DAKOTA STATE UNIVERSITY

DISSERTATION APPROVAL FORM

This dissertation is approved as a credible and independent investigation by a candidate for the Doctor of Philosophy degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this dissertation does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department or university.

Student Name: Shuvro Chakrobartty Student ID: A00529475

Dissertation Title:
A PERFORMANCE-EXPLAINABILITY-FAIRNESS FRAMEWORK FOR BENCHMARKING ML MODELS

Graduate Office Verification: DocuSigned by: Abby Chowning F44C8D9E621C417... Date: 11/13/2023

Dissertation Chair/Co-Chair: DocuSigned by: BE671FC34D9845C... Date: 11/14/2023
Print Name: El-Gayar, Omar

Dissertation Chair/Co-Chair: _____ Date: _____
Print Name: _____

Committee Member: DocuSigned by: Insu Park AD181583719C41C... Date: 11/14/2023
Print Name: Insu Park

Committee Member: DocuSigned by: Dr. Deb Tech 40B0654962AE4E3... Date: 11/14/2023
Print Name: Dr. Deb Tech

Committee Member: _____ Date: _____
Print Name: _____

Committee Member: _____ Date: _____
Print Name: _____

Submit Form Through Docusign Only
or to Office of Graduate Studies
Dakota State University

DAKOTA STATE UNIVERSITY

A PERFORMANCE-EXPLAINABILITY-FAIRNESS FRAMEWORK FOR BENCHMARKING ML MODELS

A doctoral dissertation submitted to Dakota State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Information Systems

October, 2023

By
Shuvro Chakrobartty

Dissertation Committee:

Dr. Omar El-Gayar
Dr. Insu Park
Dr. Deb Tech

DISSERTATION APPROVAL FORM

We certify that we have read this dissertation and that, in our opinion, it is satisfactory in scope and quality as a dissertation for the degree of Master of Science in Information Systems.

Student Name: Shuvro Chakrobartty

Dissertation Title: A performance-explainability-fairness framework for benchmarking ML models

Dissertation chair: Dr. Omar El-Gayar Date: _____

Committee member: Dr. Insu Park Date: _____

Committee member: Dr. Deb Tech Date: _____

ACKNOWLEDGMENT

I would like to express my sincere gratitude to all those who have contributed to the successful completion of my Ph.D. dissertation.

First and foremost, I am deeply thankful to my advisor, Dr. Omar El-Gayar, for his unwavering guidance, invaluable insights, and unwavering support throughout this journey. His mentorship has been instrumental in shaping my research and academic growth. I am also indebted to the members of my dissertation committee, Dr. Deb Tech and Dr. Insu Park, for their constructive feedback, expertise, and dedication to ensuring the quality of my work.

I extend my appreciation to my colleagues and peers for their stimulating discussions, encouragement, and camaraderie. Your diverse perspectives have enriched my research.

I am grateful to my family for their love, encouragement, and belief in my abilities. Your unwavering support has been my pillar of strength.

This dissertation would not have been possible without the collective efforts of all those mentioned above. Thank you for being an integral part of this academic journey.

ABSTRACT

Machine learning (ML) models have achieved remarkable success in various applications; however, ensuring their robustness and fairness remains a critical challenge. In this research, we present a comprehensive framework designed to evaluate and benchmark ML models through the lenses of performance, explainability, and fairness. This framework addresses the increasing need for a holistic assessment of ML models, considering not only their predictive power but also their interpretability and equitable deployment.

The proposed framework leverages a multi-faceted evaluation approach, integrating performance metrics with explainability and fairness assessments. Performance evaluation incorporates standard measures such as accuracy, precision, and recall, but extends to overall balanced error rate, overall area under the receiver operating characteristic (ROC) curve (AUC), to capture model behavior across different performance aspects. Explainability assessment employs state-of-the-art techniques to quantify the interpretability of model decisions, ensuring that model behavior can be understood and trusted by stakeholders. The fairness evaluation examines model predictions in terms of demographic parity, equalized odds, thereby addressing concerns of bias and discrimination in the deployment of ML systems.

To demonstrate the practical utility of the framework, we apply it to a diverse set of ML algorithms across various functional domains, including finance, criminology, education, and healthcare prediction. The results showcase the importance of a balanced evaluation approach, revealing trade-offs between performance, explainability, and fairness that can inform model selection and deployment decisions. Furthermore, we provide insights into the analysis of trade-offs in selecting the appropriate model for use cases where performance, interpretability and fairness are important.

In summary, the Performance-Explainability-Fairness Framework offers a unified methodology for evaluating and benchmarking ML models, enabling practitioners and researchers to make informed decisions about model suitability and ensuring responsible and equitable AI deployment. We believe that this framework represents a crucial step towards building trustworthy and accountable ML systems in an era where AI plays an increasingly prominent role in decision-making processes.

DECLARATION

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions or writings of another.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

_____

Shuvro Chakrobartty

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This chapter provides a comprehensive presentation of the research context, delineating the historical underpinnings of the research problem, clarifying the problem statement, and articulating the overarching research objectives. It starts with a detailed historical analysis of the research problem's evolution, subsequently transitioning to an exploration of the pivotal elements that have contributed to the formulation of the research objectives. Ultimately, this chapter concludes in a concise overview of the document's organizational structure and thematic progression.

1.1 Background of the Problem

Machine learning models have witnessed unparalleled success in a multitude of domains, transforming industries and driving innovation. They have been applied in diverse contexts, from healthcare diagnostics to autonomous vehicles, and have demonstrated remarkable capabilities in pattern recognition, prediction, and decision-making. However, alongside their proliferation, concerns regarding their performance, explainability, and fairness have become increasingly prominent.

Performance Evaluation in ML Models: The assessment of ML model performance has historically focused on traditional metrics like accuracy, precision, recall, and F1-score. These metrics provide valuable insights into a model's predictive capabilities. However, they may not adequately represent the real-world impact of a model's decisions, especially when the consequences of false positives or false negatives are substantial (Provost & Fawcett, 2013). Performance-centric evaluation alone may fail to address concerns about model bias, fairness, and the ability of stakeholders to trust and interpret model outputs.

Explainability and Interpretability: The opacity of many ML models, often referred to as the "black box" problem, has raised questions about their transparency and interpretability. Understanding why a model makes specific predictions is essential for building trust, especially in applications where decisions have significant consequences, such as medical diagnoses or autonomous vehicles. Recent advances in explainable AI (XAI) techniques have aimed to

illuminate the decision-making processes of complex models (Barredo Arrieta et al., 2020; Chakrobartty & El-Gayar, 2021; Guidotti et al., 2018; Gunning, 2017; Mohseni et al., 2018; Tjoa & Guan, 2020). These techniques include methods like Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017), which provide insights into model predictions and attribute them to input features.

Fairness in ML Models: As ML models increasingly influence decisions in areas like lending, hiring, and criminal justice, concerns about fairness have come to the forefront. Biases present in training data can lead to discriminatory outcomes and perpetuate existing inequalities. Various fairness-aware ML methods and metrics have been developed to address these concerns, including demographic parity, equal opportunity, and disparate impact (Hardt et al., 2016). Ensuring that ML models provide equitable predictions across different demographic groups is crucial for ethical AI deployment.

While each of these dimensions—performance, explainability, and fairness—holds its own importance in assessing ML model quality and suitability for real-world applications, they have often been treated in isolation. Researchers and practitioners have faced challenges in navigating the trade-offs between these dimensions, as optimizing one aspect may inadvertently compromise another.

In response to these challenges, this dissertation introduces a novel Performance-Explainability-Fairness (PEF) framework for benchmarking ML models. The PEF framework aims to unify the evaluation of ML models by considering performance, explainability, and fairness in an integrated manner. It provides a holistic view of model quality, allowing stakeholders to make informed decisions about model selection, deployment, and monitoring while addressing ethical and transparency concerns. This research builds upon the growing body of work in the fields of performance evaluation, explainable AI, and fairness-aware machine learning, aiming to bridge the gap between these critical dimensions and promote responsible and equitable AI development and deployment practices.

1.2 Statement of the problem

The rapid proliferation of ML models across diverse domains has ushered in an era of unprecedented automation and data-driven decision-making. While ML models have exhibited

remarkable predictive capabilities, the inherent complexity of these models poses multifaceted challenges, particularly concerning their performance, explainability, and fairness. These challenges underscore the need for a unified framework that can comprehensively assess ML models along these critical dimensions.

Performance Evaluation Challenges: The predominant focus on traditional performance metrics like accuracy, precision, and recall often overlooks nuances that impact model suitability. Such metrics may not adequately account for false positives or false negatives' real-world consequences, leading to suboptimal decision-making outcomes. Additionally, performance-centric evaluations do not address issues of model bias, which can have profound ethical and societal implications. Consequently, there is a growing demand for an evaluation framework that extends beyond conventional performance metrics and considers broader implications (Provost & Fawcett, 2013).

Explainability and Transparency Deficits: The opacity of many state-of-the-art ML models, including deep neural networks, has given rise to concerns about their interpretability. Stakeholders often require insights into why a model makes specific predictions, especially in contexts where model decisions have substantial consequences. While the field of eXplainable AI (XAI) has made significant progress in developing techniques for model interpretability, the integration of explainability into the broader evaluation of ML models remains a challenge. A framework is needed to systematically incorporate explainability assessments into model benchmarking.

Fairness and Bias Mitigation Imperatives: ML models are increasingly being deployed in applications that directly impact individuals and communities, such as lending, hiring, and criminal justice. Concerns about fairness and bias have escalated as biases present in training data can lead to discriminatory outcomes, perpetuating existing disparities and inequities. Fairness-aware ML methods and metrics have been proposed to address these concerns (Hardt et al., 2016). However, there is a lack of standardized approaches for integrating fairness evaluations into the overall assessment of ML models. This lack of a systematic approach inhibits the development of responsible AI systems.

Addressing these challenges requires a holistic framework that combines performance, explainability, and fairness evaluations into a unified benchmarking approach. Such a framework would enable practitioners, researchers, and policymakers to make informed

decisions about the selection, deployment, and ongoing monitoring of ML models while adhering to ethical and regulatory considerations.

1.3 Objectives of the dissertation

The primary objective of this dissertation is to address the pressing need for a comprehensive and unified methodology to evaluate and benchmark ML models. The overarching aim is to develop, validate, and demonstrate the effectiveness of a novel Performance-Explainability-Fairness framework that seamlessly integrates these three critical dimensions into a cohesive evaluation process for benchmarking ML models. Following are the specific outcome and deliverables of this dissertation.

Develop a Comprehensive PEF Framework: The dissertation will begin by laying the foundation for a Performance-Explainability-Fairness framework that extends beyond traditional performance metrics. The framework will encompass a wide range of evaluation techniques that account for model accuracy, precision, recall, F1-score, as well as more nuanced performance aspects such as overall balanced error rate, overall AUC. Drawing from recent advances in explainable AI, the dissertation will also incorporate state-of-the-art techniques for assessing model explainability and transparency (Fauvel et al., 2020). Furthermore, it will focus on fairness-aware ML methods and metrics (Hardt et al., 2016) to enable a holistic evaluation that identifies and mitigates potential biases within ML models.

Validation and Comparative Analysis: The dissertation will rigorously validate the proposed PEF framework across diverse domains and ML model types. It will include benchmarking against traditional performance-centric approaches to highlight the enhanced insights provided by the comprehensive evaluation. The validation process will emphasize real-world applications of finance, criminology, education, and healthcare prediction, to showcase the framework's utility and generalizability.

Trade-off Analysis and Model Selection Process: The dissertation offer perspectives on examining the trade-offs among performance, explainability, and fairness in machine learning models. By systematically analyzing these trade-offs, the research will showcase to practitioners and researchers on selecting models that align with specific use cases and ethical considerations. It will empower decision-makers with the tools to make informed choices regarding model deployment and to strike a balance between different evaluation dimensions.

Contributions to Responsible AI Deployment: The dissertation seeks to contribute significantly to the ongoing discourse on responsible AI development and deployment. It aims to provide a practical and versatile framework that promotes ethical, transparent, and unbiased use of ML models in high-stakes applications. By addressing the challenges associated with model performance, explainability, and fairness in a unified manner, the research endeavors to foster trust and accountability in AI systems.

Through rigorous research, comprehensive validation, and practical insights, this dissertation aspires to advance the field of ML model evaluation, offering a holistic and actionable approach that promotes the responsible, transparent, and equitable use of AI in decision-making processes.

1.4 Structure of the Dissertation

This dissertation adheres to a structured organizational framework. Chapter 2 initiates with an extensive exploration of the theoretical background and a comprehensive review of relevant literature. Chapter 3 elaborates on the research methodology that forms the foundation of this dissertation, with particular emphasis on the application of the design science research methodology, alongside an explanation of the guidelines outlined by Peffers et al. (2007). Chapter 4 dedicates careful attention to explaining the intricate details of the performance-explainability-fairness framework, which represents one of the central design artifacts within the scope of this study. Additionally, it includes a case study demonstration. Following this, Chapter 5 provides an in-depth discussion of the evaluation processes employed in this study. The findings are carefully examined and discussed in Chapter 6. The conclusion of this dissertation is provided by Chapter 7, which summarizes the contributions made, outlines the implications, acknowledges limits, and outlines potential directions for ongoing research in this area.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

ML models have become indispensable in solving complex real-world problems across various domains. Their widespread adoption has led to the development of numerous ML algorithms and models, each with its own strengths and limitations. Consequently, evaluating and benchmarking these models is crucial to ensure their effectiveness and reliability in practical applications. This literature review examines the key components of the dissertation's title: performance, explainability, and fairness, to establish the foundation for the proposed Performance-Explainability-Fairness framework for benchmarking ML models.

2.2 Performance Evaluation in ML Models

Assessing the performance of ML models is a fundamental aspect of model benchmarking. The assessment of a machine learning technique's performance relies on its ability to accurately predict outcomes for instances it hasn't encountered before. The choice of performance metrics plays a critical role in this evaluation. Typical metrics comprise accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) (Davis & Goadrich, 2006; Powers, 2020). Nevertheless, the choice of appropriate metrics depends on the specific characteristics of the problem at hand. As an example, in medical diagnosis, sensitivity and specificity may be more relevant (Shreffler & Huecker, 2022) while mean squared error (MSE) and R-squared are commonly used for regression tasks predicting a numeric outcome (Kuhn & Johnson, 2013).

While these metrics provide valuable insights into a model's performance, they may not always be sufficient. ML models often face challenges such as class imbalance, noisy data, and skewed distributions, which can bias results. Researchers have proposed methods to address these issues, such as Synthetic Minority Oversampling Technique (SMOTE) and other advanced sampling techniques (Guo et al., 2008; Hasib et al., 2020) and advanced evaluation techniques like precision-recall curves (Davis & Goadrich, 2006).

