases were proven fraud cases, while the other half were not.

5.1.1 Survey design

We presented the participating reviewers with two of the selected cases, which 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 with 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 were swapped around for each new reviewer.
We would ask each reviewer to take in the explanation for both cases like they usually do, and afterwards fill in a short survey consisting of eleven questions, followed by a short (5 minutes) oral review for extra feedback. Based on research by Hoffman et al. [Hof+18] concerning metrics for explainable AI, we designed questions to measure explanation satisfaction and trust, as well as the reviewers' opinions on their speed and accuracy. We also added two specific questions covering the grouping of features and the Anchor component we added. 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 [Hof+18], and uses three questions to query users on their confidence in the model, and whether they feel the model is predictable, reliable, efficient and believable. The level of trust the human reviewers have in the model is of great importance. If the reviewers do not trust the model enriched with explanations more than they trust the current rule-based explanations, then our prototype

has no benefits. However, reviewers must also not put their full trust in the explanations provided, as that would mean that the verification task they are required to do to conform to regulations is not executed properly anymore.

- **Explanation Satisfaction**
  This is defined as the 'degree to which users feel that they understand the AI system or process being explained to them.' [Hof+18]. For this, research by Hoffman et al. [Hof+18] also provides a list of questions, of which three were used. Adequate explanation satisfaction is very important for the success of our prototype. If reviewers are not satisfied by the explanation provided, they will not be inclined to consider the explanations and suggestion given by the prototype, which might result into the explanation not being used, making our prototype unnecessary.

- **Performance**
  The performance of the human reviewers is also important. Ideally, working with an interpretable prototype will improve the performance of the reviewers. To validate this, we used several performance metrics specific to the fraud use case.

  - **Speed** – Does the explanation help the reviewer get started on a case quicker?
  - **Accuracy** – Does the reviewer feel more confident in making a decision on a case thanks to the explanation?

  These two points are important for gauging the practical capabilities of the prototype. If the prototype increases the perceived speed of a reviewer, but negatively impacts their perceived accuracy, it is not successful. Conversely, if the accuracy is improved but speed is not, the prototype needs further development to help the reviewer better understand the explanation quickly.

These three groups of metrics were further developed using research by Hoffman et al. [Hof+18], and eleven questions were set up. These questions took the form of statements, with which the participant could fully disagree or fully agree with on a five point Likert scale. The complete list of survey questions can be found in Appendix A, Table 4.

5.1.2 Survey execution

We were able to run the survey with 11 different reviewers. These reviewers differed in experience and time spent in their role, but all were able to properly complete the survey. The reviewers were explained the survey in a 10 minute session, a week before conducting the survey. This session highlighted the differences between the current method and prototype method, and explained the structure of the survey and the explanations that would be presented to them.

We scheduled each reviewer separately, making sure to explain what we desired them to do and being available while they completed the survey. Afterwards, we completed the short oral part of the research by asking four questions, the first three of which focused on opinions they felt they could not 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.

In all cases, the participants were able to complete the survey and the oral questions in less than five minutes. All reviewers indicated the goal of the research was clear to them. The raw results of this survey can be found in Appendix B, Table 5.

5.2 Global interpretability and bias detection

The three tools selected for global interpretability and bias detection, namely the What if? tool, BlackBoxAuditing and Aequitas, were developed and validated differently from the local interpretability prototype. We created draft work instructions for data scientists, which show them how to apply the tools for their use case on any Python classifier in use with the organisation. These draft work instructions were then used by a data scientist to apply all three tools on a different model and data set, namely one to predict whether a customer is at risk of missing a mortgage payment. Based on their feedback, the work instructions were completed and accepted by the organisation. For these three tools, we demonstrated their use by applying them to the mortgage fraud model, and presenting their results and workings to several different functional user groups.

The proposed tools do not replace a current set of tools, as in the case of the local interpretability prototype. Therefore, it was not possible to compare against a baseline of current practice, and we chose to instead present and demonstrate the tools in two different sessions. In the first session, we presented the technical and application side of the tools, mainly the What if? tool and BlackBoxAuditing, and how these tools could help data scientists understand and prove their models. We collected several feedback points from this demo, which we then processed into a short questionnaire which was answered by the participating data scientists. The questions included in this questionnaire, as well as the raw results, can be found in Appendix C. In the second demo session, we presented the use of Aequitas to several people from Legal, Risk and Compliance departments, including the Data Privacy Officer connected to the team. We focused on the ability of Aequitas to discover false positive rate (FPR) disparity for different demographic groups, as well as the ability to extend the data with CBS data sets, which allows for a model to be tested on FPR disparity for different zip code areas, focusing on the percentage of people in that area with a non-western migration background.

6 Results

In this section, we present the results of the validation of both prototypes. Starting off, we present the results of the survey and interviews with mortgage reviewers, to see what their experiences with the prototype were. Following that, we move on to the results of the small survey following the demonstration of the prototype targeting data scientists, as well as the feedback received after presenting the prototype to the group of managers (Legal, Risk and Compliance). After presenting the results, we discuss what they mean in regards to the goals set out by the organisation as well as what can be said about the impact of the applied techniques on user experience.

6.1 Presentation of results

This subsection will be used to factually present the results from both surveys and demo sessions, for both prototypes. Firstly, we will cover the survey and interview results generated for the prototype for local interpretability, followed by the demos and subsequent surveys for the prototype for global interpretability and bias detection.

6.1.1 Prototype for local interpretability

As described in section 5.1, we presented 11 mortgage reviewers with selected cases, and asked them to compare the explanations generated by the current method with those generated by the prototype as described in Section 4.2.2. 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). The table shows the statements in English, with the original Dutch questions presented in Appendix A, Table 4. Further data displaying means, standard deviations and variances, can be found in Appendix B, Table 6.

Figure 12 also shows the responses to the 11 statements, by presenting a bar chart which shows the frequency of certain degrees of agreement. The underlying raw data can be found in Appendix B, Table 5.

