Understanding Data Scientists' use of Explainability and Interpretability tools

MSc. Thesis Computer Science

Purvesh Baghele

Understanding Data Scientists' use of Explainability and Interpretability tools

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Purvesh Baghele
born in Nagpur, India

Multimedia Computing Group
Department of Intelligent Systems
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
www.ewi.tudelft.nl

AI for Fintech Research
ING Bank N.V.
Frankemaheerd 1
Amsterdam, the Netherlands
www.ing.nl

Understanding Data Scientists' use of Explainability and Interpretability tools

Author: Purvesh Baghele
Student id: 5087589
Email: p.baghele@student.tudelft.nl

Abstract

Machine learning techniques are being used increasingly in high-risk domains such as healthcare, law, and finance. Misclassifications or mispredictions in areas like these can have serious consequences. Therefore, it is crucial to have high performing models that can make as minimal errors as possible. However, in some cases, it is not enough to know what the predictions and classifications are, it is also important to understand why a given model is making certain decisions. Explainability and interpretability tools help data scientists in understanding how an output is obtained by a machine learning model from a given input.

Model explainability and interpretability tools like SHAP (Shapley Additive Explanations) and GAMs (Generalised additive models) are becoming widely used. However, user studies that evaluate the extent to which these tools help data scientists interpret and explain their models are still uncommon. Our study attempts to find out how data scientists use these tools – their general behaviour, factors affecting their trust in the tools, and their expectations from the tools. We conduct a qualitative study consisting of Pilot interviews and Contextual inquiries. By analyzing the results, the study shows that the tools sometimes are not used at their full potential. Moreover, some participants expressed the need to also align with other stakeholders to get the full picture on the explanations and trust them. The most important factors affecting their trust were the lack of clarity in understanding the visualisations that further lead to the participants reasoning intuitively and having suspicions about the explanations, dataset, and the underlying model.

Thesis Committee:

Chair:	Prof. Dr. A. van Deursen, Faculty EEMCS, TU Delft
University supervisor:	Dr. Cynthia Liem, Faculty EEMCS, TU Delft
Company supervisor:	Dr. Flavia Barsotti, Research Coordinator, ING Bank and Scientific Fellow, IAS (Institute of Advanced Studies), University of Amsterdam

Preface

I would like to thank my supervisor Dr. Cynthia Liem for her invaluable guidance and great discussions during this thesis project. Special thanks to Dr. Flavia Barsotti for her constant availability, indispensable feedback, and support in reaching out to the participants for our study. It is a great experience to work with both of you.

I would also like to thank Elvan Kula for giving me the opportunity to pursue my thesis project at ING. I would like to thank all my amazing colleagues at ING and AI for Fintech Research for all the important feedback they gave me during my research.

Lastly, I would like to thank my family and friends for constantly keeping me motivated and enthusiastic during my studies. I especially want to thank my Mom and Dad for having faith in me and motivating me in taking the road less travelled if it is worth it.

Purvesh Baghele
Delft, The Netherlands
August 13, 2021

Contents

List of Figures

List of Tables

Glossary

1. ML: Machine learning
2. SHAP: Shapley Additive Explanations
3. GAM: Generalized Additive Model
4. EBM: Explainable Boosting Machines
5. RQ: Research Question

Chapter 1

Introduction

1.1 Motivation

Machine Learning (ML) and Artificial Intelligence (AI) are getting used everywhere today. They are also extensively used in sectors like banking, healthcare, and law [41, 8, 59]. In banking, machine learning is used to do credit-risk analysis, fraud detection, marketing etc. [15, 6, 51]. Other examples of the use of machine learning include detecting hard to discern patterns for cancer diagnosis and detection, predictive analytics, and even predicting outcomes in legal cases [13, 42, 4]. One common theme when using machine learning in all these domains is to do predictions or classifications. A machine learning model gets evaluated using some performance metrics after the model-training process. Different types of metrics are used for regression and classification problems. Regression algorithms predict continuous values e.g. house prices based on input features like house size, location etc. On the other hand, classification algorithms categorize the data into different classes, for example spam or no spam emails based on the text in the email. The underlying commonality between most of the metrics is that they compare the predictions/classifications of the model to the ground-truth value. Mean squared error (MSE), which measures how accurately a machine learning model can predict the expected outcome, calculates the average squared difference between the predicted value and the actual value [29]. Formally:

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y_j})^2 \tag{1.1}$$

where N is the total number of observations, y_j is the ground-truth value, and $\hat{y_j}$ is the predicted value by the model. Similarly, Mean absolute error [9] calculates the average absolute difference between the predicted and the actual value:

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y_j}| \tag{1.2}$$

Logarithmic loss (Log Loss) is a popular performance evaluation metric for classifica-

tion algorithms. Log loss penalises incorrect classifications. Formally:

$$LogLoss = \frac{-1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} * log(p_{ij}) \tag{1.3}$$

where N is the number of observations, M is the number of classes (into which the classification algorithm would categorize the observations), y_{ij} is the indicator for observation i belonging to j^{th} class, and p_{ij} is the probability of observation i belonging to j^{th} class. The smaller the Log loss the higher the accuracy. Other popular metrics like Precision, Recall, F-score, and Area under curve (AUC) [45] can also be used to measure the performance of a given classification algorithm.

Although these metrics are good to evaluate the performance of the model, there are other factors that are difficult to translate into mathematical functions. Consider an algorithm used to determine bail decisions as an example. This algorithm should also optimize ethics, fairness, and transparency. These aspects are difficult to quantify. Interpretability of a model and being able to explain its output to stakeholders for better transparency could help in establishing trust in our models – that they are making the right decisions under the right set of assumptions. It could also help us in diagnosing what went wrong if our model fails. Thus, explainability and interpretability could help us in ensuring that these unquantifiable aspects are achieved.

Interpretability is defined by Miller [36] as "the degree to which a human can understand the cause of a decision." It is achieved when humans can understand the internal working of the model. Montavon et al. [39] define an interpretation as "the mapping of an abstract concept (e.g. a predicted class) into a domain that the human can make sense of", e.g., images and text qualify as the domain which are interpretable as opposed to things like a word vector [33]. They further define an explanation as "the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g. classification or regression)". Examples include a highlighted text in natural language processing use cases that explains the classification. Explainability focuses on explaining how a specific input has contributed to a given output in a human-understandable way.

The terms explainability and interpretability seem relatively close in literature. Ribeiro [47] states that "an essential criterion for explanations is that they must be interpretable". Gunning et al. [20] argues that definitions of explainability and interpretability are domain-dependent; when models are fully interpretable, they give full and completely transparent explanations. One difference between explainability and interpretability can be found through the trends in literature available on these topics. Research around explainability focus more on communicating the behaviour of the model to the people more interestingly and engagingly [54, 32, 49], whereas research around interpretability deals with designing models which can explain themselves [60, 11, 18].

Nevertheless, both explainability and interpretability contribute to answering why a model is making certain decisions. Molnar [38] tries to establish why interpretability is important by trying to answer two questions: Do we want to know **what** is predicted? or Do we want to know **why** a certain prediction has been made? Similarly, Doran et al. [16] also advocates the use of explainability for answering **why** a prediction was made. The

answers to these questions, we believe, depend on the use case. If the use case is predicting the steering angle of a self-driving car by identifying the road lines then the accuracy of predictions is very important to ensure safety. In this case, knowing the "what?" is very important and it might not be necessary to consider the "why?" as long as the self-driving car does a good job of staying on the track. A more sensitive case could be deciding whether some applicants should receive a loan based on their transaction history. In this case, if data scientists are not careful about identifying why their model is making certain decisions, then a possible risk could be that the model makes decisions based on protected features like Age and Gender. Answering the "why?" along with the "what?" becomes important here. Even in the self-driving car case, if the car persistently predicts the wrong steering angle even after having a good accuracy on the test set, then data scientists would want to know why such predictions are made. In [17], the authors suggest that if data scientists can understand how the model makes decisions (thus, answering why a certain prediction is made) , then the following aspects can be addressed:

- Fairness: ensuring that our model is not biased towards an underrepresented group
- Privacy: ensuring that our model does not use any protected features while making decisions (e.g. using race to determine criminal recidivism)
- Trust: if the model is able to explain its decisions then it is more likely for a human to trust it
- Robustness: ensuring that the explanations are consistent

In literature and in practice, two widespread approaches are considered for explainabilty and interpretability: 1. Use model-agnostic post-hoc explanations and 2. Use an inherently interpretable model. Model-agnostic approaches deal with explaining the output of any black box model (hence model-agnostic). Black box models are machine learning models created by an algorithm that takes input data and output predictions/classifications based on the interactions of input variables. These interactions could become so complex that it becomes very difficult for a human to comprehend how these variables are combined to make a decision. It could still be possible to understand the working of these complex models by knowing the variables they use. One could also see the feature weights attributed by the model to understand how a decision is made. However, some machine learning models could be proprietary and we may not be allowed to see their internal working or even query the model [52]. Post-hoc explanation tools try to approximate the behaviour of a black box model by studying the relationships between the input variables and the predictions [40]. A popular post-hoc model-agnostic explainability tool is LIME (Local Interpretable Model-agnostic Explanations). LIME considers a single data point and perturbs the input feature values and observes the resulting impact on the output [48]. Another such tool is SHAP (Shapley Additive Explanations) which calculates the contribution of each feature to the final prediction by comparing the value of that prediction to the value of the same prediction at the baseline value of the features [31] (see Section 2.3 for more details).

Alternatively, we can achieve interpretability by designing models which are glass-box models. These models are called glass-box because these are inherently interpretable hence,

data scientists can understand how the model is making decisions. Examples of such models include decision trees, linear regression, and GAMs (Generalized Additive Models) [22]. Predictions and classifications by decision tree can be seen as a set of if-then rules. GAMs learn a different function for each feature. Therefore, we can see the behaviour of each feature towards the final prediction/classification.

The problem that is the primary focus of our study is - How do data scientists working with these tools use them? People can have their own biases and perceptions while using these tools and the tools themselves could be misleading as well. Post-hoc explanation techniques could potentially mislead the decision-maker. It has become apparent that these explanations are inconsistent in the sense that they change very quickly even with a small change in the input [24]. There could also be multiple explanations that seem qualitatively different for the same black box [41]. Lakkaraju and Bastani [40] showed that high fidelity explanations (how well does the explanation approximate the prediction of the black box model [38]) may not accurately reflect the bias in the black box model. The authors created high fidelity explanations for a black box that used protected features like race and gender to predict if a defendant should be granted a bail or not. Later in the same work, a user study showed that these high fidelity explanations lead to participants trusting the black box (which used protected features for decisions) whereas they did not trust the same black box without the explanations. In a similar study, Slack et al. [56] showed that it is possible to hide the biases of any classifier by allowing an adversary to craft any arbitrary desired explanation by introducing perturbations in the input. The study demonstrated with the example of the German credit dataset [5] that input can be perturbed by an adversary which could lead to explanations from popular post-hoc tools like LIME and SHAP to show uncorrelated features (like Loan Rate % Income) as the most important ones, whereas in reality, the most important features would've been protected features like Gender. Moreover, even when the explanations are totally correct, the way users reason the output from intepretability and explainability tools could lead to further issues with trust and fairness. Kaur et al. [36] conducted a user study that tried to find how data scientists use interpretable tools. Participants were presented with explanations from SHAP and a GAM. Later, the participants were asked some questions which tried to evaluate how they interpret the visualisations. It was found that participants misuse and disuse these explanations. Moreover, it was found out that the popularity of the tools have a positive correlation to the participants' confidence in finding the tool trustworthy. These studies show that users need to be careful while trusting these explanations especially when these are used in decision making in domains such as banking.