Furthermore, Gunning and Aha (2019) have highlighted an inherent tension that resides within the domain of machine learning, specifically concerning the trade-off between ML performance metrics, notably predictive accuracy, and explainability. It is often observed that the methodologies exhibiting the highest levels of performance, exemplified by Deep Learning (DL) techniques, tend to exhibit relatively lower levels of explainability. Conversely, methods renowned for their explainability, such as decision trees, frequently exhibit diminished predictive accuracy (Gunning & Aha, 2019). In contexts of critical decision-making, such as those prevalent within the medical domain, neither of these attributes can be afforded precedence over the other. As a response to this intricate challenge, researchers have sought to reconcile this tension by endeavoring to create a harmonious and balanced system. Such systems aim to optimize for both characteristics concurrently, with a particular focus on its application in the identification of patients at elevated risk of mortality (Kanda et al., 2020).

2.3 Model Explainability

The concept of explainable AI is not novel, as it finds its roots in the expert systems of the 1980s, where reasoning architectures were employed to facilitate an explanatory function within intricate AI systems (Holzinger, 2018). In expert system-based AI, the process typically begins with the codification of human knowledge, followed by the utilization of an inference engine to furnish expert decisions to non-expert users through a designated interface (London, 2019). Such systems are inherently designed to be explainable, given that the inference engine adheres to predetermined rules to arrive at decisions. However, it is noteworthy that while the initial AI systems boasted a high degree of interpretability, recent years have witnessed the dominance of opaque or “black-box” decision systems, exemplified by Deep Neural Networks (DNNs) (Barredo Arrieta et al., 2020). These black-box approaches, by their very nature, lack transparency, and consequently, they do not foster the trust and acceptance of machine learning methodologies among end-users (Holzinger et al., 2017). Conversely, transparency represents the antithesis of black-box opacity, signifying a clear and direct comprehension of the underlying mechanisms guiding a model as it deliberates upon and formulates decisions (Barredo Arrieta et al., 2020).

As ML models are increasingly deployed in sensitive domains like healthcare and finance, the need for model explainability has gained prominence. Model explainability refers to the ability to understand and interpret the decisions made by a ML model. Lack of transparency in ML models can lead to mistrust and legal or ethical concerns.

Explainability methods can be broadly categorized into model-specific and model-agnostic techniques. Model-specific methods, such as decision trees and linear regression, provide inherently interpretable models (Molnar, 2020). In contrast, model-agnostic methods, like LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017), can explain the decisions of any ML model by approximating its behavior.

Furthermore, post hoc explainability methods aim to generate explanations after the model has made predictions, while intrinsic explainability methods build interpretability into the model's architecture (Chen et al., 2018). Post hoc explainability methods are designed with the objective of producing comprehensible and insightful explanations after a machine learning model has made its predictions. These methods play a crucial role in bridging the gap between the complex, often opaque inner workings of modern machine learning models and the need for transparency and interpretability in decision-making processes. Post hoc explainability allow us to dissect and understand the rationale behind a model's predictions, shedding light on why a certain decision was made. This retrospective approach is particularly valuable in real-world applications, where decision-makers, such as doctors, policymakers, or business analysts, require not only accurate predictions but also the ability to justify those predictions and trust the underlying model. Post hoc explainability methods encompass various techniques, ranging from feature importance analysis and saliency maps to generating human-readable textual or visual explanations. In doing so, they empower users to gain insights into model behavior, detect potential biases, identify influential features, and, ultimately, enhance model performance and fairness while building trust with stakeholders. Explainability, therefore, constitutes a critical dimension of benchmarking, ensuring that models are not just accurate but also transparent and understandable.

2.4 Fairness in Machine Learning

Fairness represents a deeply cherished human value that plays a pivotal role in shaping the outcomes of various everyday decisions that have a significant impact on human lives. In

recent years, the proliferation of successful applications of AI systems has become increasingly prominent. Notably, AI methodologies are progressively integrated into a multitude of new applications designed for decision-making tasks that were historically the authority of human agents. This transition has given rise to a series of fundamental inquiries. First, there is the question of the trustworthiness of AI-driven decisions. Second, concerns emerge regarding the inherent fairness of these decisions. Collectively, these concerns raise the broader issue of whether AI-based systems are contributing to equitable decision-making or potentially exacerbating societal disparities. Within academic discourse, it is evident that a universally accepted definition of fairness remains elusive. Consequently, the determination of fairness metrics for any given ML model is contingent upon the specific contextual nuances of each situation (Mehrabi et al., 2021; Verma & Rubin, 2018). This lack of consensus can be attributed to the inherent complexity of defining fairness, compounded by the fact that stakeholders often hold divergent perspectives on what constitutes a "fair" decision across different domains of life. Moreover, a determination of fairness is highly context-dependent, with an outcome deemed equitable in one context potentially appearing inequitable in another. Nevertheless, in the realm of decision-making, fairness is typically construed as the absence of any form of bias or preferential treatment directed toward an individual or group predicated upon their intrinsic or acquired attributes (Makhlouf et al., 2021).

The issue of fairness in ML has gotten significant attention in recent years (Chakrobartty & El-Gayar, 2023). Biases in training data or the modeling process can lead to unfair or discriminatory outcomes, reinforcing existing social inequalities. To address this concern, various fairness definitions and metrics have been proposed. Notable among them are disparate impact (Zafar et al., 2017), equal opportunity (Hardt et al., 2016), and demographic parity (Dwork et al., 2012).

Furthermore, numerous algorithms have been developed to mitigate bias in ML models, such as adversarial debiasing (Zhang et al., 2018) and reweighting the training data (Kamishima et al., 2011). Benchmarking fairness in ML models necessitates the incorporation of these fairness metrics and evaluation methods into the assessment process.

2.5 Integrating Performance, Explainability, and Fairness

While each of these dimensions—performance, explainability, and fairness—is vital on its own, there is an increasing recognition that they are interrelated. A highly accurate model might not be fair, and an interpretable model might not have the best performance. Balancing these aspects is essential, and this balance varies depending on the application context.

Recent work has started to explore the trade-offs between these dimensions. In this context, Fauvel et al. (2020) have advanced an analytical framework that centers on the assessment of performance and explainability within the context of machine learning algorithms. Their framework introduces a structured set of characteristics that serve to systematize the evaluation and benchmarking of ML algorithms in relation to their performance and explainability. Furthermore, Naylor et al. (2021) have introduced a distinct framework tailored to the assessment of the boundary delineating a model's predictive performance from the quality of the explanations it provides.

MLPerf (Mattson et al., 2020; Reddi et al., 2021) is a widely recognized benchmark suite designed to evaluate the performance of machine learning systems across various tasks and hardware platforms. It covers a broad spectrum of ML workloads, including image classification, object detection, natural language processing, recommendation systems, and more. MLPerf aims to provide a standardized framework for comparing the computational efficiency and speed of ML models, thus fostering healthy competition and innovation within the field. However, it's essential to note that MLPerf primarily focuses on ML models runtime performance metrics, such as throughput and latency, and does not directly address the critical dimensions of model accuracy, explainability and fairness in machine learning. While runtime performance is undeniably crucial, issues related to model accuracy, transparency and the mitigation of algorithmic biases are equally vital in real-world applications. These aspects are not within the scope of MLPerf's evaluation, making it essential for practitioners and researchers to complement these benchmarking efforts with dedicated assessments of explainability and fairness to ensure the responsible and ethical deployment of ML systems.

Therefore, it is noteworthy that there exists a discernible gap within the existing scholarly literature concerning the adaptability of these frameworks to different classifier types, as well as their capacity to incorporate considerations of fairness alongside performance and explainability when benchmarking ML algorithms for suitability within specific usage

scenarios. Thus, our research endeavors to enhance and augment Fauvel et al.'s framework by incorporating characteristics associated with AI fairness. This augmentation extends the applicability of the framework, rendering it comprehensive for the benchmarking of ML algorithms in contexts where the algorithm's fairness constitutes a pivotal criterion for evaluation.

Benchmarking frameworks that integrate performance, explainability, and fairness can help practitioners make informed decisions about model selection, deployment, and tuning based on their specific requirements.

2.6 Chapter Summary

This literature review has provided a comprehensive overview of the key components of the dissertation's title: performance, explainability, and fairness in machine learning model benchmarking. Evaluating ML models goes beyond accuracy alone and encompasses understanding their inner workings (explainability) and ensuring equitable outcomes (fairness). The following chapter 4 will propose a Performance-Explainability-Fairness Framework for Benchmarking ML Models that synthesizes these dimensions into a unified benchmarking approach, building upon the rich body of research and methodologies discussed in this review.

CHAPTER 3

RESEARCH METHODOLOGY

This chapter outlines the research methodology employed in our study. Since the objective of this research is to construct a framework for benchmarking ML models, we have chosen a design science research methodology for this research.

3.1 Design Science Research

Design Science Research (DSR) constitutes a contemporary research methodology dedicated to the resolution of complex problems with the ultimate objective of generating transformative change. The design science paradigm is characterized by its overarching aspiration to expand the horizons of human and organizational capabilities through the conception and realization of novel and innovative artifacts (Hevner et al., 2004). This methodological approach assumes significance due to the inherent limitations of the natural science method, which primarily focuses on the observation, analysis, and explication of existing phenomena. In contrast, the creation of artifacts or the development of theoretical constructs pertinent to artificial systems is beyond the purview of natural science. Within the design-science paradigm, the pursuit of knowledge and comprehension pertaining to a specific problem domain and its potential solutions becomes intrinsically intertwined with the iterative process of conceiving, constructing, and applying the designed artifact (Hevner et al., 2004).

This integration of knowledge creation and practical application underscores the distinctive character and relevance of DSR as a research methodology. Also, the dual mandate of the DSR is outlined by (Baskerville et al., 2015): first, to use the knowledge acquired to solve problems, bring about change, or improve current solutions; second, to produce new knowledge, insights, and theoretical explanations. Therefore, unlike traditional research approaches that primarily aim to explain phenomena or develop theories, design science research is driven by the need to design, build, and evaluate practical solutions. In the context of our dissertation, the problem we are addressing is the lack of a comprehensive and integrated framework for benchmarking ML models across multiple dimensions, including performance,

explainability, and fairness. To tackle this problem effectively, we have adopted the design science research methodology.

We use the Design Science Research Methodology (DSRM) process described by Peffers et al. (2007) to guide our research. The diagram in Figure 1 describes the various components of the process model.

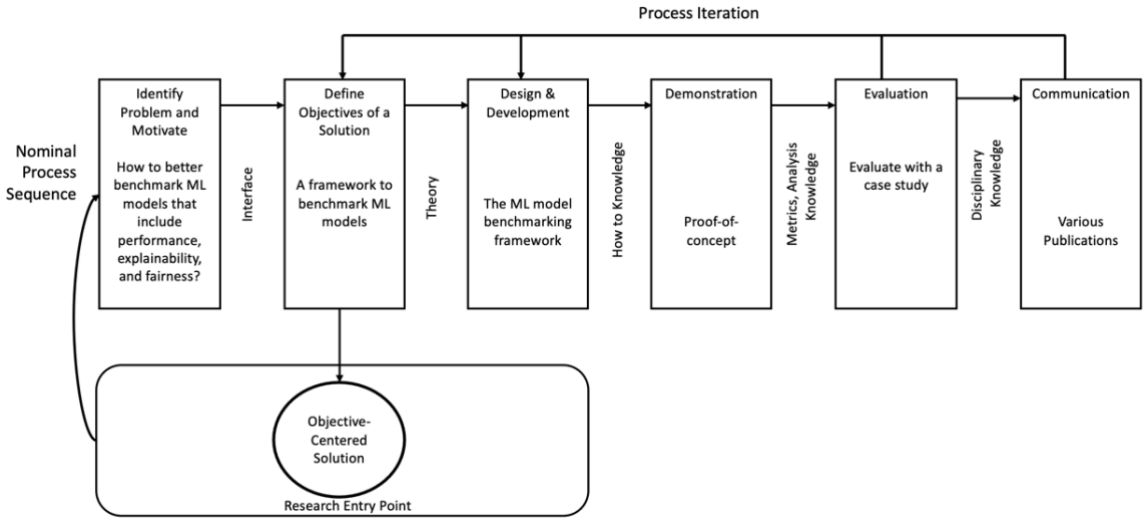

Figure 1. Design Science Research Methodology (DSRM) process

3.2 Problem Identification and Motivation

This research has an objective-centered initiation, so we motivate our research by identifying the gaps in the current ML model benchmarking approaches. The literature doesn't offer a more generic and complete framework for benchmarking ML models. Fauvel et al. (2020) introduced an analytical framework focused on the evaluation and benchmarking of ML methods with regard to their performance and explainability. This framework is designed to systematize the assessment of performance-explainability characteristics through the incorporation of a structured set of attributes. However, there is a gap and limitations of the existing framework that are summarized below.

- Fauvel et al. (2020) framework doesn't include fairness characteristics that are very important for many ML models. Without a greater understanding and performance on the fairness scale, some models may not be deployed in the real world because of their concerns with compliance and ethical issues.

- Fauvel's framework was evaluated on a specific class of Multivariate Time Series Classifiers ML model, and there are opportunities to expand the proposed framework to evaluate with other classification problems.

3.3 Define Objectives of the Solution

In Chapter 1, section titled as "Objectives of the dissertation," we have defined the objectives for building a comprehensive framework for evaluating ML models that includes performance, explainability, and fairness dimensions as well as specific outcome and deliverables. The objectives of the solution are as follows:

- We lay the foundation for a Performance-Explainability-Fairness framework that extends beyond traditional performance metrics.
- We rigorously validate the proposed PEF framework across diverse domains and ML algorithms.
- We offer perspectives on examining the trade-offs among performance, explainability, and fairness in machine learning models.
- We seek to contribute significantly to the ongoing discourse on responsible AI development and deployment.

3.4 Design and Development

The proposed artifact of this research is a framework for benchmarking ML models. Chapter 4 have demonstrated design of the proposed framework with relevant characteristics. It also demonstrated the implementation of the framework with a proof-of-concept case study demonstration as well as the evaluation with both internal and external validation process. The subsequent process in Figure 2 is employed to benchmark the models within each domain.