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

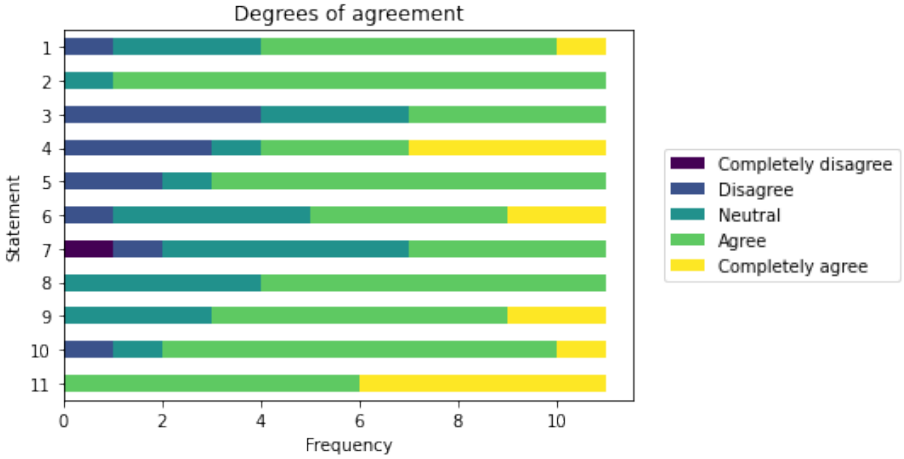

Figure 12: Figure showing the frequency of reviewers indicating certain degrees of agreement to statements, following a Likert scale.

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 5-point Likert 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 prototype explanation (*Completely agree*).

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. In the Discussion section, we go into more detail with regards to specific feedback received.

- **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.

- **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, for example when a second applicant is reportedly earning a set amount but it is unknown for how long they have been employed. Reviewers were confused as to the meaning of 'unknown', whether it meant that the value was unknown or the contribution to the risk was unknown.

- **Overview** (Performance)
  Practically all eleven reviewers were positive about the extra information provided by the prototype explanations. While stressing the importance of their own independent research, they indicated that the prototype provides them with a much better guidance as to what to investigate first compared to the current explanation method, improving their accuracy and possibly saving time.

6.1.2 Prototype for global interpretability and bias detection

In this section, we present the results from the validation of the second prototype, executed using demo sessions and subsequent surveys and feedback sessions. Firstly, we list the feedback gathered from presenting all three tools to the group of data scientists, and the results of a survey for gauging the importance of these points for the data scientists. Secondly, we present the feedback gathered from legal, compliance and risk managers who were presented with the Aequitas tool combined with CBS data.

For the data scientists, we scheduled a one-hour session, in which we used the first 30 minutes to demonstrate the What if? tool, BlackBoxAuditing and Aequitas. We opened the floor to feedback and questions during the presentation, as well as a group discussion after all three tools were demonstrated. From this group discussion, the following four concrete feedback points were distilled.

- **Example applications**
  Even though the use of all three tools was demonstrated using the mortgage fraud model and example cases, the data scientists requested more concrete examples of different applications, for example on different models and applications within the organisation.

- **Concrete thresholds**
  The prototype gives the data scientist certain indications of indirect influence (in the case of BlackBoxAuditing) and the fairness measures (in the case of Aequitas). Given the paperwork and documentation surrounding the development of new and existing model applications, the data scientists would like to have concrete thresholds for these indications, so that they can be easily checked and reported on in the paperwork. These thresholds must be defined by the Compliance department, who must consider what a reasonable fairness threshold would be before it can be implemented in the tool.

- **Clear working instructions**
  Besides the demonstrations, the data scientists indicated that clear work-

ing instructions, accompanied by examples would be helpful in applying the tools on their own models.

- **Choosing fairness measures**
  Aequitas offers many different fairness measures, and while the mortgage fraud model was demonstrated using false positive rate disparity, the data scientists requested more explanations and background with the possible fairness measures, so that they could choose the fairness measure that fits their application.

The four feedback points gathered during the session and listed above were compiled into a short survey which was filled in by the data scientists who also attended the demo session. Table 2 lists the average importance on a scale from 1 (*Not important*) to 5 (*Very important*).

Table 2: Average importance ranking given to the four points gathered during the demonstration to data scientists.

Feedback point	**Average importance (1 to 5)**
Example applications	4.25
Concrete thresholds	4
Clear work instructions	4.5
Choosing fairness measures	3.75

In Figure 13, we display the frequency of the responses gathered. The figure does not show the *Very unimportant* and *Unimportant* options, as their frequency was zero; hence, these were left out, in order to improve readability of the figure.

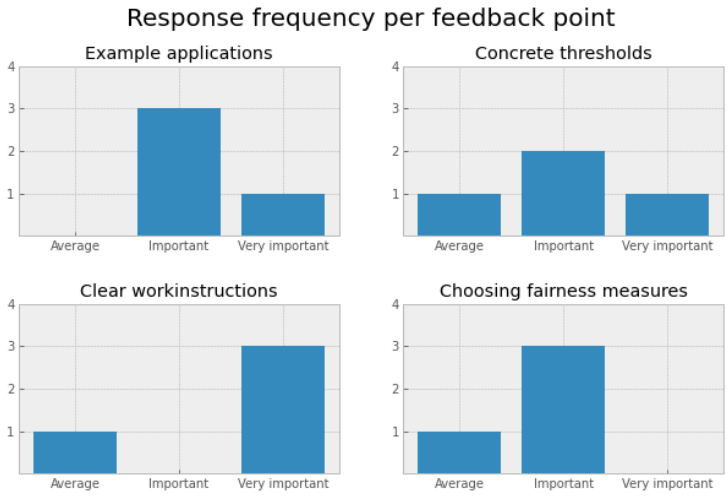

Figure 13: Bar charts showing the frequencies of the responses to the four feedback points. The *Very unimportant* and *Unimportant* responses are not shown, as their frequency was zero for all four feedback points.

The data scientists showed interest in the capabilities of both tools, and indicated that, when the points discussed above could be addressed, the prototype could be a great asset to their toolbox. Using the tools, they feel big steps in quantifying global interpretability and bias could be made, besides granting them more insight into their own models. The data scientists discussed the possibilities of requiring tool output to be included in every PID (Project Initiation Document), as they saw the added benefit of having such a quantified result.