1.2 Problem Statement

The goal of this thesis is to design a user study that analyzes how data scientists use explainability and interpretability tools. For the purpose of this analysis, the main tools considered are SHAP and GAMs. To the best of our knowledge, the literature does not yet contain strong evidence and contributions on this. Possible reasons could be linked to both awareness and expertise with these tools, or also the specific study design. Moreover, these studies

could become too cumbersome for the participants since looking at visualisations could be tiring and understanding the math behind the working of these tools could be complicated. At the same time, if users do not understand the tool, they might not trust or would not be willing to use the tool. Thus, finding the right balance between insightful design and clarity for users is important.

1.3 Main Research Questions

We list out the following research questions considered in our study:

RQ1: What is the general behaviour of data scientists while using explainability and interpretability tools?
These behaviours are a result of users' interaction with the tool. We could get the answer to this research question by answering questions like - Do they show concerns while using the tools? Do they use the tools in the intended way? Do they understand the visualisation outcome of the tools? etc.

RQ2: What are the factors (if any) which affect a user's trust when using explainability and interpretability tools?
This can include understanding if there is a close match between the user expectation from the tool, the way they use the tool and read the output. What would increase their trust in the tool? What other visualizations would the users appreciate from the tool? How do they intend to use it?

RQ3: What do users expect from explainability and interpretability tools?
Different users could have their own set of expectations and views towards what they could use these tools for. Finding out if there is a match between the user's mental model about the tools and the intended way to use the tools, and answering questions like - What would increase their trust? What other visualisations would the users appreciate from the tool? etc. would help us answering this research question.

1.4 Thesis Outline

The remainder of the thesis is structured as follows. Chapter 2 provides background and overview of existing literature regarding explainability and interpretability along with examples of other user studies. Chapter 3 describes our research design in detail. In Chapter 4 and 5, we present the two phases of our study, namely Pilot interviews and Contextual inquiry. Finally, Chapter 6 discusses the main outcomes, limitations and ideas for future work followed by the conclusion.

Chapter 2

Background and Related Work

We provide an overview of the key topics related to our study by dividing this chapter into four major sections. In the first section, we talk about the work and results of the Kaur et al. [27] study, from which we draw our inspiration from. In the second section, we give a brief overview of approaches to model explainability and interpretability followed by the formal descriptions of some explainability and interpretability tools in the third section. Finally, we discuss some user studies that have also worked on a similar topic to ours in the fourth section.

2.1 Interpreting Interpretability

Kaur et al. [27] conducted a user study in 2020 which tried to find how data scientists use interpretable tools. It included two qualitative phases and one quantitative phase. The qualitative phases included a Pilot interview which identified the issues that data scientists face in their daily work. The researcher then used these issues to conduct a contextual inquiry where the data scientists were put in a realistic setting. The data scientists were shown a dataset, a trained machine learning model on this dataset, and explanations generated by one of the interpretability tools. Based on these explanations, participants were asked to answer some questions. The last phase was a large scale Survey which tried to scale up the findings. The three phases had 6, 11, and 197 participants respectively. Following were the main findings from this study:

1. Misuse and Disuse : Participants took the explanations at the face value and did not try to dig deeper into investigating any erratic behaviour in the explanations. This is termed as misuse. Participants also did not use the tools to their full potential which qualifies as disuse.

2. Social context : The study found out the participants based their answer on social context like the popularity of the tools.

3. Misleading Visualisations : It was found that participants continued the use of the tools even after not understanding the visualisations (meaning of the axes, scales etc.) clearly.

4. Prior experience is an important factor : The study found a correlation between the experience of the participants and their confidence in understanding the explanations and their confidence in deploying the underlying model.

2.2 Model explainability and interpretability

There are two common ways in which data scientists can understand model behaviour. They either use a black box model and use a post-hoc explanation tool like SHAP or LIME [31, 47] to understand how the output of the model is produced, or use an inherently interpretable model like a Generalized Additive Model (GAM) [22]. In the former approach, it is common to use model-agnostic tools which generate explanations on top of the predictions of the black box model by perturbing the input data and seeing the change in the output [12, 47]. In the latter approach, models like GAMs try to learn a function for each individual feature in the training set. Machine learning is split on the decision as to which approach should be taken because black box models such as a neural network or a large random forest, even though not inherently interpretable, have high performance. Rebeiro et al. [48] argue explaining machine learning predictions using model-agnostic approaches because these provide flexibility in the choice of models, explanations, and representations. On the other hand, Rudin [52] argues that one should use the interpretable model instead of black box models when making high stake decisions. She argues that interpretable models having low accuracy is a myth and with the right expertise, structured data, and a good representation in terms of naturally meaningful features it is possible to have an interpretable model with high accuracy.

A similar yet slightly different argument to Rudin comes from Freitas [19] who studied the interpretability of five classification techniques. He argues that model interpretability is subjective and should be approached in a more customized way to the technique being used. There is no universal knowledge representation that can make data scientists interpret their model in a better way. In his words, the knowledge representation (explanations/visualisations) used should "depend on the user's background and subjective preferences, and characteristics of the target dataset." Regarding the user's background, users generally have difficulties in correctly interpreting Bayesian networks when compared to Decision trees. In such cases, using Decision trees could be more useful since users can know more information about their model if they find it easier to interpret. Furthermore, regarding the characteristics of target dataset, for e.g. if a dataset has many features having a single value which is relevant for the classification, decision trees would still include irrelevant values to have a balanced tree representation. This could mislead user's understanding and could lead to overfitting as well [19]. In this case, using a rule based classification algorithm could be a better choice.

2.3 Explainability and Interpretability tools

We talk about the previously discussed explainability and interpretability tools in detail in this section. We give a brief introduction to the working of SHAP and GAMs and show the

explanations generated by them [1].

2.3.1 Shapley Additive Explanations

Before getting to know what SHAP is and how do the explanations from it look like, we think it is important to get at least a high-level understanding of Shapley values – on which SHAP is based on. Shapley values [50] is a method from coalitional game theory which determines the fair payoffs of the players in a coalition of players working towards achieving a certain overall gain from their cooperation in a game. In other words, assuming that the contribution of the players p in a game is not the same, what should be the distribution (payoff) of the total value (gain) achieved through this teamwork. Shapley values determine this payoff by calculating the marginal contribution of each player and averaging it over all the combinations of players.

We can equate terms like *Players*, *Gain*, and *Payoffs* from the above definition to Feature values, the difference between the prediction of one instance and the average prediction for all instances, and the individual contribution of feature values to the final prediction respectively.

Formally, the Shapley value of a feature value is its contribution to the gain, weighted and summed over all possible feature value combinations [38]:

$$\phi_j(v) = \sum_{S \subseteq \{x_1, x_2, \ldots, x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!} (v(S \cup x_j) - v(S)) \tag{2.1}$$

where $\phi_j(v)$ is the Shapley value of each feature j, x is the feature values of the instance to be explained, S is the subset of features that form a coalition, $v(S)$ is the total expected sum of payoffs the feature values in S can obtain as a payoff, and p is the total number of features.

Molnar [38] gives a good example to understand what Shapley values are in an intuitive way: "The feature values enter a room in random order. All feature values in the room participate in the game (= contribute to the prediction). The Shapley value of a feature value is the average change in the prediction that the coalition already in the room receives when the feature value joins them."

SHAP by Lundberg and Lee [31] uses Shapley values to generate post-hoc explanations for a given ML model. One good feature about SHAP is that it is model-agnostic, meaning that the ability to generate explanations by SHAP is independent of the model.

SHAP helps us to see Global explanations (what the model learned overall from the training data; for e.g feature importance), as well as local explanations (how an individual prediction was made). Figure 2.1 shows how the ranked overall importance of each feature, as considered by SHAP, can be visualized. SHAP also shows the global trend per feature using a dependency plot (See Figure 2.2). The x-axis shows the feature value and the y-axis shows the corresponding SHAP value. These features are also plotted against some

[1] We used Home Equity Line of Credit dataset as an example to generate the explanations. Please refer to Appendix A for more details.

other feature and one can observe the pattern of this other feature with the change in colour. Moreover, it is also possible to see how an individual prediction was made by looking at the local explanation plot (Figure 2.3). This plot starts with a base SHAP value – the output value assigned to an instance if all feature values were zero. Gradually, contribution of each feature to the prediction for an instance is added. This pushes the instance to have a higher SHAP value than the base value, or lower than it. As a result we see the final plot as the contribution of each feature value in either increasing or decreasing the base value.

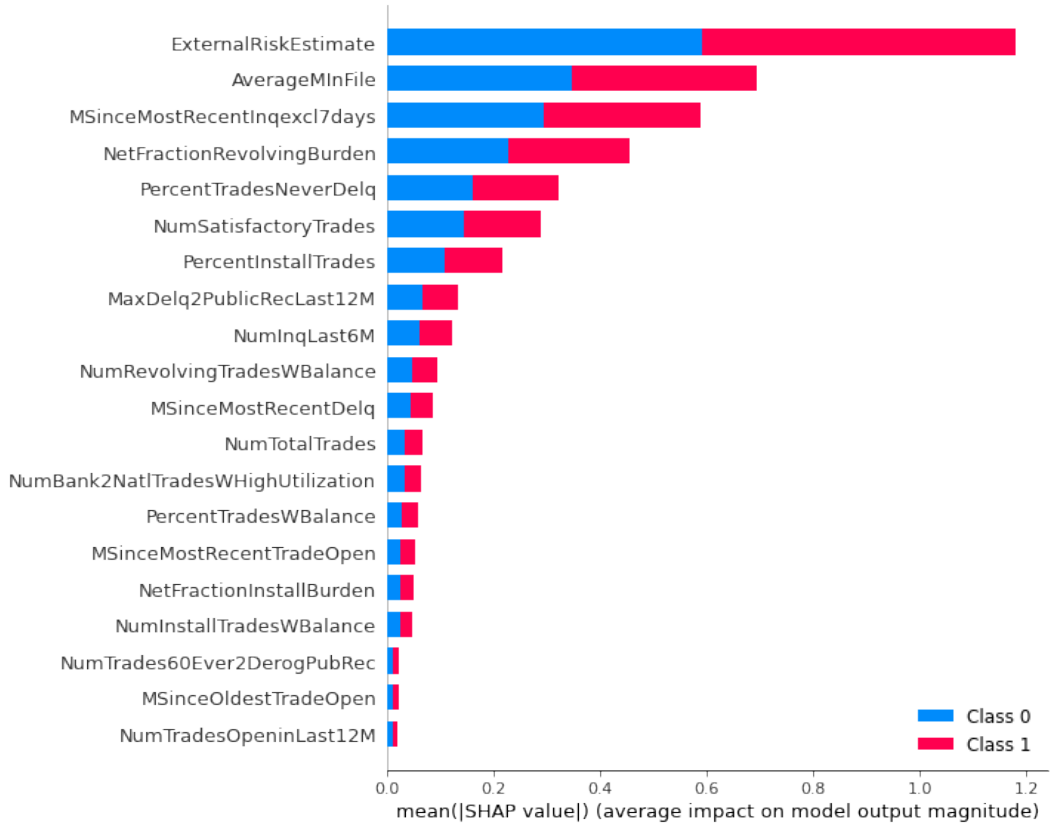

Figure 2.1: This plot is an example of feature contribution visualization. Feature contribution in SHAP is determined by calculating the mean of absolute SHAP values for each of the input features. We see that ExternalRiskEstimate has the highest mean absolute SHAP value of all the features meaning that it had the highest impact on the classification.