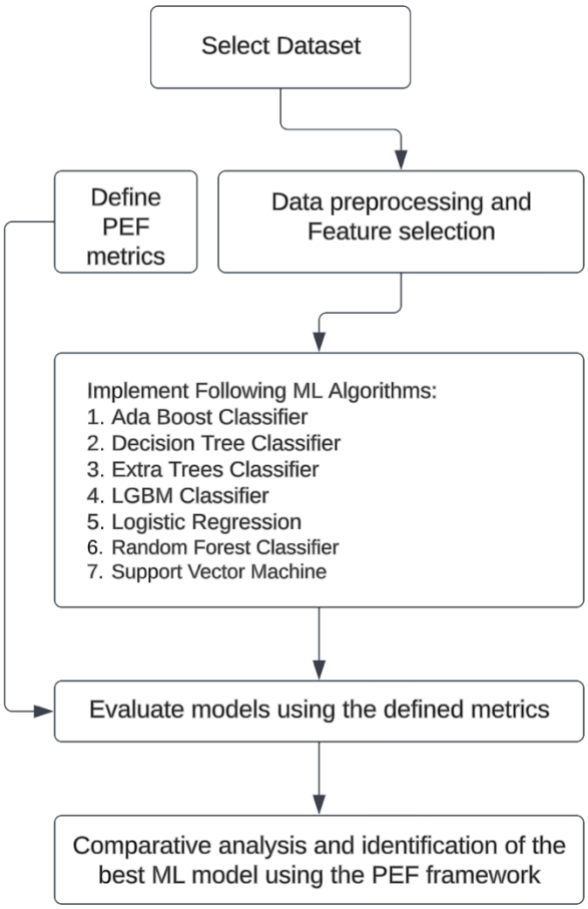

Figure 2. Model benchmarking process

3.5 Demonstration

To show how the proposed framework meet ML benchmarking, we demonstrate the research result by building a proof-of-concept implementation case study with a loan decision binary classification problem using the UCI Credit-card default dataset (Yeh, 2016).

3.6 Evaluation

The evaluation of the proposed framework is conducted systematically, following a comprehensive and structured approach. This multifaceted evaluation process is crucial to assess the effectiveness, applicability, and robustness of the Performance-Explainability-Fairness framework in the context of benchmarking ML models. The evaluation is executed in the following two distinct yet interconnected steps:

3.6.1 Internal Validation and Fine-Tuning

In this initial phase of evaluation, the primary focus is on the internal validation and fine-tuning of the PEF framework. The objectives of this step include:

- **Validation of Framework Components:** We rigorously assess the individual components of the PEF framework, including performance metrics, explainability techniques, and fairness measures. This involves evaluating the accuracy, reliability, and effectiveness of these components.
- **Integration Testing:** We examine how well the various components of the framework integrate and function together as a cohesive unit. This step ensures that the framework operates smoothly and consistently across different ML models and datasets.
- **Fine-Tuning Model Constraints:** To optimize the framework's performance, we fine-tune its parameters for both performance and fairness constraints, as well as the explainability. This fine-tuning process aims to enhance the framework's adaptability to diverse scenarios with diverse model assessment parameters for performance and fairness.

The internal validation and fine-tuning phase serve as a crucial foundation for the subsequent external validation, ensuring that the framework is robust and ready for real-world testing.

3.6.2 External Validation and Application

Building upon the insights gained from the internal validation, the external validation and application phase is designed to assess the PEF framework's performance in real-world scenarios. This phase encompasses the following key elements:

- **Application to Diverse Domains:** We apply the PEF framework to diverse domains and use cases with binary classification problems from four domains dataset including healthcare, finance, education, and criminology. This step demonstrates the framework's versatility and adaptability to different contexts.
- **Benchmarking Against Existing Approaches:** We conduct comparative benchmarking assessments, wherein we explore the comprehensiveness, effectiveness, advantages, and limitations of the PEF framework. This methodology provides us with the opportunity to explain the framework's comparative advantage in delivering

exhaustive insights pertaining to the performance, explainability, and fairness aspects of machine learning models.

- **Conduct Survey based Assessment with ML Practitioners:** We conduct a qualitative study with machine learning practitioners using a survey questionnaire. In this study, ML practitioners assess the efficacy of the proposed framework. The survey consists of multiple questions that involve comprehensive assessments of the framework's efficacy.

Through these external validation steps, we aim to demonstrate the PEF framework's real-world value, highlighting its potential to empower organizations and practitioners to make informed decisions about ML model selection and deployment while ensuring transparency and fairness.

3.7 Communication

As ongoing and iterative research, various milestones of the research have been considered for communication through various publication outlets. We have published and presented in the AMCIS 2022 Conference. The title of the paper is "Towards a Performance-explainability-fairness Framework for Benchmarking ML Models" (Chakrobartty & El-Gayar, 2022).

3.8 Chapter Summary

In summation, this chapter explored the design science research, explaining its fundamental tenets, and showcasing its relevance and application to formulate the Performance-Explainability-Fairness framework for benchmarking machine learning models. By following an iterative process of design, build, and evaluation, we've introduced a practical solution. This solution is specifically tailored to tackle the complex challenges associated with model benchmarking in the era of AI and ML. Moreover, it places a significant emphasis on addressing fairness-related aspects within the framework. In the ever-evolving landscape of research methodologies, design science research stands resolute as a valuable compass, guiding us towards innovative solutions and practical insights, thereby helping us advance knowledge.

CHAPTER 4

FRAMEWORK AND DEMONSTRATION

This study introduces an extended framework termed the "Performance-Explainability-Fairness Framework," as outlined in Table 1 and Table 2. The foundation of this framework draws from the work presented by Fauvel et al. (2020). The primary objective of this framework is to address a multitude of inquiries that may be posed by an end-user when making informed decisions based on recommendations generated by a machine learning model.

4.1 Performance-Explainability-Fairness Framework

The framework encompasses a set of evaluation characteristics, corresponding questions, and assessment values, all of which are carefully detailed in Table 1 for performance and explainability dimensions. Concerning the aspect of explainability, it is pertinent to note that Fauvel et al.'s position their framework as an extension of the fourth phase within the systematic method explained by Hall et al. (2019). Furthermore, Fauvel et al.'s (2020) framework encompasses a collection of explanatory characteristics, which are presented in Table 1, designed to structure the evaluation of ML models. It is noteworthy that this framework does not encompass application-specific implementation constraints such as considerations related to time, memory utilization, and privacy (Fauvel et al., 2020). The evaluation characteristics of the Fauvel et al.'s framework is succinctly summarized in Table 1.

Table 1. Characteristics for the performance-explainability analytical framework

Evaluation Characteristics	**Question**	**Assessment Answer Values**
Performance	"What is the level of performance of the model?"	Best, Similar, Below
Comprehensibility	"Is the model comprehensible?"	Black-box, White-box
Granularity	"Is it possible to get an explanation for a particular instance?"	Local, Global, Global & Local
Information type	"Which kind of information does the explanation provide?"	Importance, Patterns, Causal
Faithfulness	"Can we trust the explanations?"	Imperfect, Perfect
User category	"What is the target user category of the explanations?"	Domain Expert, ML Expert, Broad Audience

Here we discuss more about each of these characteristics and their assessment answer values.

4.1.1 Performance

With this characteristic the framework assesses machine learning model performance by considering various metrics such as accuracy, F-measure, and Area Under the ROC Curve. However, there's no consensus on an evaluation procedure for this aspect. The framework introduces a performance component as the initial step towards standardizing the assessment of machine learning models, evaluating their relative performance in specific applications. It categorizes performance into three levels:

- **Best:** representing the top-performing model in a given application.
  Similar: referring to models that demonstrate performance comparable to the top-performing model but are ranked second based on the same evaluation setting. This designation includes all models that do not display a statistically significant difference in performance compared to the second-ranked model under the same evaluation setting; and
- **Below:** denoting models performing less effectively than the state-of-the-art models under the same evaluation conditions.

4.1.2 Comprehensibility

This characteristic explores model comprehensibility, which relates to a user's ability to understand how a model functions and makes predictions. Comprehensibility is closely tied to model complexity, but there's no consensus on how to assess it. Models are generally categorized as "white-box," which are easy to understand, or "black-box," which are complex and harder to comprehend. Examples of "white-box" models include rule-based models and decision trees, while ensemble methods and deep learning models fall into the "black-box" category. However, not all rule-based models or decision trees are necessarily easy to understand, as human cognitive limitations impose constraints on the complexity of models that can be comprehended.

- **Black-Box:** models that are complicated-to-understand.
- **White-Box:** models that are easy-to-understand.

The framework distinguishes between "white-box" and "black-box" models from a comprehensibility perspective.

4.1.3 Granularity

This characteristic discusses the availability of explanations for specific instances in machine learning models, emphasizing the granularity of these explanations. Typically, two levels are recognized: global and local explanations. Global explanations pertain to the model's overall behavior across the entire dataset, while local explanations provide insights into a particular prediction. Certain methods offer either global or local explanations exclusively, while others, such as decision trees, can provide both types of explanations. Therefore, following three categories of granularity is used for evaluating granularity characteristic.

- **Global:** global explainability.
- **Local:** local explainability.
- **Global & Local:** both global and local explainability.

4.1.4 Information type

This characteristic explores the type of information conveyed by explanations in machine learning. The most valuable information aligns closely with human reasoning, encompassing causal and counterfactual rules. Causal rules can clarify that specific observed variables cause particular model predictions. However, machine learning typically relies on statistical associations within data and doesn't probe into causal relationships among observed and unobserved variables. These associations vary depending on the machine learning task. Therefore, a generic assessment of information type, categorized from least to most informative is used as described below:

- **Importance:** which reveals the relative importance of each dataset attribute.
- **Patterns:** which provides predefined semantic conjunctions associated with predictions; and the most informative.
- **Causal:** which presents explanations in the form of causal rules.

4.1.5 Faithfulness

This characteristic examines the trustworthiness of explanations provided by machine learning models. Trust, in this context, pertains to the extent to which end-users can rely on these explanations, i.e., how closely they relate to what the model actually computes. Explanations derived directly from the original model are inherently trustworthy. However,

some post-hoc explanation methods attempt to approximate the "black-box" model's behavior with a more explainable surrogate model. These surrogate models may not perfectly mirror the original model's behavior, introducing questions about their faithfulness. The fidelity criteria are used to measure the faithfulness by assessing how closely the surrogate model imitates the predictions of the original model. An assessment of faithfulness in two categories are used as outlined below.

- **Imperfect:** imperfect faithfulness (use of an explainable surrogate model).
- **Perfect:** perfect faithfulness.

4.1.6 User category

This characteristic digs into the target user category for explanations in machine learning. It emphasizes that the accessibility of explanations depends on the user's background and experience, as it influences how they organize information. To enhance explanation accessibility, it's important to categorize user types and determine which users will have access to the explanations. The broader the audience that can understand the explanations, the more effective they are. Consequently, the assessment proposed in the text encompasses three categories.

- **Machine Learning Expert:** who build the ML models.
- **Domain Expert:** domain experts (e.g., professionals, researchers)
- **Broad Audience:** non-domain experts (e.g., policy makers).

In an extension of the existing framework (Chakrobartty & El-Gayar, 2022) presented in Table 2, this study augments the framework originally proposed by Fauvel et al (Fauvel et al., 2020). The augmentative aspect of this extended framework is particularly dedicated to the assessment of fairness concerning ML models.

Table 2. Extended characteristics of the PEF analytical framework

Evaluation Characteristics	Question	Assessment Answer Values
Fairness Context	For whom is it fair? Is it fair for individual or group/sub-group or both?	Individual, Group, Subgroup, Both (individual, group), All
Fairness	What is the level of fairness of the model?	Best, Similar, Below

Within this expanded framework, specific evaluation characteristics are introduced, which are designed to provide insights into the fairness considerations associated with the ML model.

4.1.7 Fairness Context

In essence, this characteristic aim to address the fundamental question of *"for whom is it fair?"* The identification of the fairness context assumes significant importance due to the nuanced nature of fairness assessments. It is noteworthy that while statistical parity-based group fairness endeavors to equalize outcomes between protected and non-protected groups, the resultant outcomes may still exhibit substantial unfairness when viewed from the perspective of an individual (Dwork et al., 2012). This acknowledgment underscores the intricacies and multifaceted dimensions inherent in fairness evaluations within the context of ML models.

Furthermore, as articulated by Kearns et al. (2017), it is imperative to recognize that in the context of group fairness, a classifier may convey an impression of fairness when evaluated individually for each distinct group. However, despite such individual assessments, the classifier may still exhibit significant deviations from fairness criteria when scrutinized within the context of one or more structured subgroups. These structured subgroups are specifically constituted by unique combinations of protected attribute values encompassing all the protected attributes under consideration. In the pursuit of a comprehensive fairness evaluation, this study identifies various dimensions of fairness, including individual fairness, group fairness (Speicher et al., 2018), sub-group fairness (Kearns et al., 2017, 2019), and a synthesis of these dimensions. This multifaceted approach is instrumental in capturing and evaluating the manifold fairness considerations inherent in the model under assessment.

- **Individual fairness:** Individual fairness is a critical facet of fairness in machine learning and artificial intelligence systems. Individual fairness looks at fairness on a case-by-case basis, emphasizing the treatment of similar individuals. It implies that similar individuals, as defined by specific features, should receive similar outcomes or predictions from a model (Dwork et al., 2012).
- **Group fairness:** Group fairness, on the other hand, focuses on fairness at the group level. It entails that different demographic groups, such as gender, race, or age, should be treated equitably by the model, reducing the potential for bias and discrimination at a larger scale (Speicher et al., 2018).
- **Sub-group fairness:** Sub-group fairness probes deeper into group fairness by examining fairness within specific sub-groups defined by the intersection of multiple attributes or features. This approach ensures fairness not only among broad

demographic groups but also among more nuanced sub-populations, contributing to a more comprehensive and nuanced assessment of fairness in machine learning models. Achieving fairness across all these levels is essential to building equitable and trustworthy AI systems (Kearns et al., 2017, 2019).

4.1.8 Fairness

The characteristic associated with fairness evaluation serve the purpose of determining the extent or degree of fairness exhibited by the ML model under examination - *what is the level of fairness of the model?* It is noteworthy that a multitude of fairness concepts and notions have been developed to facilitate the assessment of an ML model's fairness. Heidari et al. (2019) offer an insightful interpretation wherein they map preexisting conceptions of algorithmic fairness, particularly within the domain of binary classification, as specific instances of the Equality of Opportunity (EOP) principle derived from economic models. Additionally, Speicher et al. (2018) contribute by delineating various fairness notions along with their corresponding fairness conditions for evaluating the fairness of an ML model. It is important to note, however, that a consensus has yet to emerge within the academic discourse pertaining to a standardized evaluation methodology for discerning the fairness exhibited by an ML model. This absence of a universally accepted evaluation technique underscores the ongoing debate and the complex nature of fairness assessment in the context of ML models.