In the second session, we demonstrated the use of Aequitas combined with CBS data to vet an ML system for potential demographic bias. During and after the presentation, we gathered feedback from all participants. All participants reported that a more streamlined Aequitas tool, possibly hosted as a webservice within the organisation, could offer great benefit for many other departments to check their models for demographic bias, since the tool does not require direct model access, merely an output table enriched with zip code data. The reports generated by the Aequitas tool could greatly benefit and streamline the internal paperwork processes regarding the deployment of new or improved models. It was decided that Legal would take the lead on the further development of the tool, and start by investigating the possibilities of the output reports. Further development of the tool has not yet taken place at the time of writing.

6.2 Discussion

In this section, we discuss the results in the context of the hypotheses set out in the introduction. First, we assess both hypotheses, structuring the subsection

accordingly. Then, we evaluate how the both prototypes help the organisation adhere to the guidelines from the Ethical Framework.

6.2.1 Assessing the hypotheses

In order to assess both hypotheses, we first discuss the results as presented in the previous subsection, 6.1. Then, we can assess the individual hypotheses.

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

By developing our prototype using two techniques for local interpretability, SHAP and Anchor, we enabled mortgage application reviewers to gain more insights into the explanations generated by the model, which offers several high fraud risk cases per day. Below, we investigate the effect of the prototype as measured by a short survey comparing the current explanation method versus the prototype method. We enrich the data in Table 1 with information gained from the short interviews. Table 1 contains the 11 statements, which the reviewers answered by indicating their degree of agreement. Degree of agreement is indicated according to a 5-point Likert scale, therefore making the questions Likert-type data, it is recommended to analyse the statements by observing the median degree of agreement indicated [SA13; Jam04]. We separate this investigation by the three metrics that are part of the hypotheses, trust, explanation satisfaction and perceived performance.

Trust
In order to evaluate trust in the model explained using the prototype, we presented the reviewers with three statements, which can be seen in Table 1. The first two statements measure the direct trust the user has, and whether they like using the prototype explanation to base their decision on. We see that the response to **statement 1** is quite positive, with only one reviewer disagreeing with the statement signalling that they do not have trust in the model as a result of the prototype explanation. In the interview, the reviewer further explained that he does prefer the prototype explanation over the current method, but that the output would have to be a lot more concrete and connected to specific risks in the mortgage fraud domain to increase his trust. In Figure 12, we can see that reviewers had differing degrees of agreement to this statement. Overall though, the large number of reviewers answering *Agree* signals that they have more trust in the model using the prototype explanation.

The response to **statement 2** is overwhelmingly positive, with not a single reviewer indicating they dislike using the explanation, and only one answering with *Neutral* (Figure 12). This reviewer is the same reviewer that indicated he does not experience trust in the model as a result of the prototype.

Finally, **statement 3** asks whether the reviewers feel that they could make the correct evaluation of a possible fraud case, by using only the explanation provided by the prototype, without conducting their own independent research. Although the median degree of agreement is *Neutral*, looking at Figure 12 we see that four reviewers *Agree* with the statement, which could indicate their independent, objective verification task as explained in the Ethical Framework [Ver21] is in danger. When reviewers find themselves easily convinced by the explanation, they could potentially neglect their verification of the model suggestion, merely forwarding its decision. This could, in this case, result in incorrect fraud suspicions being placed and applicants could be negatively impacted.

Following the overall answers to the first three statements, we can conclude that the trust in the model did increase with the use of the prototype explanation method. This could simply be due to the fact that the prototype offers far more detail than the existing approach, which leads reviewers to believe the prototype 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 and blindly forward it without performing their own research, for example by requiring them to investigate the underlying properties of the case, as indicated by the top 5 features.

Explanation satisfaction
The degree of satisfaction with the prototype explanation was measured using five statements, for which the responses are presented in Figure 12. **Statement 4** gauges whether reviewers gain insight into the decision process of the model, concretely asking them whether they understand how the model made its decision in the specific case by observing the prototype explanation. The response here is very divided, with the median answer being *Agree* but answers ranging from *Disagree* to *Completely Agree*. This indicates that, while the prototype is an improvement to the current explanation method, more gains might be made, specifically in narrowing the spread of responses. This might be achieved by a better explanation of the overall workings of an ML model, or by creating better explanations for all possible features in the model, as mentioned by several reviewers and noted earlier (Subsection 6.1.1). This would make reviewers more knowledgeable on the way ML models use features, which in turn could help improve their understanding of a model's decision making using those features and explanations the reviewers are provided.

Statement 5 directly measures user satisfaction, asking for overall satisfaction with the prototype explanation. This statement gets a median agreement of *Agree*. Figure 12 shows us that the majority of reviewers (eight out of eleven), indicate that they *Agree* with this statement, while one reviewer answered with *Neutral* and two answered with *Disagree*. These last two reviewers also specifically indicated in the interviews that they struggled with the explanation of

certain features expressing a ratio, as well as features that are marked *'unknown'*. Both of these points were noted in Subsection 6.1.1.

Statements 6 to 8 cover more specific aspects of explanation satisfaction. Overall, users find that the explanation has enough depth and detail, but the decision making process is still unclear to some. This could possibly be mediated by providing users with better information as to the workings of ML models in general, or by improving the context of the fraud indicators used by the model. Users indicated that they find some indicators to be very abstract, and struggle to connect them to concrete risk factors for the specific mortgage application. Interestingly, users report that the prototype improves their understanding of the accuracy of the model assessment. Logically, a reported increased understanding of accuracy is seemingly inconsistent with a disappointing degree of insight into the decision making process. It might be possible that users connect the presence of certain fraud indicators to a higher possibility of a fraud assessment, therefore lowering the value of the other indicators present in the explanation. Further insight into this possibility must be gained in order to make a sound assessment.

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 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 shown when looking at Figure 12, which shows us that the responses and therefore degrees of agreement cover a wide range. 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. Hence, the prototype is more satisfying to use than the existing solution, partly as a result of the mediocrity of the existing solution. 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 the prototype. Three statements, listed with their responses in Table 1, queried reviewers as to their experiences with the prototype regarding the speed of the process, how well informed the reviewer feels and finally the addition of the Anchor component, which allowed for combinations of fraud indicators that together posed a fraud risk.