2.3.2 Generalized Additive Models

Generalized Additive Models (GAMs) do a good job in capturing the non-linear relationship between the predictor variables and the target variables while being transparent about it. The idea behind them is similar to linear regression, but unlike linear regression, where the relationship between the predictor variable and the target variable is a linear combination of weight sums, GAMs assume that the outcome can be modelled by a sum of arbitrary

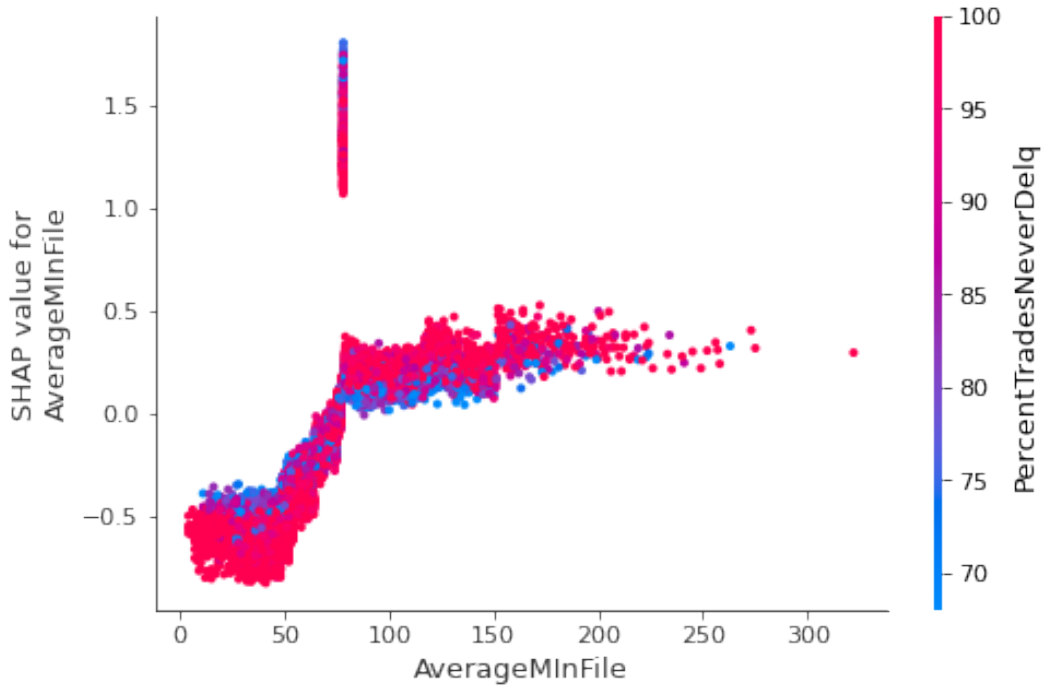

Figure 2.2: This plot highlights the trend in a feature's importance. Each feature is plotted against another, making the plot dense enough to highlight patterns. We see how the feature value of AveragMInFile impacts the SHAP value. We can also see the interaction of another feature (PercentTradesNeverDelq) with the AveragMInFile and the SHAP value. Higher SHAP values indicate that the particular value of AverageMInFile had a higher contribution to the classification.

Figure 2.3: Local explanation for an instance with predicted SHAP value of -0.26. We can see that the base value is -0.0798 and the the feature AveragMInFile with value 78 has the highest contribution in pushing it to the postive side, while ExternalRiskEstimate with the value 59 has the highest contribution of pushing it towards the negative side.

functions of each feature [38]. Each component in GAM is a function of a single input feature. Therefore, they are referred to as glassbox models because they are inherently interpretable. Mathematically, given a probability $p(x)$ =Pr$(y = 1|x)$,

$$logit\, p(x) = \alpha + f_1(x_1) + f_2(x_2) + ... + f_p(x_p) \tag{2.2}$$

where α is the model intercept (similar to the intercept in linear regression), and $logit\, p(x) = log\frac{p(x)}{(1-p(x))}$. $logit$ function converts the linear combination of covariate values into the scale of probability [34]. Here we model the probability of the classification into one of the classes as a function of the covariates. The effect of each feature $x_1, x_2, ..., x_p$ is captured by functions $f_1, f_2 ... f_p$ respectively which are estimated from the data.

GAMs can be implemented through Explainable Boosting Machines (EBMs) [43]. A publicly available implementation of this is offered by InterpretML [2]. Similar to SHAP, we can also see the global and local explanations in GAMs. Figure 2.4 shows the feature importance of the most important features according to the EBM classifier. The feature importance is ranked from highest to lowest based on the absolute score calculated by averaging the absolute value of local feature importance over all the datapoints. We can also plot the relationship between the individual feature and the predictions as shown in Figure 2.4. For the local explanations, InterpretML's EBM shows the explanation of an individual observation and the effect of each feature values on the final classification (Fig. 2.5 and Fig. 2.6)

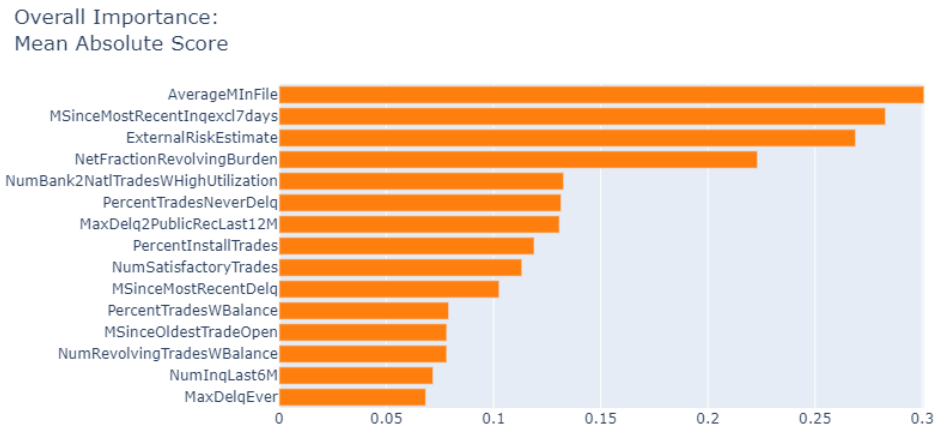

Figure 2.4: Global explanations generated by InterpretML's interpretation of GAM. This plot is similar to the Global explanation plot of SHAP seen in Fig. 2.1. Based on this plot we can say that AverageMInFile has the highest impact on the classification.

[2] https://github.com/interpretml/interpret

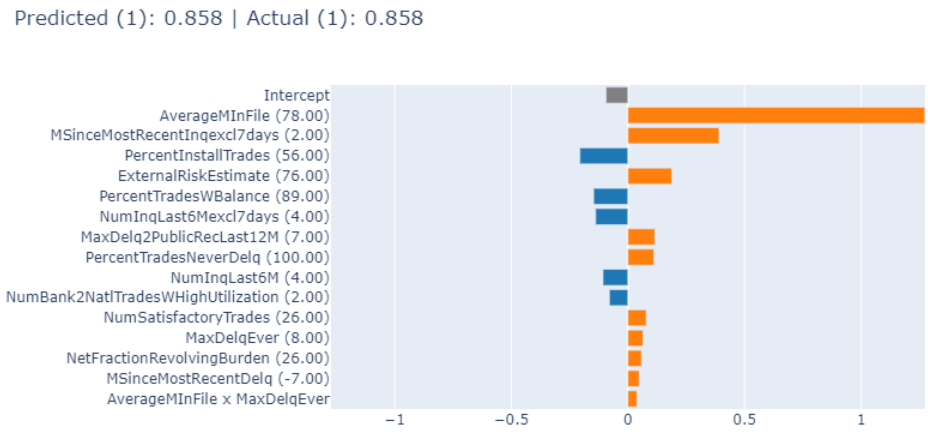

Figure 2.5: Local explanation for an observations. We can see that the actual classification for this observation was 1 (RiskPerformance = Good) and our model's prediction was also 1. The value 0.858 is the probability of this instance belonging to class 1. The explanation also shows the logit score for each feature which could be understood as the feature importance.

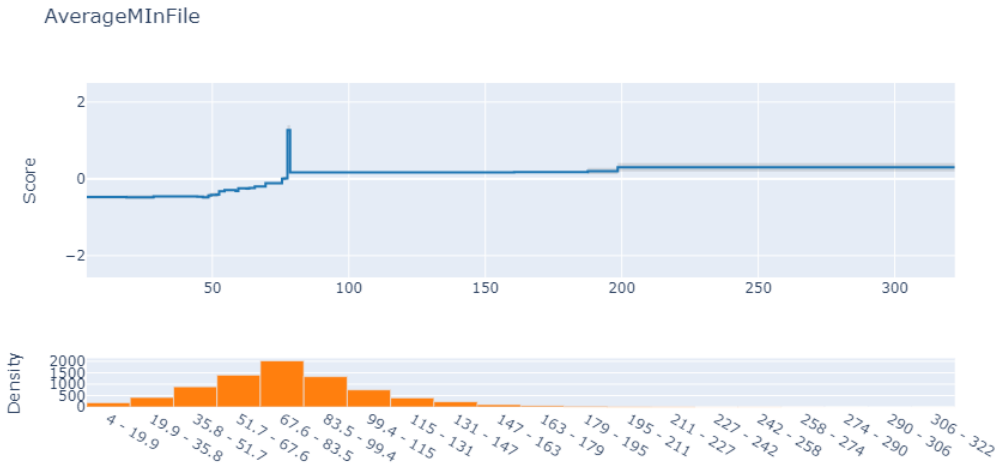

Figure 2.6: The effect of a particular feature (AverageMInFile here) on the score. We can also see the distribution of the values of the selected feature at the bottom.