The selection of an appropriate metric for assessing the fairness of an ML model is contingent upon the specific application in question. The choice of a metric is made with careful consideration of its alignment with the intended objectives of the experiments conducted within that application. This approach ensures that the fairness assessment is tailored to the unique requirements of each application, thereby precluding direct comparisons of fairness between ML models across divergent applications. It is essential to underscore that the assessment of fairness is inherently relative and context-specific, indicative of a model's fairness concerning a particular application. Furthermore, this relative assessment extends to the comparison of models against a designated top-ranked model within a given application and evaluation scenario. This conceptual framework facilitates the classification of models with respect to their fairness attributes, specifically within the context of application and evaluation. Notably, the fairness component affords the ability to compile a roster of models that have surpassed the

performance of current state-of-the-art top-ranked models within their respective applications, particularly in scenarios encompassing diverse applications characterized by similar ML challenges. Consequently, this approach identifies specific models that may warrant consideration for evaluation in novel applications. However, it is imperative to acknowledge that the superior performance of these models within their original applications does not guarantee analogous performance within new and distinct application domains. Building upon the foundational work of Fauvel et al. (2020), this study proposes a classification of fairness into three distinct categories:

- **Best**: representing the leading fair model in a given application, i.e., it ensures the highest degree of fairness. It refers to the fairness of the top-ranked model on the application based on evaluation setting such as models, fairness evaluation methods, and datasets.

  Similar: indicating models with fairness similar to the leading fair model, but of second ranked based on the same evaluation setting. It refers to all the models which don't exhibit a statistically significant fairness difference from the second-ranked model under the same evaluation setting; and

  Below: fairness is lower than the state-of-the-art models. Given the same evaluation setting, it refers to the fairness rank of the remaining models.

4.2 Case Study Demonstration

In order to illustrate the practical applicability of our framework, we provide a detailed case study utilizing a dataset from the finance domain. This case study serves as a concrete example of how our framework can be employed to assess and benchmark machine learning models in real-world scenarios. By applying our methodology to this finance dataset, we showcase its versatility and effectiveness in promoting transparency, fairness, and improved model performance in a domain where these aspects are of paramount importance. As shown in the Figure 3, multiple ML models are trained with the dataset then a set of performance, explainability and fairness criteria is assessed against the test dataset result to evaluate the models with the PEF framework which helps with identifying the best ML model.

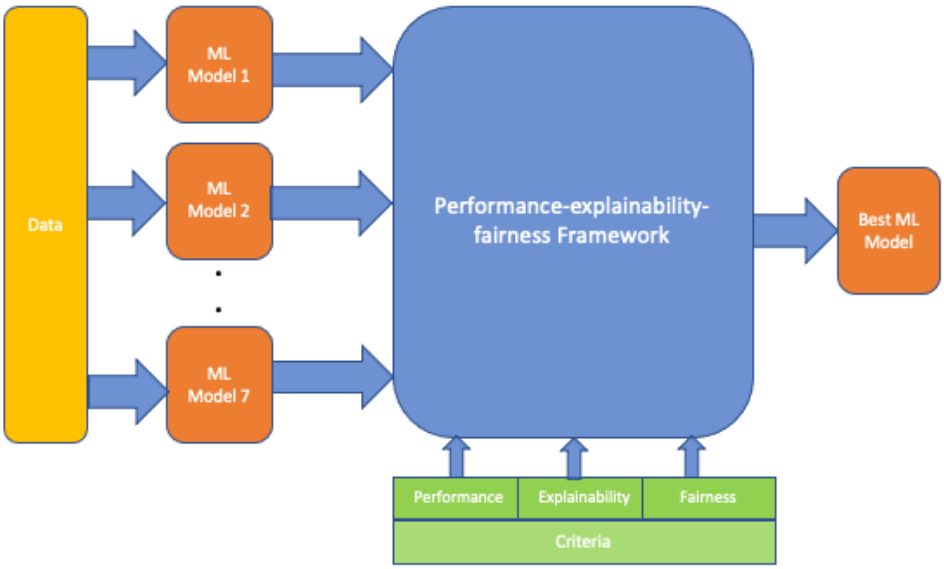

Figure 3. Model evaluation with PEF framework

4.2.1 Finance Domain

For the purpose of this study, we have chosen the binary classification problem of predicting default payments. To conduct our research, we utilized an openly accessible dataset sourced from the University of California Irvine (UCI) Machine Learning Repository, specifically the "default of credit card clients Data Set" (Yeh & Lien, 2009). The primary objective behind the acquisition of this dataset was to facilitate a comparative assessment of data mining techniques, specifically focusing on their efficacy in predicting the probability of credit card clients defaulting (Yeh & Lien, 2009). This dataset captures the credit card payment history of customers and was originally employed in a study aimed at forecasting the likelihood of customers defaulting on their payments in Taiwan.

Notably, this dataset encompasses a substantial sample size, comprising 30,000 individual records, and features a diverse array of attributes, including but not limited to age, gender, and marital status, which possess the potential to be associated with discriminatory practices. It has 6,636 samples representing the minority class labeled as "Yes," indicating customers who are likely to default in the following month. The majority class, labeled as "No," comprises 23,364 samples, indicating customers who are not expected to default in the next month. The dataset includes twenty-three explanatory variables, categorized into five static and eighteen dynamic features, with further details provided in Table 15. Notably, this dataset has

been employed in thirteen research papers (Chakrobartty & El-Gayar, 2022), where prediction accuracy results have been reported in prior studies.

The central purpose of our utilization of this dataset is to engage in predictive modeling with the aim of estimating the likelihood of credit default incidents. Following the model benchmarking process described in Figure 2, we evaluate the PEF framework. The evaluation process is described here.

4.2.2 Data Preprocessing

We first remove the ID column, we ensure the data is cleaned and have no missing values, also ensure the categorical features are marked as dataframe "category" type. To address the imbalanced distribution of classes in the dataset, we first resampled the training data to create a balanced training set. This step is crucial when dealing with imbalanced datasets, where one class significantly outnumbers the others. Next, we removed the gender feature from the dataset and split the remaining data into 70% training and 30% testing sets. This split was performed in a stratified manner to ensure that the proportion of target classes or labels is preserved in both the training and testing datasets.

4.2.3 Model Evaluation Metrics

It is imperative to acknowledge the multifaceted implications of default occurrences within the realm of credit transactions. Such occurrences have widespread ramifications, affecting not only the financial institution but also the borrower, the institution's other clients, and potentially the broader societal fabric. Lenders are confronted with financial losses in the event of client defaults, which may necessitate the transfer of such economic burdens onto their remaining clientele and investors. Additionally, borrowers who experience default may encounter considerable adversity, potentially impeding their future borrowing prospects, thereby underscoring the long-term repercussions of default occurrences. Furthermore, the persistent extension of credit to clients with a tendency for defaulting can substantially impact the overall creditworthiness of the financial institution and may even reverberate throughout the broader regional economy.

In accordance with the framework outlined in this study, we have identified a set of comprehensive fairness metrics tailored to the specific application involving the "default of

credit card client" dataset. These metrics play a crucial role in our assessment of model fairness, addressing the fundamental question of how fair the models are in their predictions. We have chosen to focus exclusively on the "gender" attribute as the demographic group of interest for constructing models and evaluating fairness metrics across different groups. This approach allows us to investigate and quantify the extent of fairness or potential disparities between these demographic subgroups within the dataset.

The subsequent metrics are sourced from Microsoft's Fairlearn toolkit (Bird et al., 2020), an essential resource designed for the analysis and enhancement of AI fairness. Fairlearn metrics package documentation (2021) provide a comprehensive understanding of the calculations associated with these metrics.

Demographic Parity Difference: Demographic parity serves as a pivotal fairness metric, denoting a scenario where a model's classification outcomes are independent of a specific sensitive feature, such as "gender" in our context. The attainment of demographic parity implies that the proportion of defaults among males is equivalent to that among females, regardless of other distinguishing characteristics within these groups. A lower value signifies a superior degree of demographic parity between the groups.

Demographic Parity Ratio: The demographic parity ratio is precisely defined as the ratio between the lowest and highest group-level selection rates, encompassing all values of the sensitive feature(s). A demographic parity ratio of 1 signifies uniform selection rates across all groups, indicating an equitable model outcome irrespective of sensitive attributes.

Equalized Odds Difference: This fairness metric evaluates whether a classifier exhibits comparable predictive accuracy across all attribute values. Equalized odds are realized when individuals, regardless of their gender, are equally likely to receive a default prediction if they meet the criteria for defaults. Likewise, they should be equally likely to receive a non-default prediction if they do not qualify for defaults. Smaller values of this metric indicate enhanced equalized odds parity between groups. An equalized odds difference of 0 suggests that true positive, true negative, false positive, and false negative rates are consistent across all demographic groups, demonstrating fairness in prediction outcomes.

In conjunction with the fairness metrics, this study also encompasses the utilization of performance metrics, which are accessible through the Python library provided by the Fairlearn

toolkit. These performance metrics are instrumental in evaluating the performance of the ML models within the context of this case study:

Overall Balanced Error Rate (BER): The Balanced Error Rate is calculated as the average of the errors incurred on each class. It offers a comprehensive assessment of classification accuracy that considers the performance across different classes, where "class" refers to the different categories or labels used in a classification problem.

Balanced Error Rate Difference: This metric quantifies the disparity in Balanced Error Rate between distinct groups. A lower value signifies a more equitable distribution of classification errors between these groups, with values approaching zero indicating a higher degree of fairness.

Overall Area Under the Curve (AUC): The Area Under the Curve (AUC) is determined by analyzing the Receiver Operating Characteristics (ROC) curve, representing a vital measure for assessing the classification model's performance. A high AUC score, closer to 1, indicates superior model performance in distinguishing between positive and negative classes.

AUC Difference: The AUC difference metric measures the distinction in AUC scores between different groups or categories. A lower AUC difference signifies a reduced gap in discriminatory power between groups, reflecting a more balanced model performance across these categories.

These performance metrics, in tandem with fairness metrics, enable a comprehensive evaluation of both the predictive accuracy and fairness aspects of the ML models under investigation. We report the performance of the models using 1) overall balanced error rate, 2) balanced error rate difference, 3) overall AUC, and 4) AUC difference. However, we use the overall AUC and balanced error rate difference to rank the models for performance. For fairness we report 1) equalized odds difference, 2) overall selection rate, and 3) demographic parity difference; while ranking the models with the equalized odds difference farness metric.

4.2.4 Model Selection

Starting with the findings from a comprehensive literature review conducted by Chakrobartty and El-Gayar (2022), as well as expanding to other classifiers, we design an experiment involving seven classifiers: 1) Adaptive Boosting (ADB), 2) Decision Tree (DT), 3) Extremely Random Trees (ET), 4) Gradient Boosting Machines (GBM), 5) Logistic Regression (LR), 6)

Random Forest (RF), and 7) Support Vector Machine (SVM). For this specific case study, we harnessed the capabilities of the Microsoft Fairlearn (Bird et al., 2020) tool, as well as its Python library, to measure the fairness of the aforementioned machine learning classifiers.

4.2.5 GridSearch and Hyperparameter Tuning

Using Fairlearn's GridSearch we perform hyperparameter tuning by systematically exploring different combinations of hyperparameters for a given machine learning estimator, in our case the estimators are seven different classifiers. These hyperparameters can include things like learning rates, regularization strengths, or tree depths, depending on the chosen estimator. We specify equalized odd as the constraints that quantify disparities or biases in model predictions across different demographic or sensitive groups – "gender" is selected for this domain specific dataset. GridSearch trains and evaluates multiple models with different hyperparameters while taking fairness metrics into account. It aims to identify models that optimize a trade-off between predictive performance and fairness. The goal is to find models that minimize disparities or biases while maintaining acceptable levels of accuracy for our model. After evaluating models using the specified fairness and performance metrics, GridSearch selects the model that best aligns with the performance and fairness criteria we've set.

4.2.6 Model Interpretability

SHAP (Shapley Additive exPlanations) is used for interpreting our ML models. It provides a unified approach to explain the output of any machine learning model by attributing the prediction to the contribution of each feature. Positive SHAP values contribute to increasing the prediction, while negative values contribute to decreasing it. Features with larger absolute SHAP values have a greater impact on the prediction.

4.2.7 Model Comparison

Instead of training a single model, we chose to train multiple models for a single classifier with GridSearch, each representing various trade-offs between the performance metric (balanced accuracy) and the fairness metric (equalized odds difference). Then the best model was selected from the candidate models for that specific classifier. This holistic approach was applied to all seven classifiers mentioned earlier, and the results of this experiment are summarized in Table 3, providing a spectrum of fairness and performance metrics for each

model used. The analysis primarily focused on the equalized odds difference, a critical fairness metric, to assess whether each classifier exhibited equitable label prediction capabilities.

Table 3. Fairness and performance metrics for default credit dataset

	Metrics	Classifiers						
		ADB	DT	ET	GBM	LR	RF	SVM
Fairness	Overall selection rate	0.382333	0.232222	0.236333	0.32	0.156111	0.991889	0.017667
	Equalized odds difference	0.004092	0.007595	0.009818	0.000656	0.008492	0.005751	0.004086
	Demographic parity difference	0.014353	0.009872	0.016979	0.012365	0.004741	0.005579	0.002378
Performance	Overall balanced error rate	0.327875	0.384994	0.305404	0.286566	0.43964	0.498017	0.486513
	Balanced error rate difference	0.000114	0.006839	0.002908	0.000499	0.004764	0.000102	0.001975
	Overall AUC	0.731333	0.615006	0.754586	0.777885	0.653943	0.723391	0.512733
	AUC difference	0.002951	0.006839	0.014316	0.011087	0.010341	0.010428	0.007165

After obtaining candidate models through grid search, for the specific machine learning algorithm, that meet the model constraints, we choose the best model that aligns with our evaluation metrics. Below Figure 4 provide the comparison of the seven different models based on the performance and fairness.