Statement 9 asks the reviewers whether they feel the prototype allows them to form their decision quicker compared to the current method. The response here is overall positive, indicating a quicker observed handling time for cases. Figure 12 and Table 1 show us a median answer of *Agree*, with no-one answer-

ing below a *Neutral* and several reviewers responding with *Completely Agree*. In the interviews, reviewers also indicated they appreciated the more in-depth explanation provided by the model, as well as the inclusion of all features used in the model.

Statement 10 measured whether reviewers felt more informed using the prototype explanation. Again, the majority of reviewers indicated feeling more informed (*Agree*), with quite a low spread of answers. One reviewer mentioned that he disagreed with this statement, stating that, while the prototype improved this feeling compared to the current method, further improvements could be made. These improvements should cover the connection to the fraud domain, giving the user more information on what they should concretely check.

Statement 11 focuses on the addition of the Anchor component. Reviewers overwhelmingly indicated that they found the addition of groups of indicators that together pose a high risk very informative. This statement received a median answer of *Agree*, with five reviewers answering *Completely Agree*. In the interviews, they again shared this sentiment and mentioned that often they find indicators that pose no risk when viewed individually but when related to other indicators combine to create a fraud risk.

According to the survey and interview results, we can conclude that the prototype has a positive impact on perceived user performance, compared to the current method for generating explanations.

Statistical analysis
As discussed above, the three separate parts of the hypothesis appear to have improved using the prototype explanation method. In order to test whether the prototype overall improved user experience, we can perform a parametric test to investigate whether the found results are statistically significant. According to Norman [Nor10], parametric tests are a valid option in our situation, in which we have Likert-type statements and a small sample size.

We perform a simple t-test, as suggested by Boone and Boone [BB12]. First, we combine all 11 Likert-type statements in our questionnaire into a Likert-scale. The purpose of the questionnaire is to measure whether the prototype positively improves the user experience compared to the current situation. This involves using the mean and standard deviation of the Likert-scale. The Likert-scale can be seen as an interval scale, meaning the *Completely Disagree* gets a score of 1 whereas the *Completely Agree* is seen as a 5. As such, H_0 in this case would be that the prototype is not an improvement over the current situation, in which case all statements would have been answered with *Neutral*, resulting in $\mu = 3$. H_1 states that the prototype is an improvement over the current situation, which would imply $\mu > 3$. In reality, we find that our Likert-scale questionnaire results in $\mu = 3.67$ and $\sigma = 0.86$. Calculating our t-test with a

significance level $\alpha = 0.05$, we find a t-value of 2.5839 and a resulting p-value of 0.013617. Therefore, since $p < \alpha$, we must reject our null-hypothesis and conclude that there is a statistically significant improvement in user experience when using the prototype explanations.

Evaluation of H1
By evaluating the three separate parts of the hypothesis covering user trust, explanation satisfaction and perceived performance, we can now answer our first hypothesis. The response to the presented prototype for local interpretability tested on the mortgage fraud model is overwhelmingly positive. Several statements had a high spread of answers, as seen in Figure 12. Interviews also indicated there is room for improvement, specifically involving explanation satisfaction. This could lower the spread of answers, leading to a more firm overall agreement.

As discussed, the improved level of trust and satisfaction with the explanation might be explained by the current method for generating explanations. The current method does not cover all features, meaning the improvement in trust could be explained by the extra detail offered by the prototype leading reviewers to believe the prototype is more knowledgeable of the underlying model. This in turn causes an increased level of trust in the explanation, but not necessarily in the model decision. Secondly, the rise in satisfaction with the explanation might be explained by the lacking capabilities of the current explanation, which means it is difficult to quantify by how much the prototype really improved user satisfaction.

Nonetheless, we can state that the prototype using techniques that allow for individual predictions to be interpretable and transparent, achieved a statistically significant improvement in trust, satisfaction and usability of ML tools with their daily users, as even the reviewers who gave differing answers indicated experiencing an improvement in using the prototype tool versus the current approach.

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.

The second hypothesis focuses on techniques that make machine learning models globally interpretable and give insights into whether a given model is (demographically) biased. The prototype built for this tasks contains three techniques: the What if? tool and BlackBoxAuditing to improve interpretability, to be used by data scientists, and the Aequitas tool enriched with CBS data to report bias measures to Risk, Compliance and Legal management. For the validation of these tools and to test our hypothesis, we presented the use of these tools to data scientists and management. In this section, we first evaluate the feedback

received from data scientists in order to test the part of the hypothesis concerning global interpretability, followed by the feedback received from management to test the suitability of a tool to test for bias.

Global interpretability
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, which can be seen in Section 6.1.2. These four points were processed into a short survey, which we distributed among the participants of the demo session in order to gauge the importance. These results are shown in Table 2 and Figure 13. Looking a the means of the importance scale displayed in the table, we can make an importance ranking:

1. Clear working instructions
2. Example applications
3. Concrete thresholds
4. Choosing fairness measures

However, we also see in Figure 13 that the data scientists disagreed in variying degrees, given the spread between *Average*, *Important* and *Very important* for all four feedback points. Therefore, it might also be interesting to look at the standard deviation for each feedback point. Table 3 shows the standard deviation per feedback point.

Table 3: Standard deviation for the importance rankings given to the four feedback points gathered during the demonstration to data scientists.

Feedback point	**Standard deviation**
Example applications	0.50
Concrete thresholds	0.81
Clear work instructions	1.00
Choosing fairness measures	0.50

The table shows that the highest ranked point, **Clear work instructions**, also has the highest standard deviation. Looking at Figure 13, we see that this is due to one data scientist ranking this point as average importance, while the others rank this as the highest importance. Given that our sample size is only 4, the standard deviation does not provide too much important information. However, the ranking and standard deviations together show that much of the uncertainties of using the tools could be addressed by developing clear work instructions for the data scientists to use, as well as supplying example applications to inspire the data scientists to use the tools. Based on this feedback, we

developed and evaluated work instructions, which were validated and accepted by the organisation (see Section 5.2). The third most important point, addressing the determining of concrete thresholds for fairness measures and indirect influence, also scores highly. This point directly connects to the hypothesis. In the group discussion, data scientists indicated that concrete thresholds and measures could greatly help them in filling in project documents, streamlining 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 agree on this ranking. While Aequitas offers many different fairness measures, the demo session used false positive rate disparity, a measure with which all data scientists were familiar. Furthermore, given their domain knowledge, it is possible they were already familiar with the other measures.