2.4 User studies

Hong et al. [24] studied how ML practitioners perceive interpretability in their workflow. The study revealed that model interpretability is an interactive process involving multiple stakeholders and taking place throughout the model lifecycle. It is also context-dependent. Hohman et al. [23] developed an interactive visualisation interface called GAMUT, which they later used as a design probe for machine learning interpretability by conducting a user study. GAMUT was designed to explore how interactive interfaces could support model interpretation. It was found that there is a high demand for better explanatory interfaces, and data scientists use global and local explanations to answer interpretability questions, and data scientists also prefer interactive explanations. Both these studies show how data scientists work with interpretability and what they prefer.

Chapter 3

Research Method

The overall goal of this study is to understand how explainability and interpretability tools can support the work of data scientists and what is the current approach of data scientists to this aspect. The thesis considers an experimental design similar to the study by Kaur et al. [27], and defines different interview phases to deal with qualitative analysis. Different phases are meant to mimic a realistic setting and they should not be too cumbersome for the participants, since they might have different backgrounds and expertise.

The study is designed in two phases, as follows:

- Phase 1: Pilot Interview, with the goal of conducting a semi-structured interview asking questions to the participants about their day-to-day work in their role as data scientists.

- Phase 2: Contextual Inquiry, with the goal of creating a realistic setting presenting to participants a Jupyter Notebook [1] built on a specific dataset, a trained ML model and explanations generated by specific tools. A questionnaire then closes the session and creates the basis for the quantitative analysis.

It is relevant to conduct a study similar to Kaur et al. [27] in the banking sector, as explainability and interpretability are critical in several applications. Machine learning in the banking sector is commonly used for fraud detection, credit risk modelling, customer analytics, and marketing. All these problems require making decisions – Is this transaction fraud? Should we accept a credit application? What would promote customer retention? etc. Whenever we come across scenarios where we have to make real-life decisions based on the output from a model, it is best to use interpretable models or explainability tools than black box models. Furthermore, careful consideration should be given to check whether the model is using protected features while making a decision. Understanding how a given model is making decisions and being able to explain its outputs could help in making sure that our algorithms also optimize unquantifiable aspects like fairness, privacy, robustness, and trust (see Section 1.1). Moreover, being able to explain model decisions might also be necessary for legal considerations. European Union's General Data Protection Regulation

[1] https://jupyter.org/

under Article 13 directs institutions to provide "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" [61] where personal data is collected from the data subject. This is commonly referred to as "Right to explanation" [55]. Thus, it is important to also understand how data scientists in the banking sectors work with explainability and interpretability tools.

While largely inspired by Kaur et al. [27], our study departs from a dedicated different dataset (as discussed Section 3.3.1). We used a credit-risk dataset that is more contextual to the domain we are targeting. Moreover, our study focused on the qualitative aspect. With fewer participants, we were able to give them the freedom and time to answer our questions in depth, leaving more space for us to go through a thorough analysis of the results.

3.1 Participants

We reached out to 148 participants via the internal mailing list. We sent one-to-one "Call for interest" emails which gave an introduction about the research, researcher, and what does it require them to do. A survey link was also attached to the mail that asked the participants about their role, the duration in this role, to what extent is machine learning a part of their daily job, and their familiarity with SHAP and GAMs. Our idea was to reach out individually to the people who filled this survey and ask for their willingness to participate in one of the phases that we thought they qualified for. It should be noted that a participant was eligible to participate in only one of the two phases i.e if a participant was already interviewed in the first phase then he/she could not participate in the second phase and vice versa. We set the following requirements to recruit the participants:

- **Phase 1: Pilot Interviews**

    1. Participant is familiar with building, evaluating, and deploying machine learning models.
    2. Participant has worked with explainability and interpretability tools. It's a plus if they have worked with SHAP or GAMs.

- **Phase 2: Contextual Inquiry**

    1. The participant has a good knowledge of machine learning techniques, toolsets and algorithms.
    2. It's a plus if they have worked with explainability and interpretability tools.

We recruited a total of 14 participants based on the responses we got. We allocated 4 and 10 participants for Phase 1 and Phase 2 respectively based on how they fit into the above mentioned requirements. Please note that we would use the notation *P1,P2...P10* whenever we mention our participants in this thesis to keep them anonymous.

3.2 Pilot Interviews

The goal of conducting these interviews was to see if the themes identified in the Kaur et al. study [27] could also be found in the domain of banking and if some other themes could be identified. The themes identified were based on the issues that data scientists faced in their regular work. Identifying these themes served two purpose: 1. It helped in understanding the issues related to data, models, debugging the model, and even some problems related to the tools that these data scientists were familiar with, and 2. It helped in setting up a realistic context for the participants in the second phase (see Section 3.3).

We used the ACM SIGSOFT Empirical standards for designing the pilot interviews [46]. We made sure that we conduct the study and design the questions in such a way that they at least had the essential attributes suggested by ACM SIGSOFT. Participants were invited for a one-to-one interview where they were asked semi-structured open-ended questions, and we recorded the interviews so that we could do our qualitative analysis over the interview transcriptions. We included high-level questions like - "What does your team do?", "How does your data science pipeline look like?", "Have you ever used SHAP/GAMs or any other explainability and interpretability tools before?", "What is the purpose of using the explainability or interpretability tool in your work?" etc. and we also asked some sub-questions for each of the above questions. For e.g. if the participant answered that they have been using SHAP then we asked more questions like - "What challenges do you face while using SHAP?", "When do you use SHAP in your pipeline?" etc. to dig deeper. The interviews took place online on Microsoft Teams[2] and only audio of the interview was retained. Each interview lasted 40 minutes on average.

3.3 Contextual Inquiry

The goal of this phase is to see how data scientists use the explainability and interpretability tools to identify the issues (if they could) that were detected in the first phase. We wanted to put the data scientists in a realistic setting where they could explore the dataset and look at the ML model along with its performance (accuracy, confusion matrix etc.).

We created a synthetic dataset that contained the issues identified from the first phase. We used this dataset to train a machine learning model. Then, we displayed the explanations generated by SHAP or GAMs to the participants. This was followed by a questionnaire. The questions in this questionnaire were designed in such a way that it encouraged the participants to look at the explanations in order to answer the questions and in turn, identify the issues that we purposefully added in the dataset. These questions were the questions one would answer using explainability and interpretability tools in a real scenario. Putting our participants in such a realistic setting helped us identify common issues that our participants faced while using explainability and interpretability tools which ultimately helped us in finding out the answers to our RQs.

[2]https://www.microsoft.com/en-in/microsoft-teams/group-chat-software

3.3.1 Dataset

The Kaur et al. study [27] used The Adult Income Dataset [2] which had attributes of people such as Age, Marital-status, Occupation, Race, Education etc. as predictor variables and their income as the target variable. This dataset is a classification dataset that has the prediction task of predicting whether a person makes above USD 50,000 a year. However, we chose not to use this dataset for the following reasons:

1. The dataset is very old and popular. As of the date of writing this thesis, it has 30k downloads [26] on Kaggle.com[3], the largest community of data scientists and machine learning practitioners.

2. This dataset is also used to give example explanations for both SHAP and GAMs on their official Github [57] [35].

Both these reasons increase the chance that some of our participants could have been familiar with the dataset. This familiarity would have worked against our research design because in the second phase we change some features of the dataset to check if the participants can identify these with the explanations shown in the Jupyter notebook. If a participant was too familiar with the original dataset (before our changes to it), then instead of using the explanations to identify what is wrong with the model or the dataset, he/she would merely identify the difference between our dataset and the original. This would lead to biased answers to our questions.

To mitigate this, we chose the Home Equity Line of Credit (HELOC) dataset. According to the FICO Community (where the dataset is from), "A HELOC is a line of credit typically offered by a bank as a percentage of home equity (the difference between the current market value of a home and its purchase price)" [1]. The predictor variables in this dataset are the features of applicants in their credit reports. Examples of these predictor variables are MaxDelqEver (Maximum Delinquency ever), AverageMInFile (Average months in File), NumTotalTrades (total number of trades) etc. (see Appendix A for the full list of features). The target variable is "RiskPerformance" which has the values "Good" or "Bad". "Good" indicates that an applicant will repay their HELOC account within 2 years. If this is not the case then that instance is classified as "Bad".

3.3.2 Contextual Inquiry Protocol

Every participant was given an Informed Consent Form and a document with information about the dataset and a short tutorial about the tool. Based on the tutorial and their initial knowledge about the tool, we asked them to fill a Trust survey by Jian et al. [25] before they could open the notebook. After this, they got access to the Jupyter notebook on our computer so that they could look at the dataset, ML model, and explanations. Their main task was to answer the questions related to the explanations. We had seven questions that were related to the themes identified in the Pilot interviews to see if they could identify these issues with the explanations. For example, we asked questions like - "How does

[3]https://www.kaggle.com/uciml/adult-census-income

the feature AverageMInFile affect output RiskPerformance?" based on theme 1 (see Figure 5.2) to see if they could identify or explain the sudden spike in that feature. In each of these questions, we asked them to rate their confidence in how correct and how reasonable was the explanation to them.

Apart from this, we had three questions regarding how they read the explanations, for example- "Can you explain the predicted value of the 10th element in the test set?", one multiple choice question which asked them to list the capabilities they feel they used related to the tool, one question about their familiarity with the dataset, and one question asking their confidence in deploying the model.

After they answered all the questions, we asked some general questions about the issues that they faced (if any) while interpreting the explanations, whether they understood what was on the axes, what extra information would they like to see from these tools etc. This was followed by filling another post-interview Trust Survey which had the same questions as the pre-interview survey, and the NASA-TLX cognitive load index survey [21].

Chapter 4

Pilot Interviews

The approach considered for this phase is similar to the one used by Kaur et al. [27], where the authors identify six themes reported in Table 4.1. Data scientists in various fields of work use different data and build different models. Since our work is in the domain of banking, where data scientists use machine learning for decision-making, we wanted to see if the same themes reported in Table 4.1 also hold true in our case and if we can identify some more themes. We used the themes identified in our study as well as the themes identified by Kaur et al. study [27] for the next phase.

Theme	Description
Missing values	Many methods for dealing with missing values (e.g. Missing values coding as a unique value or imputing with the mean) can cause biases or leakage in ML models.
Changes in data	Data can change over time (e.g., new categories for an existing feature)
Duplicate data	Unclear or undefined naming conventions can lead to accidental duplication of data.
Redundant features	Including the same feature in several ways can distribute importance across all of them, making each appear to be less important.
Ad-hoc categorization	Category bins can be chosen arbitrarily when converting a continuous feature to a categorical feature.
Debugging difficulties	Identifying potential model improvements based on only a small number of data points is difficult.