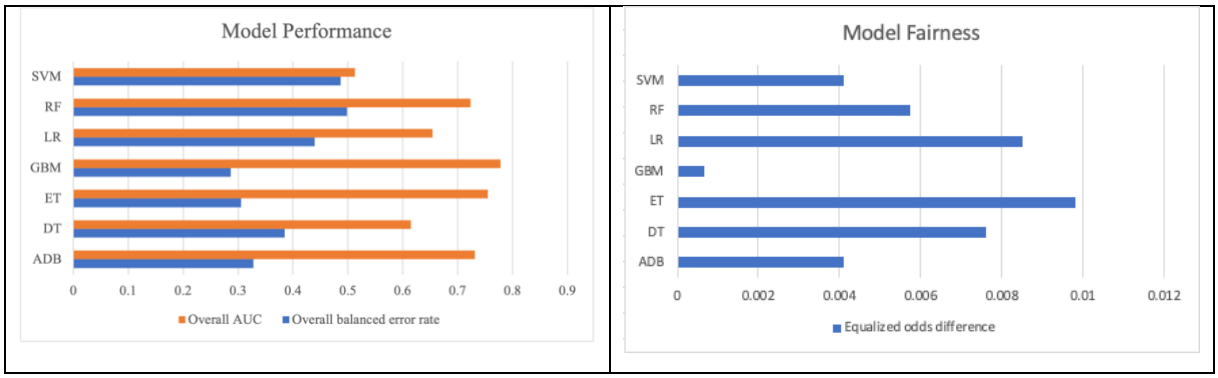

Figure 4. Comparison of seven models for default credit dataset

The experimental findings reveal that the GBM classifier-based model attains the highest overall AUC value, lowest balanced error rate value, signifying superior performance in discriminatory ability. Moreover, the GBM classifier-based model exhibits a marginal AUC difference across different gender categories, indicating consistency in its predictive performance. ET classifier-based model closely follow the GBM's having similar performance. Consequently, we recognize the GBM classifier-based model as the *best* in terms of performance. The ET classifier-based model, ranking second, demonstrates performance comparable as *similar* to the GBM's, while other models exhibit lower performance levels marking them as *below* performance.

In the realm of fairness, again GBM classifier emerges as the frontrunner, giving the lowest equalized odds difference. These metrics collectively affirm the GBM model as the

fairest and performant among the models evaluated. Notably, the SVM and ADB classifier-based models approach the second ranked fairness metrics comparing that of the GBM model, however maintaining a significantly lower AUC value, indicative of its lacking performance.

Subsequently, these results are seamlessly integrated into the extended Performance-Explainability-Fairness framework. The comprehensive summary of the extended framework outcomes for the seven default credit classifiers is presented in Table 4, encapsulating the model characteristics and their respective performances across the dimensions of performance, explainability, and fairness.

In light of the incorporation of these metrics, we construct a summary Table 4 encapsulating the characteristics and outcomes of the extended framework as applied to the identified default credit classifiers. This presentation serves as a valuable reference point, explaining how the choice of ML model may diverge contingent upon the specific requirements and trade-offs necessitated by the use case, which must reconcile considerations of performance, explainability, and fairness.

Table 4. Summary of extended framework results for default credit dataset

Evaluation Characteristics	**ADB**	**DT**	**ET**	**GBM**	**LR**	**RF**	**SVM**
Performance	Below	Below	Similar	Best	Below	Below	Below
Comprehensibility	Black-box	White-box	Black-box	Black-box	Black-box	Black-box	Black-box
Granularity	Global & Local	Global & Local	Global & Local	Global & Local	Global & Local	Global & Local	Global & Local
Information type	Feature Importance	Feature Importance	Feature Importance	Feature Importance	Feature Importance	Feature Importance	Feature Importance
Faithfulness	Imperfect	Perfect	Imperfect	Imperfect	Imperfect	Imperfect	Imperfect
User category	Domain Expert	Domain Expert	Domain Expert	Domain Expert	Domain Expert	Domain Expert	Domain Expert
Fairness context	Group	Group	Group	Group	Group	Group	Group
Fairness	Similar	Below	Below	Best	Below	Below	Similar

Regarding model comprehensibility, only the DT model qualifies as a “white-box” model, distinguished by its interpretability, whereas the remaining six classifiers, of which some are ensemble models, all fall under the category of “black-box” models.

In the realm of granularity, the DT model excels in providing explanations at both local and global level. Equally, the other classifiers, equipped with the SHAP method, demonstrate the capability to furnish both global and local level explanations. Furthermore, all classifiers yield feature importance information as part of their explanations.

In terms of faithfulness, the DT model stands out as a model with perfect faithfulness, as explanations can be directly extracted from its original model. In contrast, the other models rely

on post-hoc explanation methods employing surrogate models, rendering them imperfect in terms of faithfulness.

Turning to the fairness context, all classifiers are evaluated with respect to group fairness, with the GBM model exhibiting the highest level of fairness. The SVM and ADB model closely approaches the fairness metrics of the GBM model, and its fairness can be considered as comparable to that of the GBM model. However, the other classifiers fall below the ADB and SVM models in terms of fairness. These findings are visually illustrated in Figure 5.

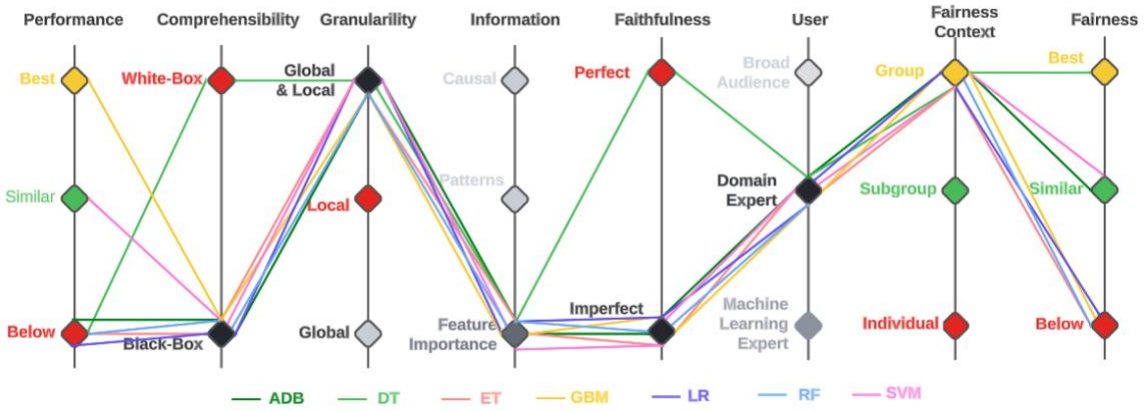

Figure 5. Parallel coordinates plot of the PEF framework result for default credit dataset

In conclusion, considering the multifaceted dimensions of the performance-explainability-fairness framework applied to the default credit score dataset, the GBM classifier-based model emerges as the most suitable choice, excelling across various characteristics and striking a favorable balance between performance, explainability, and fairness.

4.3 Chapter Summary

In this chapter, we have detailed the framework and demonstrated a practical implementation of it. We begin by thoroughly explaining the development and formulation of the PEF framework. We explore the critical components that constitute this innovative artifact: performance, explainability, and fairness. Additionally, we showcase the application of the PEF framework in a real-world scenario. Through a comprehensive case study, we provide firsthand insights into how the framework operates, how it facilitates model selection, and how it ensures that AI systems are not only accurate but also transparent and equitable.

CHAPTER 5

EVALUATION

With the case study of four domains dataset, we demonstrate the application of the performance-explainability-fairness framework for benchmarking ML models. Also, we evaluate the efficacy of the proposed framework by applying it to different binary classification problems from these four domains dataset. This confirms the framework's applicability under different fairness constraints, sensitive attributes, and class ratios within the datasets. The class ratio represents the number of observations belonging to one class vs. those belonging to the other. We evaluate the identified classifiers for each dataset against the new characteristics, emphasizing fairness. The following Table 5 shows a list of four different datasets from four domains. Note that we used the finance domain dataset in previous section for a case study demonstration, so in this chapter we evaluate the framework by using the other three datasets.

Table 5. Dataset from four different domains for evaluation

Dataset	Location	Domain	Sensitive attribute	Target class	Class ratio	Example count
Credit card clients	Taiwan	Finance	Gender	Default payment	60% female, 40% male	30,000
Diabetes	USA	Healthcare	Race	Readmit in 30 days	75% Caucasian, African American 19%, Other 6%	101,766
COMPAS recidivism	USA	Criminology	Race	Two-year recidivism	51% African-American, 34% Caucasian, 9% Hispanic, 5.2% Other, 0.45% Asian, 0.25% Native American	7,214
Law School Admission	USA	Education	Race	Pass the bar exam	85% White, 16% Non-White	20,798

The model evaluation process shares three common aspects across all datasets, as outlined in section 5.1. Further details on the specific evaluation steps for each dataset are described in their respective sections.

5.1 Evaluation Common Steps

5.1.1 Model Selection

In our evaluation process, we have employed a diverse set of seven machine learning algorithms, each serving as a distinct tool in our analytical toolbox. These algorithms include: 1) Adaptive Boosting (ADB), 2) Decision Tree (DT), 3) Extremely Random Trees (ET), 4) Gradient Boosting Machines (GBM), 5) Logistic Regression (LR), 6) Random Forest (RF), and

7) Support Vector Machine (SVM). These algorithms were thoughtfully selected for their versatility and compatibility with Fairlearn's GridSearch API, a valuable utility that facilitates grid search hyperparameter optimization. The primary aim of this optimization process is to identify machine learning models that strike a balance between predictive performance and fairness. This functionality is particularly well-suited for addressing fairness concerns within the realm of machine learning tasks. By leveraging this tool, we can systematically explore a range of hyperparameter configurations for each algorithm to determine the best-performing models. This grid search optimization process is essential for tailoring our models to achieve not only high predictive accuracy but also fairness in their predictions. As such, it plays a crucial role in ensuring that our models meet ethical and fairness criteria while delivering meaningful insights and predictions.

5.1.2 GridSearch and Hyperparameter Tuning

Within our methodology, we harness the power of Fairlearn's GridSearch as a critical component for hyperparameter tuning in our machine learning models. This process involves a systematic exploration of various hyperparameter combinations for a given machine learning classifier. The specific hyperparameters under consideration may encompass factors such as learning rates, regularization strengths, or tree depths, depending on the chosen estimator.

In our case, we prioritize the "equalized odds" constraints, which serve as quantifiable measures of disparities or biases present in the model predictions across different demographic or sensitive groups. For our analysis, a specific sensitive group is chosen based on the context of our domain-specific dataset, as outlined in Table 5. GridSearch methodically trains and evaluates multiple models with distinct hyperparameter configurations, all while conscientiously taking into account fairness metrics. It's important to note that when employing GridSearch, we begin by excluding the sensitive feature from the dataset before performing the dataset split into training and test subsets. Within the GridSearch framework, the fit() method plays a pivotal role by facilitating the fitting of the training dataset to an estimator, with the sensitive features explicitly passed as parameters to the fit() method. Its overarching objective is to pinpoint models that strike an optimal balance between predictive performance and fairness. We seek to identify models that minimize disparities or biases in predictions while still maintaining an acceptable level of accuracy for our models.

Following a rigorous evaluation process that leverages our designated fairness and performance metrics, GridSearch adeptly singles out the model that aligns most closely with the established performance and fairness criteria we've defined. This ensures that the models generated through this process not only excel in terms of their predictive capabilities but also adhere to the ethical and fairness standards we've set, thus facilitating responsible and unbiased machine learning outcomes.

5.1.3 Model Interpretability

We employ SHAP as a vital tool for interpreting our machine learning models. SHAP offers a comprehensive and unified framework for explaining the output of various machine learning models. It achieves this by breaking down the prediction into individual feature contributions, allowing us to understand how each feature influences the model's output. In this context, SHAP values play a pivotal role. Positive SHAP values signify contributions that increase the model's prediction, while negative values indicate contributions that decrease it. The magnitude of SHAP values is equally important, as features with larger absolute SHAP values exert a more substantial impact on the model's predictions. This approach enables us to gain a nuanced understanding of which features drive the model's decision-making process and to what extent, enhancing our ability to interpret and make informed decisions based on the model's outputs.

5.2 Healthcare Domain

The Diabetes 130-Hospitals Dataset (Clore & Strack, 2014) encompasses a decade of clinical care records from 130 healthcare facilities and integrated delivery networks in the United States. Each entry corresponds to a patient's hospital admission for a diabetes diagnosis, with stays ranging from one to fourteen days. These hospital encounters involve various aspects, including laboratory tests, medication administration, and medical procedures. The dataset provides a comprehensive set of attributes, including patient demographics, diagnostic information, details about diabetic medications, the number of healthcare visits in the year prior to the admission, and payer-related data. Additionally, it tracks whether patients were readmitted post-discharge and, if so, whether this readmission transpired within 30 days of their initial release. The details of the dataset description are described in Table 16.

5.2.1 Data Preprocessing

We first create outcome variables, two binary outcome variables, readmit_30_days and readmit_binary, are created based on the readmitted column. The attribute readmit_30_days is True if a patient was readmitted within 30 days, and readmit_binary is True if a patient was readmitted (not equal to NO). Then we replace missing values denoted by "?" in columns like age, payer_code, medical_specialty, and race with appropriate labels like an empty string, Unknown, or Missing. We make appropriate recoding for categorical variables. Several categorical variables are recoded. For example, admission_source_id is recoded into Referral, Emergency, or Other. The attribute age is grouped into categories like 30 years or younger, 30-60 years, and Over 60 years. We also clean medical codes, the discharge_disposition_id is transformed to Discharged to Home if it originally had a value of 1. We also do re-coding of medical specialties and primary diagnosis, the medical_specialty column is recoded into a limited set of specialties or Other. The primary_diagnosis column is also recoded into categories such as Respitory Issues, Diabetes, Genitourinary Issues, and Musculoskeletal Issues. We also create several binary features, such as medicare and medicaid based on the payer_code. Additionally, binary features like had_emergency, had_inpatient_days, and had_outpatient_days are created based on numerical columns.

Ultimately, a subset of columns is selected to form the final dataset, which is the processed DataFrame that contains the columns needed for further analysis and modeling. These columns include demographic information, medical information, and the outcome variables created earlier. Overall, we perform a series of data cleaning, recoding, and feature engineering tasks to prepare the dataset for further analysis and machine learning applications.

Subsequently, we excluded the "race" feature from the dataset and divided it into a 70% training set and a 30% test set. The dataset split was conducted in a stratified manner to maintain the proportion of target classes or labels in both the training and testing datasets. However, due to the dataset's imbalances, we initially performed a resampling of the training data to generate a new, balanced training dataset. This approach proves especially valuable when dealing with imbalanced datasets, where one class is significantly more prevalent than others.

5.2.2 Model Evaluation Metrics

We report the performance of the models using 1) overall balanced error rate, 2) balanced error rate difference, 3) overall AUC, and 4) AUC difference. For the purpose of comparing model performance, we employ two key metrics: overall AUC and balanced error rate difference.