Bias detection
In the second demo session, we presented the Aequitas tool to Risk, Compliance and Legal management. These participants showed a great response, discussing among themselves the possibilities of including a quantified risk measure in documentation paperwork such as the PID, for new and existing machine learning systems. They shared their struggles with properly understanding and comparing the degree of bias between different systems. Further more, they explained it was difficult to decide what degree of demographic bias was allowed. Hence, they were very glad to have the beginnings of a tool that allowed them to not only understand what demographic groups were disadvantaged, but also understand what indicator in the model was the probable cause for this bias. They decided the next step would be to decide what degrees of bias were acceptable for different indicators, for example gender or probable migration background. If executed properly, they anticipated the addition of Aequitas to the evaluation process of new and adapted systems could help streamline this process.

Evaluation of H2
By evaluating the different parts of this hypothesis with different user groups, we can now assess the second hypothesis. Based on the demo session for data scientists and the four points distilled from the discussion, we can conclude that the inclusion of these tools in their workflows could improve their efficiency and facilitate easier documentation. It would also make the processes involved in bringing a new or adapted model to production easier. The points covering work instructions and the development of proper thresholds can be tackled, as also indicated by the participants in the management demo session. These participants showed great interest in the tool, and decided to tackle the step of setting proper thresholds for different sources of possible bias. The deployment of a working Aequitas tool could help improve and streamline the processes surrounding the deployment of ML tools. Therefore, we can conclude that techniques that allow for ML tools to be globally interpretable and demonstrably bias free 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.

6.2.2 Adherence to the Ethical Framework

As mentioned in Section 3.1, there are a number of ethical guidelines to which the organisation must adhere. In this section, we evaluate to which degree the developed prototypes enable the organisation to follow the relevant guidelines. For each guideline identified earlier, we first discuss whether a certain prototype or tool helped, specifically.

> **(7) The insurer ensures adequate quality (including integrity, correctness, representativeness) of (training) data used for data-driven applications.**
> The global prototype, specifically the Aequitas tool, 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 tools 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.**
> The local prototype 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. The local prototype 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 tools used in the global prototype, specifically the What If? tool and Aequitas, help this process as proven in the last subsection. 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. The local prototype and partly the global prototype offer the ability to provide a client with more information on the results of an application pertaining to their case. On request, the local prototype 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 tools included in the global prototype are used to complete the Project Initiation Document, and evaluation indicated that certain thresholds involving tool output will be developed. Therefore, the global prototype 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 the prototype for global interpretability, helps both the data scientist as well as management 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 the global prototype) 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 the prototypes. However, the tools included in the global prototype will be included in the main artefact of this accountability system, the Project Initiation Document. The outputs of these tools 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 the global prototype) is suitable for interpretation by executives and auditors, as proven in the demo session for

the global prototype 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, as detailed in Figure 2, the organisation adheres to this guideline. The Aequitas tool from the global prototype 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 from the global prototype, 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.

As illustrated above, both the local prototype as well as the global prototype help the organisation adhere to the guidelines set forward in the Ethical Framework for insurers. The global prototype, specifically the Aequitas tool, 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. Therefore, we can conclude that not only do the techniques in the prototype improve user experience and allow for the streamlining of internal processes, they also aid the organisation in tailoring their processes to further adhere to the Ethical Guidelines.

7 Conclusion

This thesis set out to explore the effects of post-hoc, model-agnostic techniques to make black-box machine learning models interpretable, as well as to measure the degree of demographic bias such a system contained. These effects focused specifically on the people working with ML models daily, and how they could benefit from interpretable and demonstrably bias-free systems. In order to explore these questions, two hypotheses were conceived:

1. Techniques that allow for individual predictions to be interpretable and transparent improve trust, satisfaction and usability of ML tools with its 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.

In cooperation with a large Dutch insurance organisation, a case study using a model for estimating fraud risk for mortgage applications was performed. This model gives 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, addressing the points mentioned in the explanation. The existing method for explanations was purely rule-based, and was replaced by a prototype method using a combination of SHAP [LL17] and Anchor [RSG18] to provide detailed explanations containing all possible features and high risk combinations of features in the model. This prototype was validated using a survey and short interviews with eleven reviewers from within the organisation, who were queried for trust, explanation satisfaction and perceived performance with the prototype explanations. It was found that the prototype method achieved a statistically significant improvement in all three measures, leading us to conclude that the first hypothesis is correct.

Another prototype containing work instructions and examples for three tools (the What if? tool [Wex+20], BlackBoxAuditing [Adl+18] and Aequitas [Sal+18]) was developed. The potential of these tools was demonstrated to data scientists and representatives from risk, legal and compliance departments. All parties agreed that the concrete insights that could be achieved using these tools would enable them to streamline processes concerned with the initiation or adaptation of models, specifically by incorporating measures into the Project Initiation Document, a document used to document and validate a model, containing information on its decision making process and degree of bias. These findings lead us to believe the second hypothesis is also correct.

Furthermore, we concluded that the prototypes aided the organisation in adhering to guidelines concerning the ethical use of machine learning in the insurance domain, covering transparency, accountability and technical quality [Ver21].

Specifically, the prototypes aided increased adherence to Ethical Guidelines 7, 14, 15, 18, 19, 20, 21, 23, 24, 25, 26.

7.1 Academic relevance

The findings from the research described in this thesis confirm the potential of tools for interpretable AI in practice. The opportunity to implement and validate these tools in a business environment, with a wide range of different stakeholders, some familiar with machine learning and others not, validates the fact that these tools fulfil their purpose, namely to make ML systems more interpretable, both on a local and global level.