Table 4.1: The six themes identified by Kaur et al. [27]

We conducted audio-recorded interviews with 4 participants (see Table 4.2) and transcribed the interview to text. After we had our transcriptions in place, we went through the audio recording of each participant and corrected any mistakes, typographical errors, and grammar in the transcriptions. Braun and Clarke [10] suggests different phases to do a thematic analysis on your data in a systematic way. We found this very insightful and hence we chose to do our analysis using the phases suggested by them. These phases are

Participant	Experience with ML (in months)	Extent to which ML used on a daily (on a scale of 1-7)	Hours spent using SHAP	Hours spent using GAMs
P1	40	5	10-20	<10
P2	60	7	<10	<10
P3	40	5	20-50	<10
P4	36	5	10-20	<10
Average	44	5.5	10-20	<10

Table 4.2: Particpants' familiarity with SHAP and GAMs. Participants were more familiar with SHAP than GAMs

sequential i.e, they use the results from the previous phase and build upon them. We would now go through each phase of the thematic analysis and describe how we identified the final themes.

- Phase 1: Familiarising yourself with the data
  We went through the transcriptions while listening to the audio recordings. The advantages of going through the transcripts while listening to the recording were twofold: 1. We understood the features like tone, voice modulation, and emphasis on things much better by listening to the audio when compared to the text; 2. We were able use our understanding of these features (tone, emphasis etc.) to edit the transcriptions so that it became easier for us to understand the next time we read it. In this phase, we also put the transcriptions in a structured question-answer format.

- Phase 2: Generating initial codes
  In this phase, we did a systematic analysis of the data by coding the information in simpler terms. We did this twice for every participant so that we do not miss out on any useful information. Throughout the process, we kept in mind the research questions and the purpose of doing pilot interviews in the first place. We added comments if we wanted to paraphrase some lines, and highlighted the text in the transcription itself if they seemed relevant. For example, when we asked participant P4 about why did they not like the tool, they replied: "Because it was not that obvious what the meaning was of the graph." which we coded as "Did not understand the visualizations properly."

- Phase 3: Searching for themes
  The aim of this phase is to review the coded data and identify areas of similarity and overlap between the codes [10]. We listed the codes and highlighted texts from all the participants' transcriptions and listed them next to each other in a 4-column table (one column for each participant). This facilitated us in comparing the codes and finding similarities between them easily since we could see them side by side in just one document. We used colour coding to identify similar codes. For example, codes like - "Worked with textual data from feedback" and "Used Natural Language Processing"

identified from two different participants were coded with the same colour because they had something to do with working with textual data.

- Phase 4: Reviewing potential themes
  This phase involved reading over the transcriptions again to see if the identified themes meaningfully captured the aspect of the entire dataset. We tried identifying the most common and relevant themes that came up from our coding.

- Phase 5: Defining and naming themes
  After colour coding the common themes, we tried to club the similar themes under a common name. We were careful that the identified themes do not overlap, they have a singular focus, and they come from at least two different participants.

After going through these phases, we identified the following 9 themes:

1. Use of personal data

Participants mentioned their use of data about people like demographics, personal data collected through marketing campaigns and feedback forms. This theme was common in 3 participants. When asked what type of data they dealt with, P1 replied -

> "So if we are working for KYC, perhaps we deal with models that use transactional data. If we are validating models for customer interactions, for example, we deal with data that it's more relating to marketing, so more personal data or data that is related with marketing campaigns"

On the same question P2 replied -

> "I've also worked with the data that came out of the warehouse and this is then usually the normal bank data where you have for mortgage contracts like information about the length of the contract, interest rate and other demographics about the individual"

It is evident that our participants dealt with sensitive data in their work. Doshi-Velez and Kim [17] state that an important use of interpretability is to check if any protected features are used by the model to make predictions. However, it is still important to see how people perceive the explanations by the explainability and interpretability tools. Kaur et al. [27] found out that data scientists in their study found the explanations reasonable even though a protected feature like "Age" was used to make decisions by the model. Therefore, it becomes important to see if such things are also common within other sensitive domains.

2. Use of textual data

All the participants dealt with some kind of textual data. P1 mentioned -

> 'So we're starting to work also quite a lot with text data...so data can come from review data"

whereas P3 mentioned -

> "(on data in forms and surveys) And the open text is the place where basically data scientist comes in"

3. Look for outliers/anomalies in the data

Three participants mentioned that they looked for things that do not make sense, and anomalies in the data. When asked if they performed any checks at different stages in the pipeline, P2 replied -

> "We look for outliers and missing data."

and when asked about what issues they use the explainability and interpretability tools for, P1 replied -

> "...so when you're looking at transactions, you want to investigate anomalies and whenever you spot an anomaly, you send it to the business"

while P4 replied -

> "If you find something weird is going on, then I find it helpful or see where the error is."

Kumar defines anomaly detection as - "Anomaly detection attempts to quantify the usual or acceptable behaviour and flags other irregular behaviours as potentially intrusive" [28]. In banking, these anomalies could be point anomalies, contextual anomalies, or collective anomalies [3]. Finding anomalies could help in preventing fraud in domains like credit card transactions, mortgage applications, insurance claims etc. Therefore, it was interesting to see how data scientists inferred the explanations by the explainability and interpretability tools that showed anomalies in the dataset.

4. Use explainability and interpretability tools for feature importance

All participants mentioned that they have used explainability and interpretability tools to know the feature importance. P4 mentioned -

> "I think they're good for if you work together with the business"...[later] "For example, I use [feature importance] when explaining the results to the business which I had to do in previous projects so that they understand - Oh if this feature is important, that makes sense"

Data scientists could see the features that contribute the most to a particular decision by the tool and they could try to check it with the stakeholders to see if such a behaviour is normal. In fraud detection, to determine if a transaction is fraudulent, features like the number of transactions, location, and amount transferred would be more relevant (thus, they should have a higher feature importance) than the features like balance of the recipient, and weekday [58].

5. Difficulty in understanding the output from explainability and interpretability tools

All the participants mentioned that they had difficulties while understanding the output from the tools. This included figuring out scales, understanding what the values on the axes represent etc. P1 said -

> "The output is not as directly interpretable as you will have with other statistics techniques"

When asked why the experience with SHAP was not good to P4, they replied -

> "Because it was not that obvious what the meaning was of the graph."

This issue is very important because then it could lead the participants into incorrectly interpreting the visualisations. As a consequence, they could trust an unfair model and deploy it. We tested this in the second phase to see if the participants correctly understood the explanations and how did it affect their confidence in deploying the model.

6. Missing data is a problem

Two participants mentioned that they faced issues like leakages in models with missing data. P2 also suggested a way to handle missing data -

> "So one way is to impute and I guess they suggested to just take averages [and impute]."

One should be careful while handling missing data. Missing data imputation if there exists a highly correlated feature can lead to the imputation being useless or even harmful [7]. It could also lead to oversimplifying the model if the amount of missing data is increased. We tested this in the second phase by imputing missing data with the average value for a feature to see how much wary the participants were about the effect of this imputation on the explanations.

7. Data and model changes over time

Participants were also quite critical about the changes of features, data, or models over time. This includes consistency of predictions and stability of features over time. When asked about any other issues they face in their work, P1 replied -

> "So how it [variables with higher impact] changes? What is the stability per individual observation? How Stable are the top five most important features in terms of SHAP, LIME or whatever technique"

P2 mentioned -

> "[on population stability checks] So if the data that we received, if it's stable over the time period that we have"

8. Understanding effects of variables on the outcome

Participants were concerned with understanding the cause-effect relationship between the predictors and the target variable. When asked what kind of questions did they try to answer using SHAP, P2 replied -

> "As an economist, I'm interested in sort of causality...what is driving what"

On the question of when would you use GAMs, P1 replied -

> "If I have some project where I really want to understand or directly interpret the effect of the variables that have on the target, I could consider GAMs."

Chapter 5

Contextual Inquiry

We recruited a total of 10 participants for this phase (Table 5.1). Five out of ten randomly selected participants were given the Jupyter Notebook containing SHAP and the other half were given the GAM notebook. These participants were called for an online interview via Microsoft Teams. We set up our experiment in a Jupyter notebook which contained the dataset, a trained ML model, explanations, and a questionnaire. Separate notebooks were created for SHAP and GAMs. Usually, this phase would have taken place in-person where we would have presented the notebook to the participants and they would have answered the questions in front of us. Conducting the contextual inquiry in this way would have made it easier for the participants to ask any questions they had while reading the notebook. Since it was not possible to conduct this phase physically, we chose to conduct it online where the participants were given access to the Jupyter Notebook on our computer via screen share. Conducting this phase online, however, did not compromise any benefits we would have gotten if the interview was physically in person.

During the Contextual Inquiry, we asked the participants to think out loud when they were trying to answer the questions in the Jupyter Notebook and instead of typing the answers in the cells of the notebook; which - 1. Takes time and 2. Is difficult to type while accessing someone else's screen, we asked them to just say the answer. An advantage of this was that the participants gave more information about how they were reading the explanations, the doubts they had, the basis to their answers, the problems they faced while answering the questions, and what they would like to see other than the visualisations that we showed them.

Table 5.2 lists the themes resulting from the Pilot interviews that we could incorporate into our Contextual inquiry by manipulating the dataset. On average, a theme was identified by 3 participants with the minimum and maximum being 1 and 9 respectively, and each participant identified 2 themes on average. Participants' average confidence in understanding the explanations was 4.91 with a standard deviation of 0.6, and their average confidence in finding the explanations reasonable was 4.57 with a standard deviation of 0.79. Although this number is high, when asked to rate on a scale of 1-7 how much likely is it that they would deploy the model based on the given explanation, the average rating was only 2.85. Participants stated reasons like accuracy being too low, not being able to see feature correlations, and not understanding the explanations clearly. Some participants also stated their

Participant	Experience with ML (in months)	Extent to which ML used on a daily (on a scale of 1-7)	Hours spent using SHAP	Hours spent using GAMs
P5	26	5	50-100	<10
P6	120	7	<10	<10
P7	48	6	20-50	<10
P8	87	6	10-20	<10
P9	48	6	<10	<10
P10	24	6	10-20	<10
P11	26	7	<10	<10
P12	90	6	<10	<10
P13	36	5	10-20	<10
Average	56	6	11-28	<10

Table 5.1: Particpants' familiarity with SHAP and GAMs in Phase 2. Similar to Phase 1, participants were more familiar with SHAP than GAMs. We only listed the details of 9 participants instead of the actual amount of participants in Phase 2 (10) lacking the answer from one participant.

willingness to check with the business if the explanations make sense.

We did not see a significant difference in the answers of Trust survey before and after the contextual inquiry. Participants' average trust in the tool (SHAP and GAMs) was 4.33 with a standard deviation of 0.712 before the contextual inquiry and 4.18 with a standard deviation of 0.97 after.

The results of NASA-TLX questionnaire shows that the mean overall load for participating in the Contextual Inquiry was 61.63 (out of 100) with a standard deviation of 6.8. This was 62.95 (s.d = 4.47) and 59.32 (s.d = 8.18) for SHAP and GAM respectively. The highest rating was given to Performance (how successful do the participant think he/she was in accomplishing the goals of the task) which received a mean rating of 68.5 (s.d = 11.41) and the lowest rating was received by Physical Demand (how much physical activity was required) with a mean rating of 11 (s.d = 13.75).