To ensure fairness, we inquire about which demographic groups may face a disproportionate and adverse impact. Past research indicates that individuals of varying racial and ethnic backgrounds may experience differential effects. When patients who would benefit from a care management program are not recommended for it, it leads to allocation-related issues. In the context of a classification scenario, these instances are referred to as False Negatives. Therefore, from the patient's perspective, the primary concerns in these situations are related to allocation, specifically the occurrence of false negatives, where someone who would benefit from the program is not recommended and may subsequently face readmission. For fairness, we report the following metrics: 1) false negative rate difference, 2) overall selection rate, 3) equalized odds difference, and 4) demographic parity difference. However, to to quantify harms and benefits and assess the fairness of the models, we use the false negative rate and the overall selection rate.

- **False Negative Rate:** This measures the proportion of patients who experience readmission within 30 days but were not advised to participate in the care management program, representing the extent of harm incurred.
- **Overall Selection Rate:** This calculates the overall fraction of patients who receive recommendations for the care management program, irrespective of whether they experience readmission within 30 days or not. This quantifies the program's overall benefit, assuming that all patients derive similar benefits from the additional care.

There are few reasons for including selection rate in addition to false negative rate. We would like to monitor how the benefits are allocated, focusing on groups that might be disadvantaged. The auxiliary metrics, like selection rate, may alert us to large disparities in how the benefit is allocated, and allow us to catch issues that we might have missed.

5.2.3 Model Comparison

We used GridSearch to systematically develop multiple models for each classifier, fine-tuning them to address the trade-off between performance (balanced accuracy) and fairness (equalized odds difference). We then selected the best model from the candidate models for each classifier. We applied this tailored approach to all seven classifiers that we employed. The results of this experiment are documented in Table 6, which provides a comprehensive overview of the diverse fairness and performance metrics evaluated for each model. A pivotal focus was placed on the false negative rate difference, a critical fairness metric for this use case. We undertook this intensive scrutiny to ascertain whether each classifier could deliver equitable predictions, particularly in the context of label assignment. By focusing on this key fairness metric, we committed to ensuring that our models operate with fairness, equity, and accuracy, upholding ethical standards.

Table 6. Fairness and performance metrics for diabetes dataset

	Metrics	Classifiers						
		ADB	DT	ET	GBM	LR	RF	SVM
Fairness	Overall selection rate	0.426709	0.509941	0.983196	0.408857	0.62308	0.440303	0.524354
	Demographic parity difference	0.115093	0.068996	0.019172	0.139724	0.026097	0.074173	0.030508
	False negative rate difference	0.00976	0.06434	0.029968	0.064615	0.087667	0.055333	0.040853
	Equalized odds difference	0.117274	0.066285	0.029968	0.139999	0.087667	0.074372	0.040853
Performanvce	Overall balanced error rate	0.431412	0.470987	0.497481	0.40468	0.423654	0.428655	0.418143
	Balanced error rate difference	0.054452	0.011714	0.010151	0.046504	0.03313	0.048705	0.006697
	Overall AUC	0.59481	0.529013	0.540736	0.629808	0.576346	0.600574	0.581857
	AUC difference	0.09093	0.011714	0.074487	0.05701	0.03313	0.067556	0.006697

After obtaining candidate models through grid search that meet the model constraints, we choose the best model that aligns with our evaluation metrics for the specific machine learning algorithm. Below Figure 6 provide the comparison of the seven different models based on their performance and fairness.

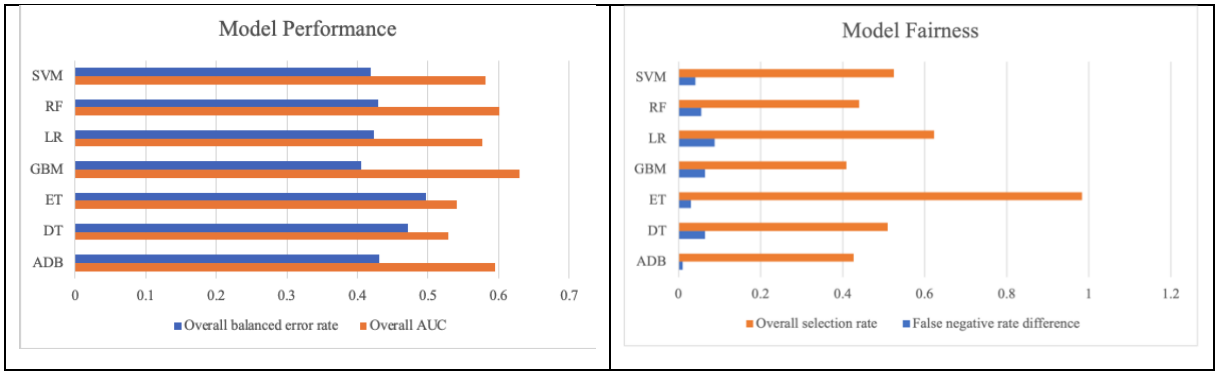

Figure 6.Comparison of seven models for diabetes dataset

The experimental findings reveal that the GBM classifier-based model attains the highest overall AUC value, lowest overall balanced error rate value, signifying superior performance in discriminatory ability. Consequently, we designate the GBM classifier-based model as the *best* in terms of performance. The ADB, SVM and RF classifier-based models provide the second-best performance closer to the GBM's and they are identified as having *similar* performance. The ADB, RF and SVM classifier-based models provide the second ranked performance, so we recognize them as *similar* to that of the GBM model's performance.

In the realm of fairness, ADB classifier-based model emerges as the frontrunner, giving the lowest false negative rate difference and an overall moderate selection rate, affirming the ADB classifier-based model as the fairest among the classifiers evaluated. Therefore, we recognize ADB as *best*. Notably, the GBM model slightly better than the performance metrics of the ADB model, however maintaining a higher false negative rate difference, indicative of its lacking fairness.

Subsequently, these results are seamlessly integrated into the extended Performance-Explainability-Fairness framework. The comprehensive summary of the extended framework outcomes for the seven classifiers is presented in Table 7, encapsulating the model characteristics and their respective evaluation across the dimensions of performance, explainability, and fairness. This presentation serves as a valuable reference point, explaining how the choice of ML model may diverge contingent upon the specific requirements and trade-offs necessitated by the use case, which must reconcile considerations of performance, explainability, and fairness.

Table 7. Summary of extended framework results for diabetes dataset

Evaluation Characteristics	**ADB**	**DT**	**ET**	**GBM**	**LR**	**RF**	**SVM**
Performance	Similar	Below	Below	Best	Below	Similar	Similar
Comprehensibility	Black-box	White-box	Black-box	Black-box	Black-box	Black-box	Black-box
Granularity	Global & Local	Global & Local	Global & Local	Global & Local	Global & Local	Global & Local	Global & Local
Information type	Feature importance	Feature importance	Feature importance	Feature importance	Feature importance	Feature importance	Feature importance
Faithfulness	Imperfect	Perfect	Imperfect	Imperfect	Imperfect	Imperfect	Imperfect
User category	Domain Expert	Domain Expert	Domain Expert	Domain Expert	Domain Expert	Domain Expert	Domain Expert
Fairness context	Group	Group	Group	Group	Group	Group	Group
Fairness	Best	Below	Similar	Below	Below	Similar	Similar

Regarding model comprehensibility, only the DT model qualifies as a "white-box" model, distinguished by its interpretability, whereas the remaining six classifiers, of which some are ensemble models, all fall under the category of "black-box" models.

In the realm of granularity, the DT model excels in providing explanations at both global and local level. Equally, the other classifiers, equipped with the SHAP method, demonstrate the capability to furnish both global and local level explanations. Furthermore, all classifiers yield feature importance information as part of their explanations.

In terms of faithfulness, the DT classifier-based model stands out as a model with perfect faithfulness, as explanations can be directly extracted from its original model. In contrast, the other models rely on post-hoc explanation methods employing surrogate models, rendering them imperfect in terms of faithfulness.

With respect to the fairness context, all classifiers are evaluated with respect to group fairness, with the ADB model exhibiting the highest level of fairness. The ET, RF and SVM classifier-based models closely approach the fairness metrics of the ADB model, and its fairness can be considered as *similar* to that of the ADB model. However, the other classifiers fall below these models in terms of fairness making them identified as *below*. These findings are visually illustrated in Figure 7.

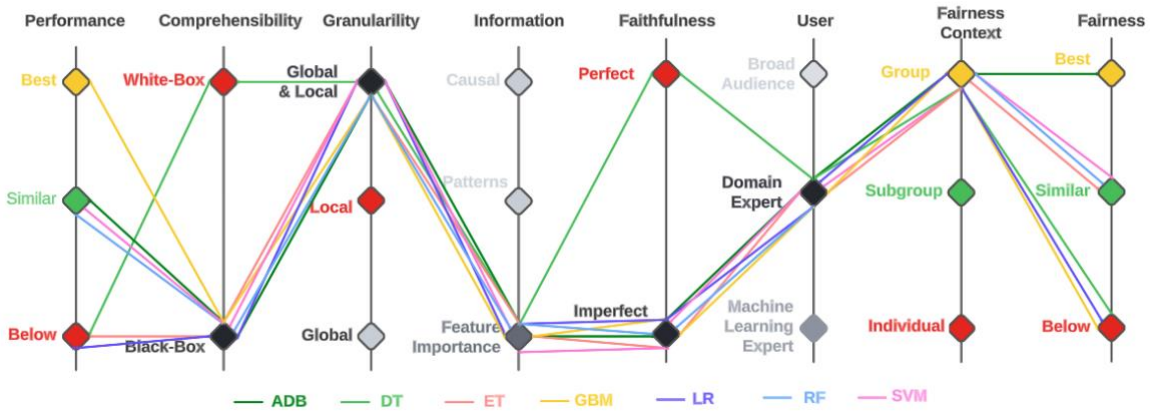

Figure 7. Parallel coordinates plot of the PEF framework result for diabetes dataset

In conclusion, considering the multifaceted dimensions of the performance-explainability-fairness framework applied to the diabetes dataset, the ADB classifier-based model emerges as the most suitable choice, excelling across various characteristics and striking a favorable balance between performance, explainability, and fairness.

5.3 Criminology Domain

ProPublica's COMPAS (Larson et al., 2016) recidivism dataset comprises 7,918 examples, making it a substantial and valuable resource for studying the complex issue of

recidivism prediction. In this dataset, the primary task involves predicting recidivism, which refers to whether an individual is likely to re-offend based on a comprehensive set of features. These features encompass various aspects of an individual's criminal history, such as prior offenses, jail and prison time, as well as demographic information. Additionally, the dataset includes COMPAS risk scores, which are crucial factors used in the prediction process. An important ethical consideration in this dataset is the inclusion of race as a protected attribute, highlighting the need to address potential fairness and bias concerns in predictive models applied to criminal justice scenarios. Analyzing and mitigating bias in recidivism prediction is essential to ensure equitable decision-making in the criminal justice system. The details of the dataset description are described in Table 17.

5.3.1 Data Preprocessing

We first ensure the data is cleaned and have no missing values, also ensure the categorical features are marked as dataframe "category" type. As indicated by Larson et al. (2016), in specific scenarios involving this dataset, challenges emerged when attempting to match charges with COMPAS scores. Consequently, it was deemed necessary to exclude such cases. Hence, we removed rows that were deemed irrelevant to the model, as certain instances might have involved alternative reasons for the charges. Ultimately, a subset of columns is selected to form the final dataset, which is the processed DataFrame that contains the columns needed for further analysis and modeling. These columns include demographic information, criminal record information, and the outcome variables. Overall, we perform a series of data cleaning, recoding, and feature engineering tasks to prepare the dataset for further analysis and machine learning applications. Subsequently, we excluded the "race" feature from the dataset and partitioned it into a 70% training set and a 30% testing set. We conducted the dataset split in a stratified fashion to maintain the proportion of target classes or labels in both the training and testing datasets. This approach proves especially valuable when working with imbalanced datasets, where one class significantly outnumbers the others.

5.3.2 Model Evaluation Metrics

We report the performance of the models using 1) Overall balanced error rate, 2) Balanced error rate difference, 3) Overall AUC, and 4) AUC difference. For the purpose of

comparing model performance, we employ two key metrics: Overall AUC and Balanced error rate difference. For fairness, we report 1) Overall selection rate, 2) Equalized odds difference, 3) Demographic parity difference. However, for assessing the fairness of the models, we utilize the Equalized odds difference and the Overall selection rate.

5.3.3 Model Comparison

In our model training process, we adopted a systematic approach by training multiple models for each classifier using GridSearch. Each of these models was fine-tuned to strike a balance between two critical metrics: performance (measured by balanced accuracy) and fairness (assessed through the equalized odds difference). This tailored approach was applied to all seven classifiers, resulting in a comprehensive experiment. The findings from this experiment are documented in Table 8, where we've provided an array of fairness and performance metrics for each of the models that we selected from GridSearch for each classifier.

Table 8. Fairness and performance metrics for COMPAS recidivism dataset

	Metrics	Classifier						
		ADB	DT	ET	GBM	LR	RF	SVM
Fairness	Overall selection rate	0.49622	0.5027	0.014039	0.026998	0.570734	0.580454	0.24514
	Equalized odds difference	0.139189	0.153846	0.076923	0.069767	0.107933	0.121795	0.192308
	Demographic parity difference	0.115551	0.167725	0.029412	0.031847	0.146032	0.174603	0.203175
Performance	Overall balanced error rate	0.116045	0.098041	0.488934	0.472521	0.106046	0.138916	0.328703
	Balanced error rate difference	0.063187	0.041667	0.03864	0.032181	0.053966	0.078139	0.119963
	Overall AUC	0.946433	0.910816	0.902573	0.975493	0.970057	0.956121	0.790762
	AUC difference	0.125	0.125	0.117216	0.125	0.25	0.25	0.22754

After obtaining candidate models through grid search that meet the model constraints, we choose the best model that aligns with our evaluation metrics for the specific machine learning algorithm.

Figure 8. Comparison of seven models for COMPAS recidivism dataset

The Figure 8 shows comparison of the seven different models based on their performance and fairness.

The experimental findings reveal that the GBM classifier-based model attains the highest overall AUC value, moderate overall balanced error rate value. The LR classifier-based model also obtain comparable AUC value and even lower overall balanced error rate value, signifying superior performance in discriminatory ability. Consequently, we designate the GBM as *best* and LR classifier-based model as *similar* in terms of performance.

In the realm of fairness, GBM and ET classifier-based models emerge as the frontrunner, giving the lowest equalized odds difference. These metrics collectively affirm the GBM model as the fairest and performant among the classifiers evaluated. Notably, the GBM and ET based models show a conservative selection rate which affirms that they have less potential of incurring harm. However, ET maintains a significantly lower overall AUC value, indicative of its lacking performance.