It was found that deploying a prototype in a business environment requires making a trade-off between performance and extensiveness, as demonstrated by the need to truncate the data set to the top 15 features for use in the Anchor model. More so, in cases where the decision making process is not impacted, as in this use cases where reviewers rarely use any feature outside this top 15, the gain in execution time is essential. Furthermore, the business does not always want all the possible information offered by a technique for interpretable ML, as it may impact the objective task executed by the reviewers. The business is fearful of influencing the reviewer too much, endangering the human oversight over model decisions. Experiments with KernelSHAP versus TreeSHAP showed that, even though the two use different underlying approaches, they achieve an overlap of 80% in the top five most important features.

Even in cases with mature ML deployment processes, efficient use of existing techniques for interpretable ML and bias detection can uncover previously unknown biases, such as the identification of significant bias against female applicants in the mortgage process. Prototypes incorporating these techniques offer new insights into existing models, forcing stakeholders to see bias in a new light, namely by setting thresholds on the degree of bias instead of viewing it as a binary decision. Overall, it can be concluded that efficient tooling for interpretable ML enables an organisation to improve the communication process between different stakeholders within the organisation, when it comes to deploying ML models. The data scientists responsible for the development of the model are able to better deliver documentation regarding bias and interpretability to stakeholders responsible for the acceptance of the model, such as officers from Legal, Compliance or Risk departments. These acceptance stakeholders in turn are able to better understand the model, therefore enhancing the overall acceptance process for new and existing ML systems.

Furthermore, the methodology described in this research allows for the application of prototypes using techniques for interpretable ML in other domains, where similar systems are in use. The confirmation that existing techniques can be applied to any ML model to create global interpretability as well as gain insights into possible bias means systems such as SyRI can be verified and tested

before being put into practice, which might prevent people from being falsely investigated for fraud.

7.2 Limitations

There are several limiting factors that hampered this research. First off, due to the fact that the organisation does not actively track whether a mortgage application suspected of fraud by the model was later confirmed to be fraud, we were unable to verify whether reviewers were able to correctly identify fraud cases. This meant we could not ascertain whether the prototype for local interpretability improved the accuracy of reviewers.

Due to limited access to reviewers, we were able to test only one prototype for local interpretability. As described earlier, the organisation preferred to test the 'objective', purely factual version we developed. We would have preferred to test the results of the 'subjective' version as well, which included contextual information helping reviewers place the case among other applications. However, the limited access to reviewers meant we were only able to verify the objective version.

The prototype for local interpretability uses SHAP to generate feature importance. In this research, we identified that TreeSHAP enables a large improvement in complexity, while being comparable to KernelSHAP in performance. The mortgage application uses a tree ensemble, meaning we were able to apply TreeSHAP. Although the general prototype described in this research is still applicable for other models, the necessity to use KernelSHAP will limit performance, as the increased complexity will cause higher running times.

As for the global prototype, the chosen bias method, Aequitas, is not yet fully in production. However, we were able to confirm the potential of the tool with all involved parties.

Furthermore, privacy and data protection laws might have limited the precision of this tool. By using the CBS data to gather information on the composition of different migration background based on zip code, we were able to determine whether a model was more biased towards applicants in areas with a higher non-western migration background. However, the fact that an applicant is located in an area with a higher percentage of non-western migration backgrounds does not mean the applicant's background is also non-western. This is a limitation that must be kept in mind when interpreting the results of the tool.

7.3 Future work

Based on findings in this research, we recommend several areas for further study. Firstly, future work could investigate the difference achieved when applying the prototype for local interpretability as suggested in this research in a situation

where a robust baseline is already in place. For example, it would be interesting to see the gains in trust and explanation satisfaction in a business situation where a limited textual explanation (based on SHAP or LIME [RSG16]) is already present, compared to the SHAP and Anchor combination used by our prototype for local interpretability.

Secondly, developing a method to quantify the increase in trust and gains, versus the current situation of being able to establish that there was an increase would be helpful. This could help compare different methods for local interpretability, by seeing by how much they are able to improve user experience. This method could then be used to determine what is more beneficial to the reviewers, a 'subjective' method containing contextual information, or a purely factual, 'objective' method. Furthermore, it would help to be able to also rate methods on the performance gains they enable. By validating methods using a data set that includes the eventual fraud verdict, so whether a case was found to be fraudulent or not, it would be possible to determine by how much a method helps reviewers in making the right decision. This would be done by seeing whether applications explained with the proposed method helped reviewers correctly judge given mortgage applications.

Thirdly, further research into quantifying degrees of bias could be done. In our research reported here, we developed a method that could determine whether or not a model is negatively biased towards applications by clients living in predominantly more non-western neighbourhoods. This could be improved by establishing a data set containing exact migration background. This data set could be fully anonymous and controlled by a central, independent organisation, for example the Dutch Association of Insurers. Using this data set, all insurers would be able to test their models for bias using the same data, effectively creating an audit for all models.

Finally, research into the impact of manually created grouping could be undertaken. This research attempted to evaluate whether manually created groups, based on domain knowledge, were helpful to reviewers. However, this question received varying responses, making drawing a conclusion impossible. This research could be expanded by using the knowledge of domain experts to create a number of groups of indicators that are related to one another, and evaluate whether reviewers find these manual groups, as opposed to automatic groups as identified by Anchor, helpful in their research.

Lastly, the further application of SHAP and Anchor as a combination for local interpretability could be explored. It would be interesting to see whether other domains benefit from the truncated Anchor addition, and whether the explanation provided by both techniques combined is as high quality as it was in our research. Furthermore, the fact that the truncated version of Anchor lost no business value in our case, does not mean this translates to other domains. It would be interesting to explore this further.

References

[AB18] A. Adadi and M. Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

[Adl+18] Philip Adler et al. "Auditing black-box models for indirect influence". In: *Knowledge and Information Systems* 54.1 (2018), pp. 95–122.

[ANP20] ANP. "Overheid stopt met gebruik SyRI na uitspraak rechter". In: *Het Parool* (Feb. 5, 2020). URL: https://www.parool.nl/nieuws/overheid-stopt-met-gebruik-syri-na-uitspraak-rechter~bbf3993a/ (visited on 02/04/2021).