At the end of the Contextual inquiry, we asked the participants to list the capabilities of the explanation system they just used out of a list of 11 capabilities. Figure 5.1 shows the responses we got to this question. Out of these 11 capabilities only 6 were the actual capabilities of the tools. The participants did a good job in identifying these since the top 6 capabilities in the Figure 5.1 are indeed the actual capabilities of the tools.

Before going into the results of this phase, we want to acknowledge that one limitation of this phase was time. We designed the study in a way that it doesn't take more than one hour to complete. However, in practice, tasks like learning about the dataset and the features, debugging the model, and analysing the explanations are long processes that take days, if not weeks, to complete. Therefore, we did not expect the participants to dig too deep into finding the reasons behind the issues while reading the explanations. We considered it a good answer if they even mentioned what they would like to investigate or what they

Theme	Description	Incorporation into Contextual Inquiry	Num.
Missing values	Many methods for dealing with missing values (e.g., Missing values coding as a unique value or imputing with the mean) can cause biases or leakage in ML models.	Replaced the value for "AverageMInFile" feature with 78 (the mean) for 10% of the data points with RiskPerformance "Good", causing the predictions to spike at 78. Asked about the relationship between "AverageMInFile" and "RiskPerformance"	9/10
Duplicate data	Unclear or undefined naming conventions can lead to accidental duplication of data.	Modified "MaxDelq2PublicRecLast12M" to have duplicate values: "unavailable delinquency", "unknown delinquency", "delinquency not known". Asked about the relationship between "MaxDelq2PublicRecLast12M" and "RiskPerformance".	1/10
Redundant features	Including the same feature in several ways can distribute importance across all of them, making each appear to be less important.	Included two features, "TotalTrades" and "NumTotalTrades" that represent the same information. Asked about the relationship between each of these and "RiskPerformance."	2/10
Ad-hoc categorization	Category bins can be chosen arbitrarily when converting a continuous feature to a categorical feature.	Converted "MSinceOldestTradeOpen" into a categorical feature, binning arbitrarily in -10-0, 0-100, 100-200, 200-300, 300-400, 400-500, 500-600, 600-1000. Asked about the relationship between "MSinceOldestTradeOpen" and "RiskPperformance".	1/10
Debugging difficulties	Identifying potential model improvements based on only a small number of data points are difficult.	Asked people to identify ways to improve accuracy based on local explanations for 20 misclassified data points.	4/10
Outliers in data	Outliers can be present in data that could mean and standard deviation and in turn reduce the power of statistical tests.	Identified outliers in "ExternalRiskEstimate" by IQR method. Asked the relationship between "ExternaRiskEstimate" and "RiskPerformance"	1/10
Difficulty in understanding the output from the tools.	There is a difficulty in figuring out scales, figuring out what the values on the axes represent etc.	Asked participants whether they understand what is on the X-axis and Y-axis in global explanations.	3/10

Table 5.2: The final list of themes. We incorporated each theme into our contextual inquiry as described in the third column. The last column shows the number of participants who identified the corresponding theme.

Figure 5.1: Responses by participants on what they think are the capabilities of the tool (SHAP or GAMs) they used. On y-axis, we can see the list of capabilities that we asked the participant to choose from, and on the x-axis is the count of participants who chose a particular capability.

would do differently to make them confident about their answers. For example, when asked about the relationship between AveragMInFile and RiskPerformance (which has a spike in the visualizations around the value 78, see Figure 2.2), (P6, SHAP) answered -

> "To answer this question clearly, I would just remove that section[with the spike] and estimate again so I will drop those data points and then see the behaviour because to me it seems like that erratic line might be contributing to[the feature] being the most important one".

An answer like this made us confident in believing that the participant was interpreting the tool correctly.

We analysed the results in a deductive and an inductive way. Deduction helped us in finding out if the themes identified by other studies existed in our findings as well, and induction helped us in finding out any new themes. These themes express the general behaviour (in line with our RQ1) of data scientists while using the explainability and interpretability tools (see Section 1.3). We now see the themes that were common in the answers of our participants while they were answering the questions in the Jupyter Notebook.

1. Misuse and Disuse

Parasuram and Riley [44] define Misuse as "over-reliance on automation, which can result in failures of monitoring or decision biases". In our case, we were looking for instances where participants relied too much on the visualization and took the visualizations for their face value instead of using the tool to figure out why certain issues are surfacing. Unlike Kaur et al. [27], we did not find very strong indications of participants misusing the tools. As specified earlier, we considered it an intended use of the tool even if the participants just stated that they found something suspicious and they would like to investigate further,

instead of actively trying to find the cause of something suspicious. If this was not the case then the answer qualified as a misuse. When asked how does *ExternalRiskEstimate*(which is a consolidated indicator of risk markers) affected the output *RiskPerformance*, (P7, GAM) answered -

> "It's very clear that lower values impact negatively while higher values impact positively."

and rated 7 (on a scale of 1-7) when asked about their confidence that they understood the explanations correctly and their confidence whether they felt that the explanation was reasonable. Moreover, (P8, SHAP) answered -

> "OK, so with the increasing number of total trades the probability of the RiskPerformance to be one is increasing"

and rated 6 to their confidence in understanding the explanations correctly as well as finding them reasonable on being asked how does *NumTotalTrades* affects the *RiskPerformance*. In the former case, even though the effect of outliers in *ExternalRiskEstimate* was clear in the explanation (they had a perfect score of 2 on a scale of -2 to 2), the participant didn't find it suspicious and neither tried to investigate it. Similarly, in the latter case, the participant didn't try to investigate if *NumTotalTrades* and *TotalTrades* were the same features even after getting confused when the same question was asked for *TotalTrades* earlier in the inquiry.

When asked about the effect of *ExternalRiskEstimate* on the target variable, (P9, SHAP) became sceptical about the use of the feature -

> "My understanding of this feature is it's some other model that tries to predict the Risk factor of the client, so obviously if that other model is accurate, it will tell you if the client is at risk of being a bad client in the next 24 months, so yeah it is using this feature as your main predictor makes sense"

These qualify as misuse because not only did the participants fail to identify any issue with the visualization but also rated their confidence as high. However, we believe that this is not a strong indication of misuse because we didn't find similar instances in other participants' responses.

We found a significant amount of evidence that indicates disuse. Disuse is termed as an "underutilization of automation" [44]. When participants disuse a tool, they do identify issues in the model or the dataset but they either ignore it or don't try to investigate further, thus not using the tools to their full potential. When asked about the effect of *AverageMInfile* on the output *RiskPerformance*, (P1,GAM) said -

> "Yeah, this makes sense to me and then we know that this is an important feature. So I understand this one. The jump is interesting, though. Why does that jump happen at that peak? That's... that's interesting."

and gave a rating of 7 to both their confidence in understanding the explanation and their confidence in finding the explanation reasonable. (P3, GAM), while talking about the the effect of *NumTotalTrades* on the output *RiskPerformance*, got confused if this feature was the same as *TotalTrades* (which it was) -

> "Well, this one is TotalTrades, right? Now we are with the NumTotalTrades, yeah? This is a different one, right?"

But instead of investigating further, the participant went ahead and gave the explanation of the effect of this feature on the target variable which was a different explanation than *TotalTrades'* effect on the target variable. Similar themes were found in three other participants' responses to the questions. (P9, SHAP) had doubts that *TotalTrades* and *NumTotalTrades* are "invariable", but didn't investigate it further, (P7, GAM) had doubts if outliers were removed from the dataset but didn't point out the possibility of the existence of outliers as seen from the explanations, (P8, SHAP) also acknowledged the spike in *AverageMInFile* around the value 78 but didn't express their willingness in investigating why the spike was there.

2. Reliability on social context

One of the findings of the Kaur et al. [27] study was that some participants in their study based their answers on the popularity of the tools. Their trust was increased if they knew the tool was popular or if it showed them something which they have not seen before. However, none of the participants in our study based their answer on either the popularity of the tool or their familiarity with the tool. This theme does not hold true in our case.

3. Misleading Visualizations

There were clear shreds of evidence of participants' confusion in understanding the visualizations. (P1, GAM) had a lot of confusion in understanding the scales and scores-

> "What does the score mean? I see that the values here [global explanation per feature] are between 0 and 0.3. And then we said scores. So these are not scores then? These are mean absolute scores. So what does that mean? That is it normalized?"

(P5, GAM) had trouble in trying to find the exact values for the Y-axis-

> "So one of the things that's a bit annoying I guess is that the Y-axis, are - 2, 0, and 2, but there's nothing in between. I guess this is 0.8 and maximum and then minus 0.1 at minimum."

Instead of just seeing the Mean absolute score, some participants also wanted to see how the features impact negatively the score. When asked which features they think are the most important ones in affecting the target variable, (P7, GAM) replied -

> "I would say that the first four of course, but then...but since it's just the positive impact, I am kind of struggling to understand if it's correct or not"

(P4, SHAP), (P9, SHAP) also stated their interest in seeing the negative impact of features and confusion in understanding the scales respectively.

We also found that the participants used their **Intuitions** and had **Perceived Suspicions** when looking at the explanations. Both Intuition and Perceived Suspicions were identified as "Factors that affect willingness to deploy" in the Kaur et al. study [27]. Instead of categorizing these themes under Factors affecting willingness to deploy, we address them as their own themes because we believe that all the themes identified from the answers we got contribute in some or the other way to the participant's willingness to deploy. We now present the themes that we identified that were not in the Kaur et al. study [27].

4. Intuitive reasoning

Some participants gave answers based on their prior experience with working with ML. Sometimes these answers were not based on the explanations themselves but based on the data or the techniques that they would like to use to analyse the data. This include correlation analysis between the variables. (P4, SHAP) on coming across the feature *TotalTrades* became suspicious of any leakages in the dataset -

> "You already know that basically this client, when it is going to apply for a lot more loans...you only apply for more loans if you didn't default in the past. So if that was the case, you would have leakage. I would expect that if there was leakage, it would've been more important. Given that this is relatively low SHAP value, I don't think there's leakage going on."

(P9, SHAP) had doubts if *AveragMInFile* could be a proxy for Age of the customer -

> "It also is the fact that if the client has been there for 20 years, then it's probably not 20 years old himself and that is a bit older, so it's also proxy just for the age of the client. And the risk of older people is less than of younger people just simply by the fact that they have more money in hand."

Two more cases were found where the participants used their intuition in finding if an explanation was reasonable or not. In all such cases, the participants also tried to solve their doubts using explanations. If this would have not been the case then it would have been a case of Disuse.

5. Perceived Suspicions

A lot of participants expressed suspicions over the dataset and the model. These suspicions were associated with knowing the meaning of the values of the features, model performance, explanations, and feature meaning.