Subsequently, these results are seamlessly integrated into the extended Performance-Explainability-Fairness framework. The comprehensive summary of the extended framework outcomes for the seven classifiers is presented in Table 9, encapsulating the model characteristics and their respective performances across the dimensions of performance, explainability, and fairness. This presentation serves as a valuable reference, illustrating how the selection of a machine learning model may vary based on the specific requirements and trade-offs inherent in the use case. It is essential to strike a balance among performance, explainability, and fairness considerations to make informed model choices.

Table 9. Summary of extended framework results for COMPAS recidivism dataset

Evaluation Characteristics	**ADB**	**DT**	**ET**	**GBM**	**LR**	**RF**	**SVM**
Performance	Below	Below	Below	Best	Similar	Below	Below
Comprehensibility	Black-box	White-box	Black-box	Black-box	Black-box	Black-box	Black-box
Granularity	Global & Local	Global & Local	Global & Local	Global & Local	Global & Local	Global & Local	Global & Local
Information type	Feature Importance	Feature Importance	Feature Importance	Feature Importance	Feature Importance	Feature Importance	Feature Importance
Faithfulness	Imperfect	Perfect	Imperfect	Imperfect	Imperfect	Imperfect	Imperfect
User category	Domain Expert	Domain Expert	Domain Expert	Domain Expert	Domain Expert	Domain Expert	Domain Expert
Fairness context	Group	Group	Group	Group	Group	Group	Group
Fairness	Below	Below	Similar	Best	Below	Below	Below

Regarding model comprehensibility, only the DT model qualifies as a “white-box” model, distinguished by its interpretability, whereas the remaining six classifiers, of which some are ensemble models, all fall under the category of “black-box” models.

In the realm of granularity, the DT model excels in providing explanations at both global and local level. Equally, the other classifiers, equipped with the SHAP method, demonstrate the capability to furnish both global and local level explanations. Furthermore, all classifiers yield feature importance information as part of their explanations.

In terms of faithfulness, the DT model stands out as a model with perfect faithfulness, as explanations can be directly extracted from its original model. In contrast, the other models rely on post-hoc explanation methods employing surrogate models, rendering them imperfect in terms of faithfulness.

In terms of the fairness context, all classifiers are evaluated with respect to group fairness, with the GBM model exhibiting the highest level of fairness. The ET and LR models closely approach the fairness metrics of the GBM model, and ET's fairness can be considered as comparable to that of the GBM model. However, the other classifiers fall below the GBM models in terms of fairness. These findings are visually illustrated in Figure 9.

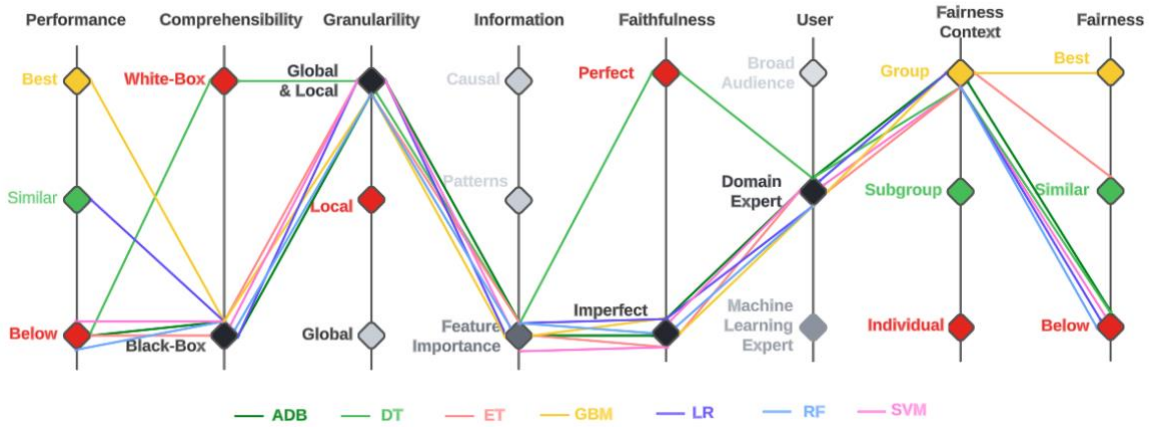

Figure 9. Parallel coordinates plot of the PEF framework result for recidivism dataset

In conclusion, considering the multifaceted dimensions of the performance-explainability-fairness framework applied to the COMPAS recidivism dataset, the GBM classifier-based model emerges as the most suitable choice, excelling across various characteristics and striking a favorable balance between performance, explainability, and fairness.

5.4 Education Domain

The Law School Admissions Council's National Longitudinal Bar Passage Study (Wightman, 1998), conducted in 1998, is a robust dataset comprising 20,649 examples. In this

dataset, the primary objective is to predict an individual's likelihood of passing the bar exam, a critical milestone in a lawyer's career. The dataset encompasses a diverse range of features that capture various aspects of a candidate's educational background, law school performance, and demographics. Of particular ethical importance is the inclusion of race as a protected attribute, emphasizing the significance of addressing fairness and potential bias in predictive models. Ensuring fairness in bar exam passage predictions is essential to promote equity and diversity in the legal profession. The details of the dataset description are described in Table 18.

5.4.1 Data Preprocessing

Our data preparation process begins with a meticulous cleaning phase, where we thoroughly scrutinize the dataset to ensure that it is free from any missing values. Additionally, we take care to categorize the relevant features within the dataframe as the "category" data type, aligning them for subsequent analysis. In our data preparation journey, we undertake a sequence of vital tasks, encompassing data cleaning, recoding, and feature engineering. These endeavors collectively serve to refine and optimize the dataset, making it well-suited for further analysis and the application of machine learning techniques. At this step in this process, we remove of the "race" feature from the dataset. Subsequently, we partition the dataset into two subsets: a 70% training set and a 30% test set. This division is performed with careful consideration, employing a stratified approach to maintain the integrity of class proportions or labels in both the training and testing datasets. This approach proves particularly beneficial when dealing with imbalanced datasets, where one class may significantly outnumber the others. By preserving this balance, we enable our models to learn and make predictions in a fair and representative manner, regardless of the class distribution.

5.4.2 Model Evaluation Metrics

We report the performance of the models using 1) Overall balanced error rate 2) Balanced error rate difference 3) Overall AUC and 4) AUC difference. For the purpose of comparing model performance, we employ two key metrics: Overall AUC and Balanced error rate difference. For fairness we report 1) Overall selection rate, 2) Equalized odds difference, and 3) Demographic parity difference. However, for assessing the fairness of the models, we utilize the Equalized odds difference and the Overall selection rate.

5.4.3 Model Comparison

To comprehensively train our models, we adopted a systematic approach by training multiple models for each classifier using GridSearch. Each of these models was fine-tuned to strike a balance between two critical metrics: performance (measured by balanced accuracy) and fairness (assessed through the equalized odds difference). We meticulously executed this calibration process for all seven classifiers we employed in our analysis. The results of this experiment are documented in Table 10, which provides a comprehensive overview of the various fairness and performance metrics that were assessed for each model. The metrics in this table reveal how each model performed in terms of both predictive performance and fairness. Of particular significance is the "equalized odds difference," a fundamental fairness metric that we closely scrutinized. We did this to ascertain whether each classifier could deliver equitable predictions, especially concerning label assignment. This thorough assessment aligns with our overarching objective to ensure that our models are not inadvertently biased or discriminatory in their predictions, safeguarding fairness, and ethical considerations.

Table 10. Fairness and performance metrics for admission dataset

	Metrics	Classifiers						
		ADB	DT	ET	GBM	LR	RF	SVM
Fairness	Overall selection rate	0.476744	0.515988	0.579215	0.433866	0.457849	0.069767	0.484738
	Equalized odds difference	0.149758	0.189369	0.156042	0.071318	0.140789	0.008555	0.083213
	Demographic parity difference	0.265349	0.304264	0.30867	0.108933	0.251458	0.046659	0.120831
Performance	Overall balanced error rate	0.235465	0.292151	0.255087	0.246366	0.22093	0.438953	0.227471
	Balanced error rate difference	0.03793	0.006312	0.001207	0.044155	0.068611	0.006585	0.045801
	Overall AUC	0.84459	0.707849	0.840172	0.753634	0.77907	0.81363	0.772529
	AUC difference	0.022369	0.006312	0.062378	0.044155	0.068611	0.019697	0.045801

After obtaining candidate models through grid search that meet the model constraints, we choose the best model that aligns with our evaluation metrics for the specific machine learning algorithm.

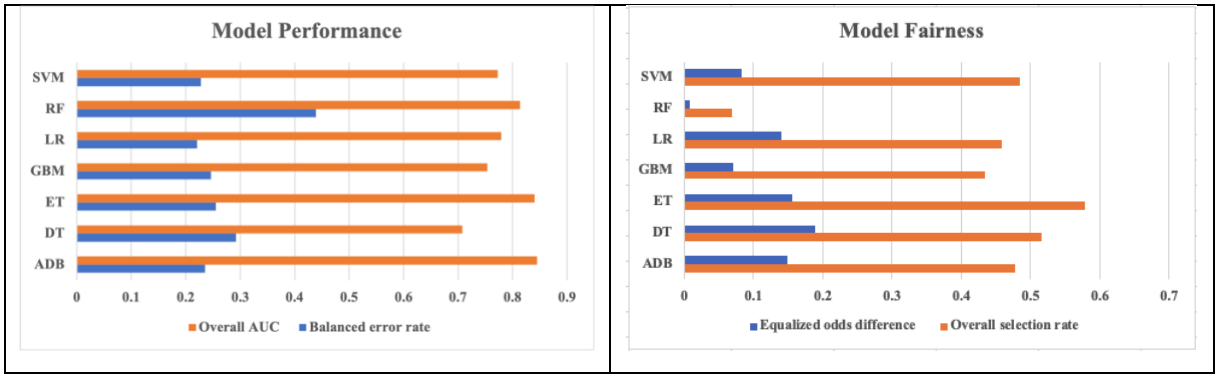

Figure 10. Comparison of seven models for admissions dataset

The Figure 10 shows the comparison of the seven different models based on the performance and fairness.

The experimental findings reveal that the ADB classifier-based model attain the highest overall AUC value, moderate overall balanced error rate value, signifying superior performance in discriminatory ability. For the ET and RF based models, while it's slightly lower in AUC value than the ADB, we consider ET and RF as providing *similar* performance that of ADB. Consequently, we designate the ADB classifier-based models as *best*, ET and RF classifier-based model as *similar* in terms of performance.

In the realm of fairness, RF classifier-based models emerge as the frontrunner, giving the lowest equalized odds difference. These metrics collectively affirm the RF model as the fairest and good enough performant among the classifiers evaluated.

Following this, the findings are smoothly incorporated into the expanded Performance-Explainability-Fairness framework. The all-encompassing summary of the extended framework's results for the seven classifiers is provided in Table 11, encapsulating both the model attributes and their corresponding performances across the dimensions of performance, explainability, and fairness. This presentation serves as a valuable reference, illustrating how the selection of a machine learning model may vary based on the specific needs and trade-offs inherent in the use case, which must carefully balance considerations of performance, explainability, and fairness.

Table 11. Summary of extended framework results for admissions dataset

Evaluation Characteristics	**ADB**	**DT**	**ET**	**GBM**	**LR**	**RF**	**SVM**
Performance	Best	Below	Similar	Below	Below	Similar	Below
Comprehensibility	Black-box	White-box	Black-box	Black-box	Black-box	Black-box	Black-box
Granularity	Global & Local	Global & Local	Global & Local	Global & Local	Global & Local	Global & Local	Global & Local
Information type	Feature Importance	Feature Importance	Feature Importance	Feature Importance	Feature Importance	Feature Importance	Feature Importance
Faithfulness	Imperfect	Perfect	Imperfect	Imperfect	Imperfect	Imperfect	Imperfect
User category	Domain Expert	Domain Expert	Domain Expert	Domain Expert	Domain Expert	Domain Expert	Domain Expert
Fairness context	Group	Group	Group	Group	Group	Group	Group
Fairness	Below	Below	Below	Below	Below	Best	Below

Regarding model comprehensibility, only the DT model qualifies as a "white-box" model, distinguished by its interpretability, whereas the remaining six classifiers, of which some are ensemble models, all fall under the category of "black-box" models.

In the realm of granularity, the DT model excels in providing explanations at both global and local level. Equally, the other classifiers, equipped with the SHAP method, demonstrate the

capability to furnish both global and local level explanations. Furthermore, all classifiers yield feature importance information as part of their explanations.

In terms of faithfulness, the DT model stands out as a model with perfect faithfulness, as explanations can be directly extracted from its original model. In contrast, the other models rely on post-hoc explanation methods employing surrogate models, rendering them imperfect in terms of faithfulness.

Regarding the fairness context, all classifiers are evaluated with respect to group fairness, with the RF classifier-based model exhibiting the highest level of fairness. Only the GBM classifier-based model somewhat approach the fairness metrics of the RF model, we still recognize as *below* fairness compared to that of RF's. However, the other classifiers fall below the RF model in terms of fairness. These findings are visually illustrated in Figure 11.

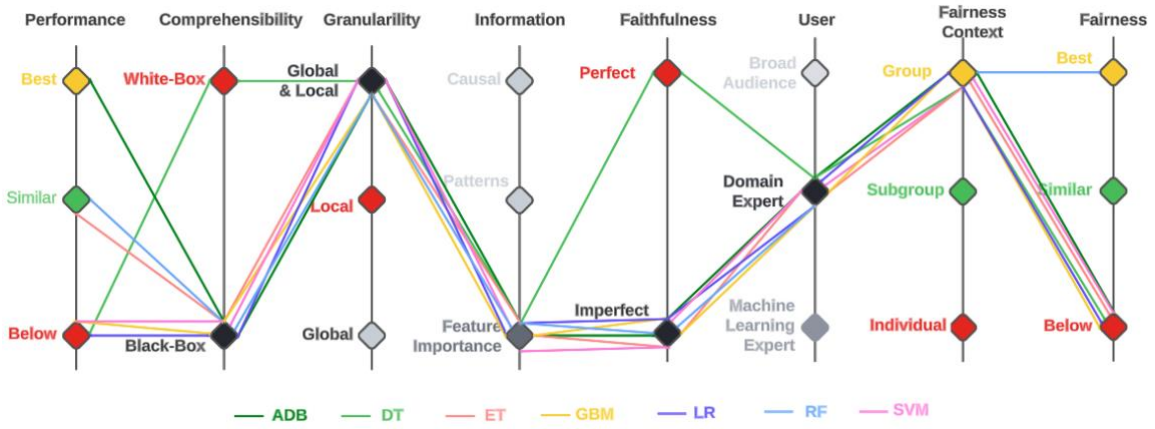

Figure 11. Parallel coordinates plot of the PEF framework for admissions dataset

In conclusion, considering the multifaceted dimensions of the performance-explainability-fairness framework applied to the law school admission dataset, the RF classifier-based model emerges as the most suitable choice, excelling across various characteristics and striking a favorable balance between performance, explainability, and fairness.