[BB12] Harry N Boone and Deborah A Boone. "Analyzing Likert Data". In: *Journal of Extension* 50.2 (2012), pp. 1–5.

[BKB17] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. "Interpretability via model extraction". Presented as a poster at the 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017). 2017.

[CH20] Simon Caton and Christian Haas. "Fairness in Machine Learning: A Survey". Pre-print, available at https://arxiv.org/abs/2010.04053. 2020.

[CMB18] Giuseppe Casalicchio, Christoph Molnar, and Bernd Bischl. "Visualizing the feature importance for black box models". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2018, pp. 655–670.

[Den19] Houtao Deng. "Interpreting tree ensembles with intrees". In: *International Journal of Data Science and Analytics* 7.4 (2019), pp. 277–287.

[DSB18] D. Doran, S.C. Schulz, and T. R. Besold. "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives". In: Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017. CEUR, Mar. 2018. URL: https://openaccess.city.ac.uk/id/eprint/18660/.

[Gui+18] Riccardo Guidotti et al. "Local rule-based explanations of black box decision systems". Pre-print, available at https://arxiv.org/abs/1805.10820. 2018.

[Gui+19] Riccardo Guidotti et al. "A Survey of Methods for Explaining Black Box Models". In: *ACM Computing Surveys* 51.5 (Jan. 2019), pp. 1–42. DOI: 10.1145/3236009. URL: https://doi.org/10.1145/3236009.

[Hof+18] Robert R. Hoffman et al. "Metrics for Explainable AI: Challenges and Prospects." Pre-print, available at https://arxiv.org/abs/1812.04608. 2018.

[Jam04] Susan Jamieson. "Likert scales: how to (ab)use them". In: *Medical Education* 38.12 (Dec. 2004), pp. 1217–1218. DOI: 10.1111/j.1365-2929.2004.02012.x. URL: https://doi.org/10.1111/j.1365-2929.2004.02012.x.

[Lak+19] Himabindu Lakkaraju et al. "Faithful and customizable explanations of black box models". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 131–138.

[LL17] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4768–4777. ISBN: 9781510860964.

[Meh+21] Ninareh Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". In: *ACM Computing Surveys* 54.6 (July 2021). ISSN: 0360-0300. DOI: 10.1145/3457607. URL: https://doi.org/10.1145/3457607.

[Mol20] Christoph Molnar. "SHAP (SHapley Additive exPlanations)". In: *Interpretable machine learning: A guide for making black box models explainable*. Leanpub, 2020.

[MQB18] Yao Ming, Huamin Qu, and Enrico Bertini. "Rulematrix: Visualizing and understanding classifiers with rules". In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2018), pp. 342–352.

[MS21] Milad Moradi and Matthias Samwald. "Post-hoc explanation of black-box classifiers using confident itemsets". In: *Expert Systems with Applications* 165 (2021), p. 113941.

[Nor10] Geoff Norman. "Likert scales, levels of measurement and the "laws" of statistics". In: *Advances in Health Sciences Education* 15.5 (Feb. 2010), pp. 625–632. DOI: 10.1007/s10459-010-9222-y. URL: https://doi.org/10.1007/s10459-010-9222-y.

[PB19] Eliana Pastor and Elena Baralis. "Explaining black box models by means of local rules". In: *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. 2019, pp. 510–517.

[pym] pymetrics. *AuditAI*. URL: https://github.com/pymetrics/audit-ai.

[Rai20] Arun Rai. "Explainable AI: From black box to glass box". In: *Journal of the Academy of Marketing Science* 48.1 (2020), pp. 137–141.

[RBB20] Dilini Rajapaksha, Christoph Bergmeir, and Wray Buntine. "LoRMIkA: Local rule-based model interpretability with k-optimal associations". In: *Information Sciences* 540 (2020), pp. 221–241.

[Ros+20] R. Roscher et al. "Explainable Machine Learning for Scientific Insights and Discoveries". In: *IEEE Access* 8 (2020), pp. 42200–42216. DOI: 10.1109/ACCESS.2020.2976199.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""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*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.

[RSG18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 2018.

[SA13] Gail M. Sullivan and Anthony R. Artino. "Analyzing and Interpreting Data From Likert-Type Scales". In: *Journal of Graduate Medical Education* 5.4 (Dec. 2013), pp. 541–542. DOI: 10.4300/jgme-5-4-18. URL: https://doi.org/10.4300/jgme-5-4-18.

[Sal+18] Pedro Saleiro et al. "Aequitas: A bias and fairness audit toolkit". Published as a background and development guide, available at https://arxiv.org/abs/1811.05577. 2018.

[Shi21] Donghee Shin. "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI". In: *International Journal of Human-Computer Studies* 146 (2021), p. 102551.

[Tra+17] Florian Tramer et al. "Fairtest: Discovering unwarranted associations in data-driven applications". In: *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE. 2017, pp. 401–416.

[Ver21] Verbond van Verzekeraars. "Ethisch Kader". 2021. URL: https://www.verzekeraars.nl/media/7541/ethisch-kader.pdf (visited on 09/14/2021).

[Vid+16] Marina M-C Vidovic et al. "Feature importance measure for non-linear learning algorithms". In: *Proceedings of the 30st International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain, 2016.

[Wex+20] James Wexler et al. "The What-If Tool: Interactive Probing of Machine Learning Models". In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), pp. 56–65. DOI: 10.1109/TVCG.2019.2934619.

Appendices

A Survey questions

In this Appendix we list all statements posed to the mortgage reviewers during the survey. Table 4 below lists the statements in Dutch, as well as their English translation.

Table 4: Degree of agreement for all questions covering trust in the model, explanation satisfaction and whether the proposed explanation would help the reviewer complete their task.