(P3, GAM) had confusion with the values of the feature *MaxDelq/PublicRecLast12M* (Max delinquency/Public Records Last 12 Months) which had values like - 30 days delinquent, 60 days delinquent, unknown delinquency etc.

> "If you have 60 days delinquent, is 30 days delinquent not part of 60 days because 30 days is within 60 days range. So that would be my first question

> with this category. Or is it that the 30 days means from zero to 30 days, the 60 days...that mean 30 to 60 days?"

(P6, GAM) was clear in what they wanted to see to trust the model, when asked what do they think is the most important feature that affects the RiskPerformance, the participant replied -

> "So for me to really trust this, for each feature I would prefer to see the density of the important scores taken from all the observations and then to see how does this really look like. Does it look like there's some kind of normal distribution order? Some kind of skews? But at least having some kind of variance. This will be absolutely reliable."

Similar suspicions were found by (P4, SHAP) who was sceptical if the model is overfitting.

6. Dependence on stakeholders

A lot of participants expressed the need to consult with stakeholders on topics like feature importance, local explanations, and validity of features to see if they were in line with the business point of view. These stakeholders are the people who formulate the problem statement and know the in-depth meaning of the features and their behaviour. When asked to explain the predicted value of the 10th element in the test set, (P1, GAM) answered -

> "The probability comes out as good. We can take a look at all these features and see if it makes sense from a business point of view. So I think that's the most interesting thing to do. So if I'm talking to someone, then I can take a look at what the model predict and based on these feature values, it would be good."

On answering which features do they think are the most important for prediction, (P10, SHAP) said -

> "So we are confident that this is the list that we want to consider, right? But at the same time just consult it with the business or have our own understanding about if this makes sense? "

(P4, SHAP) also expressed their willingness to consult with the stakeholders in knowing if there were some specific cases they could be interested in. In total, we found this theme in 4 participants.

7. Not trusting the explanations standalone

We found that a lot of participants wanted to see something more than just the explanations to increase the confidence in their answers. (P4, SHAP) wanted to see a plot where we can see the positive as well as the negative impact of features on the SHAP value. (P3, GAM) wanted to check the correlation between the features -

"Well, if we only look at the mean absolute score...[Í'm] a little bit hesitant because yeah, it gives them in a somewhat in isolation and we did not check the correlation. It could very well be that this feature is really important, but it's because there's another feature in it that's giving it an extra boost."

(P10, SHAP) wanted to see the monotonicity of the features -

"One thing which I always like to do before you do any sort of this explanation is to see if there are any monotonically increasing features and are we including that in the model."

Chapter 6

Discussion and Conclusion

We first answer the research questions addressed in Section 1.3 by analysing the results of our two phases followed by a discussion of main findings, threat to validity, limitations, and future work. We finally conclude our thesis by suggesting ways in which our results could be used and give value in the banking context.

6.1 Answering our research questions

The answers to our research questions lie in the analysis of the themes identified in our two phases:

RQ1: What is the general behaviour of data scientists while using interpretability tools?

From the results of our study, we found that the most prominent themes that capture the behaviour of data scientists while using interpretability tools, especially SHAP and GAMs are: 1. Disuse - not using the tool to its full potential, 2. Intuitive reasoning - use of prior knowledge and/or intuition to find if the explanations are reasonable, 3. Perceived Suspicions - scepticism or doubts about the data or the model that were not answered by tools, 4. Dependence on stakeholders - checking with the business people if the behaviour of the model is reasonable, 5. Not trusting the explanations standalone - needing more information to trust the model than just explanations, and 6. Misleading visualisations - not being able to correctly read the explanations. This list is not at all exhaustive. We believe that more research would be needed to see if such behaviour is also exhibited by data scientists working in other domains.

RQ2: What are the factors (if any) which affect a user's trust when using interpretability tools?

We found that Misleading visualisations, Perceived suspicions, and Intuitive reasoning are the most important reasons that affect user's trust when using interpretability tools. Misleading visualisations which include difficulty in understanding the meaning of the axes, doubts regarding scales in the graph, confusion over the meaning of GAM score etc. lead the participants into having low confidence in their understanding of the explanations and finding them reasonable. The presence of Perceived suspicions and Intuitive reasoning in-

dicates that there was a demand for more information apart from the one provided by the explanations that would make the participants trust the underlying model more. As a result, it leads to a low deployment score.

RQ3: What do data scientists expect from interpretability tools?

Data scientists expect more clarity on the meaning of visualisations and expect more capabilities that would help them investigate the issues in detail. Features like nearest neighbours of individual instances and feature correlations were the most frequently mentioned in the answers of the participants. Some participants also wanted to see the same visualisations with some modifications like - seeing the negative contributions of a feature to the SHAP value in the global explanation plot (Fig. 2.1), and class (0 or 1) to be colour coded in the SHAP partial dependency plot (Fig. 2.2). This indicates that users also demand some control in displaying the visualisations from these tools.

6.2 Main findings

The most common theme identified by our study was *Disuse*, which was common in the response of 5 participants. Other common themes identified by our study include - *Misleading Visualizations*, *Intuitive reasoning*, *Dependence on Stakeholders*, and *Not trusting the explanations standalone*, all being identified in the answers of 4 participants. Finally, the themes that were the most uncommon are *Misuse* and *Reliability on social context* which was identified in the answers of 3 and 0 participants respectively.

We now discuss the themes starting from the least frequently identified to the most frequently identified theme. We can rule out *Reliability on social context* since it was not identified in our study. An important skill of a data scientist is to have intense curiosity [14] and being critical about looking at data and debugging the issues related to either the data itself or the model. Therefore, we consider it a positive thing since in an ideal world there should not be any bias in data scientists related to the popularity of the technology they are using. It is also good to see that *Misuse* is one of the least identified themes. This shows that most of the participants do identify some issues with the help of the interpretability tools.

Moving on to the most common themes, it is a combination of positive and negative themes. We deem that in the case of *Misleading visualisations* and *Intuitive reasoning* we might have some risks. If data scientists have confusion on what the axis shows or what the score represents, then it would have consequences on how they would perceive the visualisations leading to inaccuracy in reading explanations. On the other hand, *Intuitive reasoning* could potentially lead to *Disuse*. If a model is discriminating against an underrepresented feature and the person looking at this trend in the explanations thinks that it is intuitively reasonable, then potential disuse could be the risk. A solution to prevent this would be to check with other stakeholders if such correlations make sense in reality. Two out of the four participants who used *Intuitive reasoning* to answer the questions also expressed their willingness to check if some irregularities make sense with the business/stakeholders.

We consider it a positive thing that most commonly, participants had *Perceived Suspicions* about the data and the model, *Dependence on stakeholders* to check if things make

sense to them as well, and they did *Not trust the explanations standalone* and wanted to check for other things as well to make sure that the explanations are consistent.

The most common theme was *Disuse*. This indicates that it is an issue that is persistent when data scientists use explainability and interpretability tools, since it was also identified by Kaur et al. [27]. We consider it as risky because it can also represent a behaviour of under-utilizing the tools and this is an issue that is strictly caused by a mismatch between the participants' mental models and the conceptual model of the tools.

6.3 Limitations

There are some limitations to the results that we would like to talk about in this section. They are the following:

1. Lack of quantitative data

   We had a small sample of participants so that we can focus on qualitative analysis. Therefore, it is uncertain if our results would stay valid if the findings are scaled up. Moreover, findings like participants' accuracy of answers, which tool is easier to read, and the role of mental models were difficult to find.

2. Time constraint

   Most of our interviews took approximately one hour. In a real scenario, going through the phases of a data science pipeline takes days. We feel like this was a limitation because in some instances, participants did not try to understand the meaning of the features in the dataset and some participants did not read the tutorial and the description of the dataset before the Contextual inquiry. This would have been different in real life because data scientists working on a project would have spent hours knowing about the dataset, building and debugging the models, and looking at the explanations. Therefore, we acknowledge that we are uncertain as to what extent were we successful in capturing the nuances of how data scientists use these tools.

3. Use of a different dataset

   The meaning of features in the dataset used in the original study was much easier to know since the features were self-explanatory. Features like Age, Marital Status, and Education are much easier to interpret than features like MaxDelqEver (Maximum Delinquency Ever), NumTotalTrades (Number of total trades), and NetFractionRevolvingBurden (revolving balance divided by credit limit). For features like the latter, it is important to know the meaning of terms like "Trades" and "Delinquency". Intuitively, we believe that the use of easily understandable features could have resulted in more cases of Misuse and Disuse because a participant could have had their own perceptions about what kind of effect of a feature should have on the target variable.

6.4 Threats to Validity

Lewis[30] defines five threats to validity that can happen in Qualitative research. We would now discuss how these threats to validity could be present in our research and how we try to mitigate them.

1. Descriptive Validity

   This threat concerns the data gathered through the qualitative methods in a research. The data gathered should be accurate and complete. In our case, this data is the interviews and transcriptions. It is important to record all the things that a participant says instead of short comments and keywords taken as notes by the researchers. To mitigate this, we recorded both, the Pilot interviews and the Contextual Inquiry interviews. Moreover, we encouraged participants to say their answers instead of typing them in the Contextual inquiry phase and we recorded the whole session.

2. Interpretation Validity

   It is important for the researcher to understand the perspective of the answers of the participants and the researcher should limit herself/himself from leading the answers of the participants in a certain way. To avoid compromising this validity, Lewis [30] recommends having open-ended questions and advises not to have questions that direct the participant to answer in a certain way which would have not been their natural response. Therefore, the questions in our research were mostly open-ended except for the ones where we wanted the participants to rate their confidence in understanding the explanations. We also gave the participants access to the Jupyter notebook on our screen so that they can scroll, click, and write in the notebook and read the questions themselves.

3. Theory validity

   The researcher should prevent herself/himself from trying to fit the data to a certain theory that they have before they begin the analysis. To this front, we try to present all the data instead of presenting only the data that seem to support the hypothesis. This is evident in the results of our Pilot Study where we presented the themes that we did not even use later in the contextual inquiry.

4. Researcher Bias

   Researchers may have their own biases. It could be related to the participant's background, the support of a theory over others, or the analysis of data. We have tried our best to not have any bias like this in our study, however, we do have to acknowledge that there could be a bias in the analysis of data since all the analysis was done only by one researcher and we might have omitted some important information. This was a constraint on the study as the data was confidential and could not be shared with anyone else apart from the researcher. However, we tried our best to look at all the information in-depth, iterating over the interviews and transcripts again and again.

For example, we listened to the interview recordings once while editing the transcripts and adding comments, and again while analysing the transcripts while trying to identify some patterns.

5. Reactivity

   This concerns the participants being conscious that they are being observed and that affecting their responses. This was indeed a threat because the contextual inquiry phase entailed us being with the participant and answering any questions they might have while trying to look at the explanations and answering the questions. This was also evident by one instance where a participant, after answering the question, said - "I give boring answers, right?". We could have sent the Jupyter notebooks to the participants and have them answer the questions in their own time. This would have completely eliminated this threat. However, we decided to go for direct interviews to have a direct discussion with the people involved. This also meant that we saved time instead of conducting the whole process via email and at the same time followed the way of thinking of the participants. To mitigate this further, we informed the participants that there are no right answers to the questions and they should feel comfortable in answering the questions in whichever way they like. We also turned off our camera if it was necessary.