Dutch statement	English translation
Ik heb vertrouwen in de uitleg. Ik heb het gevoel dat het model goed werkt.	I trust the explanation. I feel like the model is working well.
Ik gebruik de uitleg graag om beslissingen te maken.	I like using the explanations to make decisions.
Ik heb het gevoel dat wanneer ik alleen deze uitleg gebruik ik de juiste beslissing zal maken.	I feel like I will make the correct decision only using this explanation.
Door de uitleg begrijp ik hoe het model tot zijn oordeel komt.	The explanations make me understand how the model reaches its judgement.
Ik ben tevreden met de uitleg.	I am satisfied with the explanation.
De uitleg van de uitkomst heeft voldoende detail.	The explanation is sufficiently detailed.
De uitleg van hoe het model werkt lijkt compleet.	The explanation on how the model works seems sufficient.
De uitleg van de uitkomst vertelt mij hoe nauwkeurig het model is.	The explanation of the result tells me how accurate the model is.
Ik kom sneller tot een beslissing omdat ik de uitleg bij een case heb.	I reach a decision quicker because the case has an explanation.
Ik kan een beter geinformeerde beslissing maken omdat ik de uitleg bij een case heb.	I am able to make a better informed decision because the case has an explanation.
Ik vind de toevoeging van combinaties van risicoindicatoren bij een uitleg belangrijk.	I find the addition of combinations of risk indicators to the explanation important.

B Survey results

In Table 5 below we list the raw results of the validation of the local prototype. For each of the statements, the degree of agreement for all reviewers is noted. In bold we show the degree of agreement which the highest number of reviewers indicated.

Table 5: Raw results for all statements used in the survey to validate the prototype for local validation. The mode for each statement is marked in bold.

Statement	**Degree of agreement**				
	Completely disagree	Disagree	Neutral	Agree	Completely agree
Trust in the model					
1 - I trust the explanation. I feel like the model is working well.	0	1	3	**6**	1
2 - I like using the explanations to make decisions.	0	0	1	**10**	0
3 - I feel like I will make the correct decision only using this explanation.	0	**4**	3	**4**	0
Explanation satisfaction					
4 - The explanations make me understand how the model reaches its judgement.	0	3	1	3	**4**
5 - I am satisfied with the explanation.	0	2	1	**8**	0
6 - The explanation is sufficiently detailed.	0	1	**4**	**4**	2
7 - The explanation on how the model works seems sufficient.	1	1	**5**	4	0
8 - The explanation of the result tells me how accurate the model is.	0	0	4	**7**	0
Performance					
9 - I reach a decision quicker because the case has an explanation.	0	0	3	**6**	2
10 - I am able to make a better informed decision because the case has an explanation.	0	1	1	**8**	1
11 - I find the addition of combinations of risk indicators to the explanation important.	0	0	0	**6**	5

Table 6 shows the mean degree of agreement per statement, as well as standard deviation and variance.

Table 6: Average answer, standard deviation and variance for all questions covering trust in the model, explanation satisfaction and whether the proposed explanation would help the reviewer complete their task.

Statement	**Mean (1 to 5)**	**Std. dev.**	**Variance**
Trust in the model			
1 - I trust the explanation. I feel like the model is working well.	3.64 (Agree)	0.77	0.60
2 - I like using the explanations to make decisions.	3.91 (Agree)	0.29	0.08
3 - I feel like I will make the correct decision only using this explanation.	3.00 (Neutral)	0.85	0.73
Explanation satisfaction			
4 - The explanations make me understand how the model reaches its judgement.	3.73 (Agree)	1.21	1.47
5 - I am satisfied with the explanation.	3.55 (Agree)	0.78	0.61
6 - The explanation is sufficiently detailed.	3.64 (Agree)	0.88	0.78
7 - The explanation on how the model works seems sufficient.	3.09 (Neutral)	0.90	0.81
8 - The explanation of the result tells me how accurate the model is.	3.64 (Agree)	0.48	0.23
Performance			
9 - I reach a decision quicker because the case has an explanation.	3.91 (Agree)	0.67	0.45
10 - I am able to make a better informed decision because the case has an explanation.	3.82 (Agree)	0.72	0.51
11 - I find the addition of combinations of risk indicators to the explanation important.	4.45 (Agree)	0.50	0.25

C Data scientist survey

In Table 7 the four statements used in the short survey for the data scientists are listed. We show the statements in Dutch, as well as their English translation.

Table 7: Feedback points gathered during the data scientist demo, and how they were posed in the survey, along with their English translation.

Feedback point	**Dutch statement**	**English translation**
Example applications	Het beschrijven van een duidelijke toepassing voor de gepresenteerde tools.	Describing clear applications for the presented tools.
Concrete thresholds	Het ontwikkelen van duidelijke thresholds en eisen voor fairness measures (in het geval van Aequitas) en indirecte invloed (in het geval van BBA), zodat de tools een duidelijke bijdrage kunnen leveren aan bijvoorbeeld een PID.	Developing concrete thresholds and standards for fairness measures (in case of Aequitas) and indirect influence (in case of BBA), so that the tools can make a clear contribution to for example a PID.
Clear work instructions	Het ontwikkelen van een werkinstructie aan de hand van een voorbeeld model, wat ik als voorbeeld kan gebruiken bij het toepassen van de tools op mijn eigen modellen.	Developing a workinstruction with an example model, which I can follow in applying the tools on my own models.
Choosing fairness measures	Het onwikkelen van duidelijke uitleg bij het kiezen van de juiste fairness measures (bij Aequitas) die het meest toepasselijk zijn op mijn model en usecase. Daarmee zou de keuze tussen False Positive Rate disparity, False Discovery Rate parity, False Omission Rate disparity, etc. makkelijker zijn.	Developing a clear explanation to help choose the proper fairness measures (for Aequitas) which are most applicable to my model and usecase. This would make the decision between False Positive Rate disparity, False Discovery Rate disparity, False Omission Rate disparity, etc. easier.

In Table 8, we display the raw resuls gathered during the short survey. For each of the four points, we show the different important rankings given by the surveyed data scientists.

Table 8: Raw results gathered concerning the importance of the four points gathered during the demonstration to data scientists.

	Importance				
Feedback point	Very unimportant	Unimportant	Average	Important	Very important
Example applications	0	0	0	3	1
Concrete thresholds	0	0	1	2	1
Clear work instructions	0	0	1	0	3
Choosing fairness measures	0	0	1	3	0