6.5 Future Work

There is some work that we think can be done in the future. This includes :

- Quantitative analysis

  Just like the Survey in the Kaur et al. study [27], it would be interesting to see if the results we got could be scaled up. This could quantify the accuracy of participants at reading the visualizations and it would also be interesting to see the correlation between the themes which is difficult to conclude with confidence with a small amount of data.

- Using a textual dataset

  It would be interesting to see if the results hold true when working with a textual dataset.

- Designing tools for deliberative reasoning

  Data Scientists' disuse of the tools identified by both the studies suggests that there is still a lot to improve by the designers of the tools to encourage people in using the tools for deep analysis instead of just using the tools at the surface level.

We think that the results obtained are quite useful for people working with explainability and interpretability tools in banking. We believe that a way to reduce the risk of disuse of these tools could be studying the documentation and getting acquainted with the technical aspects and language at an earlier stage. In this way, raising awareness can be

beneficial. It was good to see that most of our participants were willing to consult with the stakeholders regarding the explanations provided by the tool and their meaning. This encourages space for getting feedback and ensures that model behaviour is consistent with any feature engineering done before. Moreover, data scientists could build their tools on top of existing explainability and interpretability tools to have the control we saw our participants wanted. Answers from our RQ2 indicate that a more comprehensive documentation on explainability and interpretability tools would be desirable to guide the user. Developers of the explainability and interpretability tools could provide a wider range of visualisations to choose from, and flexibility to the users in modifying the visualisations to their needs (changing the variables on the plot, colour etc.). This would cater to the different needs of data scientists working in different domains. Furthermore, there exists the problem of explanations not having perfect fidelity i.e there is an inherent risk of explanations inaccurately representing the original model [37]. Explanations sometimes can be inconsistent as well. The outcome of the explainability tool could flag the user's short credit history as the reason for rejection. However, it is still possible to see other denied credit applications associated to long credit history [52]. Therefore, data scientists or companies/institutions working in sensitive domains should promote research to develop better explainability and interpretability tools that minimise such inconsistencies and have high fidelity. In addition to this, in some cases there could be a mismatch between the end user's mental model and the intended way to use these tools as found out by our study and other user studies [27, 24, 23]. Developers of the tools and the research community, in general, could also promote knowledge sharing and stakeholder participation. In different domains, end-users could be different from the people who build these models – auditors, lawyers, business stakeholders in the context of banking. Different end users could also comprehend fairness differently [53]. Therefore, it is important to work in the direction of understanding user requirements and designing tools by incorporating their expertise.

Bibliography

[1] Fico community. https://community.fico.com/s/explainable-machine-learning-challenge. Accessed: 12-08-2021.

[2] Uci machine learning repository: Adult data set. https://archive.ics.uci.edu/ml/datasets/adult. Accessed: 12-08-2021.

[3] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md Rafiqul Islam. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278–288, 2016.

[4] Benjamin Alarie, Anthony Niblett, and Albert H Yoon. Using machine learning to predict outcomes in tax law. *Can. Bus. LJ*, 58:231, 2016.

[5] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[6] John O Awoyemi, Adebayo O Adetunmbi, and Samuel A Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI)*, pages 1–9. IEEE, 2017.

[7] Gustavo EAPA Batista and Maria Carolina Monard. An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence*, 17(5-6):519–533, 2003.

[8] Andrew L Beam and Isaac S Kohane. Big data and machine learning in health care. *Jama*, 319(13):1317–1318, 2018.

[9] Alexei Botchkarev. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv preprint arXiv:1809.03006*, 2018.

[10] Virginia Braun and Victoria Clarke. Thematic analysis. 2012.

[11] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Interpretable deep models for icu outcome prediction. In *AMIA annual symposium proceedings*, volume 2016, page 371. American Medical Informatics Association, 2016.

[12] Mark Craven and Jude Shavlik. Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, 8:24–30, 1995.

[13] Joseph A Cruz and David S Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2:117693510600200030, 2006.

[14] Thomas H Davenport and DJ Patil. Data scientist. *Harvard business review*, 90(5):70–76, 2012.

[15] Rober Hunter DAVIS, DB Edelman, and AJ Gammerman. Machine-learning algorithms for credit-card applications. *IMA Journal of Management Mathematics*, 4(1):43–51, 1992.

[16] Derek Doran, Sarah Schulz, and Tarek R Besold. What does explainable ai really mean? a new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*, 2017.

[17] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[18] Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, Marcel van Gerven, and Rob van Lier. *Explainable and interpretable models in computer vision and machine learning*. Springer, 2018.

[19] Alex A Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10, 2014.

[20] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science Robotics*, 4(37), 2019.

[21] Sandra G Hart. Nasa task load index (tlx). volume 1.0; paper and pencil package. 1986.

[22] Trevor J Hastie. *Generalized additive models*. Routledge, 2017.

[23] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019.

[24] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26, 2020.

[25] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1):53–71, 2000.

[26] Kaggle. Adult income dataset. https://www.kaggle.com/wenruliu/adult-income-dataset/activity, Oct 2016.

[27] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[28] Sandeep Kumar. *Classification and detection of computer intrusions*. PhD thesis, Purdue University, 1995.

[29] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[30] John Lewis. Redefining qualitative methods: Believability in the fifth moment. *International Journal of Qualitative Methods*, 8(2):1–14, 2009.

[31] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[32] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018.

[33] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

[34] Darryl I MacKenzie, James D Nichols, J Andrew Royle, Kenneth H Pollock, Larissa Bailey, and James E Hines. *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Elsevier, 2017.

[35] Microsoft. interpretml/interpret. https://github.com/interpretml/interpret/blob/develop/examples/python/notebooks/Interpretable%20Classification%20Methods.ipynb, 2021. Accessed: 12-08-2021.

[36] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[37] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288, 2019.

[38] Christoph Molnar. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.

[39] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

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

[41] Lkhagvadorj Munkhdalai, Tsendsuren Munkhdalai, Oyun-Erdene Namsrai, Jong Yun Lee, and Keun Ho Ryu. An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability*, 11(3):699, 2019.

[42] B Nithya and V Ilango. Predictive analytics in health care using machine learning tools and techniques. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 492–499. IEEE, 2017.

[43] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

[44] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

[45] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

[46] Paul Ralph, Sebastian Baltes, Domenico Bianculli, Yvonne Dittrich, Michael Felderer, Robert Feldt, Antonio Filieri, Carlo Alberto Furia, Daniel Graziotin, Pinjia He, et al. Acm sigsoft empirical standards. 2020.

[47] 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*, pages 1135–1144, 2016.

[48] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[49] Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *Ieee Access*, 8:42200–42216, 2020.

[50] Alvin E Roth. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

[51] Pumitara Ruangthong and Saichon Jaiyen. Bank direct marketing analysis of asymmetric information based on machine learning. In *2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 93–96. IEEE, 2015.

[52] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[53] Debjani Saha, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, and Michael Tschantz. Measuring non-expert comprehension of machine learning fairness metrics. In *International Conference on Machine Learning*, pages 8377–8387. PMLR, 2020.

[54] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

[55] Andrew Selbst and Julia Powles. "meaningful information" and the right to explanation. In *Conference on Fairness, Accountability and Transparency*, pages 48–48. PMLR, 2018.

[56] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[57] Slundberg. slundberg/shap, Dec 2020.

[58] Branka Stojanović, Josip Božić, Katharina Hofer-Schmitz, Kai Nahrgang, Andreas Weber, Atta Badii, Maheshkumar Sundaram, Elliot Jordan, and Joel Runevic. Follow the trail: machine learning for fraud detection in fintech applications. *Sensors*, 21(5):1594, 2021.

[59] Harry Surden. Machine learning and law. *Wash. L. Rev.*, 89:87, 2014.

[60] Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. Making machine learning models interpretable. In *ESANN*, volume 12, pages 163–172. Citeseer, 2012.

[61] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10:3152676, 2017.

Appendix A

Dataset

Home Equity Line of Credit (HELOC)

A HELOC is a line of credit typically offered by a bank as a percentage of home equity (the difference between the current market value of a home and its purchase price). The customers in this dataset have requested a credit line in the range of $5,000-150{,}000$. The fundamental task is to use the information about the applicant in their credit report to predict whether they will repay their HELOC account within 2 years. This prediction is then used to decide whether the homeowner qualifies for a line of credit and, if so, how much credit should be extended.

- output variable :
  - **RiskPerformance**: The target variable to predict is a binary variable called RiskPerformance. The value "Bad" indicates that a consumer was 90 days past due or worse at least once over a period of 24 months from when the credit account was opened. The value "Good" indicates that they have made their payments without ever being more than 90 days overdue.
- input features:
  - Categorical:
    - **MaxDelq2PublicRecLast12M** - Max delinquency/Public Records Last 12 Months.
    - **MaxDelqEver** - Max Delinquency Ever
  - Numerical:
    - **ExternalRiskEstimate** - consolidated indicator of risk markers
    - **MSinceOldestTradeOpen** - number of months that have elapsed since first trade
    - **MSinceMostRecentTradeOpen** - number of months that have elapsed since last opened trade
    - **AverageMInFile** - average months in file

    - **NumSatisfactoryTrades** - number of satisfactory trades
    - **NumTrades60Ever2DerogPubRec** - number of trades which are more than 60 past due
    - **NumTrades90Ever2DerogPubRec** - number of trades which are more than 90 past due
    - **PercentTradesNeverDelq** - percent of trades, that were not delinquent
    - **MSinceMostRecentDelq** - number of months that have elapsed since last delinquent trade
    - **NumTotalTrades** - total number of trades
    - **NumTradesOpeninLast12M** - number of trades opened in last 12 months
    - **PercentInstallTrades** - percent of installments trades
    - **MSinceMostRecentInqexcl7days** - months since last inquiry (excluding last 7 days)
    - **NumInqLast6M** - number of inquiries in last 6 months
    - **NumInqLast6Mexcl7days** - number of inquiries in last 6 months (excluding last 7 days)
    - **NetFractionRevolvingBurden** - revolving balance divided by credit limit
    - **NetFractionInstallBurden** - installment balance divided by original loan amount
    - **NumRevolvingTradesWBalance** - number of revolving trades with balance
    - **NumInstallTradesWBalance** - number of installment trades with balance
    - **NumBank2NatlTradesWHighUtilization** - number of trades with high utilization ratio (credit utilization ratio - the amount of a credit card balance compared to the credit limit)
    - **PercentTradesWBalance** - percent of trades with balance

Following Special Values are present in the dataset :

- -9 : No Bureau Record or No Investigation
- -8 : No Usable/Valid Trades or Inquiries
- -7 : Condition not Met (e.g. No Inquiries, No Delinquencies)