5.5 Result

The results of our evaluation are promising, indicating that the framework holds substantial potential in real-world applications. One of the notable achievements was our ability to discern and select the most suitable classifier for each dataset, guided by the principles of the PEF (Performance, Explainability, Fairness) framework. Importantly, it is crucial to highlight

that each binary classification model presented unique challenges and opportunities, necessitating customized performance and fairness criteria tailored to the specific sensitive attribute under consideration.

To comprehensively assess the framework's adaptability, we ventured into diverse scenarios, spanning different fairness constraints, sensitive attributes, and class ratios within the datasets. This wide-ranging exploration underscores the framework's adaptability and applicability in addressing the nuanced fairness concerns that can vary greatly from one context to another.

Table 12. Result of the application of the PEF framework on four domains dataset

Dataset	Domain	Sensitive attribute	Target class	Best Classifier
Credit card clients	Finance	Gender	Default payment	GBM
Diabetes	Healthcare	Race	Readmit in 30 days	ADB
COMPAS recidivism	Criminology	Race	Two-year recidivism	GBM
Law School Admission	Education	Race	Pass the bar exam	RF

The result of our efforts is encapsulated in the Table 12, which showcases the results for all four binary classification tasks. These results not only reflect the efficacy of the PEF framework in selecting the best models but also highlight the positive impact of considering fairness in decision-making processes. In essence, this evaluation underscores the framework's potential to facilitate equitable and effective decision-making across a spectrum of domains and applications.

In addition to our comprehensive framework development and evaluation, we aimed to assess its practical implications by conducting a survey involving eight participating machine learning practitioners. This survey was designed to assess the practical effectiveness and utility of the framework. Below, we provide an in-depth analysis of the survey results, shedding light on the perspectives and experiences of ML practitioners, and how the framework aligns with their needs and objectives in the ever-evolving field of machine learning.

5.5.1 Survey Results for the Framework Characteristics

In order to comprehensively evaluate the framework's effectiveness and user-friendliness, a set of survey questions was administered. These questions were carefully designed to assess participants' perceptions regarding the framework's clarity, relevance, and

potential areas for enhancement. In the subsequent analysis, we provide a concise summary of the survey results, offering valuable insights into how the framework aligns with user expectations and highlighting any opportunities for refinement and improvement.

The survey responses overwhelmingly indicate that the key characteristics of the framework, including performance, comprehensibility, granularity, information type, faithfulness, user category, fairness context, and fairness, are universally deemed relevant and clear by participants, with 100% agreement. This consensus underscores the framework's effectiveness in addressing these aspects. However, there is a notable minority opinion (12.5%) that suggests the potential inclusion of additional dimension currently missing from the framework, highlighting an opportunity for further refinement and expansion based on specific needs and perspectives.

5.5.2 Survey Results for Utility

The survey also aimed to gather valuable insights on participants' perspectives regarding the fairness benchmarking framework. Through using a five-point Likert scale (Strongly Agree, Agree, Undecided, Disagree, Strongly Disagree), respondents expressed the importance of fairness characteristics for their team's models and the utility of the performance-explainability-fairness benchmarking framework. Additionally, they provided feedback on the ease of following the benchmarking process and their inclination to integrate the framework into their projects. These responses offer a comprehensive understanding of the framework's practical significance, usability, and potential adoption in real-world machine learning projects.

Table 13. Survey results for the utility of the PEF framework

No	Quesiton answered on a rating scale	Answers				
		Strongly Agree (%)	Agree (%)	Undecided (%)	Disagree (%)	Strongly Disagree (%)
10	The fairness characteristics of the framework are important for the models my team develops and deploys.	62.5	25	12.5		
11	The performance-explainability-fairness model benchmarking framework is useful.	87.5	12.5			
12	The performance-explainability-fairness model benchmarking process is easy to follow.	62.5	25		12.5	
13	I am inclined to use this framework in my project.	37.5	50	12.5		

The survey results, shown in Table 13 provide valuable insights into participants' perceptions of the ML model benchmarking framework:

- **Fairness Characteristics Importance:** We find 62.5% of respondents strongly agree that the fairness characteristics of the framework are important for the models developed and deployed by their team. Additionally, 25% of respondents agree with this statement, signifying a noteworthy level of significance placed on fairness. However, 12.5% of respondents remain undecided on the matter.
- **Usefulness of Framework:** 87.5% of participants strongly agree that the performance-explainability-fairness model benchmarking framework is useful, while 12.5% agree. This suggests that all of respondents find the framework valuable for their work.
- **Ease of Use:** Similarly, 62.5% of respondents strongly agree that the performance-explainability-fairness model benchmarking process is easy to follow, while 25% agree. This indicates that most participants perceive the framework as user-friendly. However, 12.5% of respondents disagreed and indicated that there are room for improvements to make it easier to follow.
- **Inclination to Use:** Interestingly, 37.5% of participants strongly agree that they are inclined to use this framework in their projects, while 50% agree. This demonstrates a strong willingness among a majority of respondents to integrate the framework into their work. However, 12.5% of respondents remain undecided on the matter.

Overall, the survey results highlight a positive reception of the fairness benchmarking framework, with a notable emphasis on its usefulness and ease of use. While there is some variance in opinions, the majority of respondents express a clear interest in adopting the framework for their projects, emphasizing the potential value it offers in promoting fairness and transparency in machine learning models.

5.5.3 Survey Results for the Strength and Weakness

These survey questions were designed to collect valuable feedback and insights from participants regarding the assessed framework. The first question aimed to identify the strengths and positive aspects of the framework, providing insight into its most effective features. The second question sought suggestions for improvement, allowing participants to contribute ideas and recommendations for enhancing the framework's usability and effectiveness. Collecting this feedback enables the refinement and optimization of the framework based on real-world user perspectives, fostering continuous improvement and relevance.

Table 14. Survey results for the strength and weakness of the PEF framework

No	Open ended question	Answer
15	What are the strengths of the framework?	1. The emphasis on fairness. 2. The multiple characteristics with which the ML models can be evaluated to identify the best fit for the use case. 3. Adding fairness as a new dimension to the existing framework. 4. The inclusion of the 3 dimensions to evaluate the ML model. 5. It takes care of an important aspect of ML model development which is fair ML development. 6. As stated by the author, the previous framework was missing fairness as an aspect of machine learning models. We have found that fairness needs to be built into ML models going forward. This new framework may provide developers a system for including fairness in future ML models 7. Inclusion of Fairness as a new dimension in the evaluation criteria. 8. Adding the additional classifiers adds much needed fairness characteristics.
16	How can the proposed framework be impr	1. Maybe look at if explainability can be expanded. 2. It looks good so far. 3. None on the framework, but I wonder if tools can be developed to help use the framework. 4. I would recommend to add a better visual representation of the original vs additional characteristics of the previous and new framework. 5. N/A 6. None

The respondents, answers summary shown in Table 14, highlighted several strengths of the framework in their open-ended responses:

- **Inclusion of Three Dimensions:** Participants appreciated that the framework encompasses three dimensions to evaluate machine learning models, emphasizing the comprehensive nature of the assessment.
- **Addressing Fair ML Development:** The framework was recognized for addressing a crucial aspect of machine learning model development, specifically fairness, which is seen as increasingly important in the field.
- **Filling a Gap:** Respondents noted that the framework addresses a gap in previous approaches by incorporating fairness as a fundamental aspect of model evaluation. They view this as vital for future machine learning model development.
- **Innovative Dimension:** The inclusion of fairness as a new dimension in the evaluation criteria was seen as a notable and innovative feature of the framework.

The open-ended responses regarding improvements to the proposed framework were generally positive and succinct:

- **Expanding Explainability:** One respondent suggested exploring the possibility of expanding the framework's focus on explainability, indicating an interest in further enhancing this aspect.
- **Positive Evaluation:** Another participant expressed satisfaction with the framework as it stands, suggesting that no major improvements are immediately apparent.
- **Supportive Tools:** One respondent suggested exploring the development of supportive tools to facilitate the practical use of the framework.
- **Visual Representation:** Another respondent recommended that the framework's visualization be improved to clearly distinguish between extended characteristics.
- **No Specific Suggestions:** Some respondents did not provide specific improvement suggestions, indicating that they found the framework to be comprehensive and satisfactory as presented.
- **No Needed Changes:** A few participants explicitly stated that they believe no changes or improvements are needed, indicating a high level of satisfaction with the framework's current design.

5.6 Chapter Summary

In our thorough evaluation of the PEF framework, we used four different datasets from various fields to showcase its adaptability. We closely examined the framework's ability to balance the trade-offs between performance, explainability, and fairness, emphasizing its versatility. For each dataset, our quest for the best machine learning model involved a thorough analysis of different approaches. After exploring various avenues, we identified a model that successfully met all the predefined criteria. This comprehensive evaluation highlights the effectiveness of the PEF framework in tackling real-world challenges across diverse domains. It also emphasizes the framework's capacity to maintain a balance between model performance, transparency, and equitable results.

CHAPTER 6

DISCUSSION

6.1 Introduction

This dissertation comprehensively explores ML model benchmarking, emphasizing performance, explainability, and fairness. It aims to create a unified framework addressing these dimensions for responsible AI deployment. The research focuses on two primary objectives.

First, we have developed and rigorously assessed the Performance-Explainability-Fairness framework, which serves as the cornerstone of this dissertation. This framework aims to provide a systematic approach for benchmarking ML models, enabling us to evaluate and compare their performance, explainability, and fairness comprehensively. Through a series of empirical investigations, we have uncovered valuable insights into the performance and fairness dynamics of various ML models, shedding light on their strengths and limitations.

Second, we have demonstrated the practical utility of the PEF framework in real-world applications, with a specific focus on the prediction of default credit card payments, along with three more datasets across a total of four different domains. By applying the framework to these use cases, we have highlighted its effectiveness in guiding model selection, thereby contributing to the responsible implementation of ML algorithms in critical decision-making scenarios.

As we delve into the discussion chapter, we will navigate the intricacies of our research findings and their implications. This chapter is structured to facilitate an in-depth examination of the framework's key components, including performance metrics, explainability levels, and fairness assessments. Additionally, we will explore the framework's applicability to various ML models and the challenges posed by real-world data biases.

6.2 Performance Evaluation

Here we discuss the performance dimension of the PEF framework from a few different perspectives.

Results of Performance Evaluations using the PEF Framework: The application of the Performance-Explainability-Fairness framework has yielded valuable insights into the

performance of various machine learning models. These evaluations encompassed an extensive analysis of predictive accuracy using AUC score and balanced error rates, providing insights into their robustness, and generalization capabilities, offering ample view of model performance.

Analysis of Different ML Models' Performance: The benchmarking process involved the systematic assessment of seven distinct ML algorithms for the respective dataset: 1) Adaptive Boosting (ADB), 2) Decision Tree (DT), 3) Extremely Random Trees (ET), 4) Gradient Boosting Machines (GBM), 5) Logistic Regression (LR), 6) Random Forest (RF), and 7) Support Vector Machine (SVM). Each resulting model underwent rigorous evaluation, enabling a comparative analysis of their respective performances.

Highlighting Significant Findings, Trends, and Patterns: The performance evaluations uncovered several noteworthy findings and discernible patterns. Notably, in case of the default credit dataset, the GBM model demonstrated the highest overall AUC, closely followed by ET, ADB and RF, while DT, SVM and LR exhibited comparatively lower performance levels. These results unveil variations in model performance, offering valuable guidance for model selection in specific applications.

Therefore, the performance evaluations conducted within the PEF framework have provided in-depth insights into the strengths and weaknesses of each ML model. These assessments contribute to a more nuanced understanding of model performance, aiding practitioners in making informed decisions regarding model deployment in real-world contexts.

6.3 Explainability Assessment

In this section, we discuss the results of the explainability assessments conducted within the Performance-Explainability-Fairness framework, aiming to shed light on the understandability of machine learning models and their implications on model trust and transparency. We also compare the explainability of models with varying levels of performance.

Presentation of Explainability Assessments within the PEF Framework: To evaluate the explainability of the machine learning models, we employed a range of techniques, including feature importance analysis, explanation methods (such as SHAP). These assessments were carried out on each model in our benchmarking dataset. Explainability features are represented by the comprehensibility, granularity, information type, faithfulness, and user category

characteristics. These characteristics provide us the option for model selection trade-off based on the explainability requirements for a specific use case.

Extent of Understandable Explanations: The results of our explainability assessments indicated varying levels of understandability across different models. While some models offered clear and interpretable explanations for their predictions such as DT classifier-based models for each of the dataset, others provided explanations that were less transparent and harder to comprehend such as other ensemble-based models with classifiers like ET, GBM, ADT etc. This discrepancy in the extent of understandability highlights the importance of incorporating explainability into the model development process.

Implications of Explainability on Model Trust and Transparency: Explainability plays a pivotal role in enhancing model trust and transparency. Models that could provide more understandable explanations are generally associated with higher levels of trust among users and stakeholders. When a model's decision-making process is transparent and comprehensible, users are more likely to trust its predictions and recommendations. This has significant implications, especially in high-stakes applications like healthcare and finance, where decision-making processes need to be justifiable and transparent.

For example, in our finance domain use case with default credit dataset, considering the inherent complexity of our decision problem, our objective is to optimize both performance and fairness, with a degree of flexibility regarding explainability whenever possible. However, in scenarios where absolute faithfulness is paramount and performance expectations can be more lenient, the Decision Tree (DT) emerges as the sole suitable choice. Conversely, when performance and fairness share equal importance, Gradient Boosting Machine (GBM) stands out as the preferred option. These observations underscore the importance of assessing use cases and aligning model selection with specific metric priorities in a balanced and nuanced manner.

Moreover, explainability can aid in identifying potential biases or ethical concerns within the model. Models that offer clear explanations for their predictions allow practitioners to pinpoint the source of bias or discrimination and take corrective actions, thereby promoting fairness in AI systems.

Comparison of Explainability Across Models with Different Performance Levels: One of the intriguing findings from our study is the correlation between model performance and explainability. We observed that models with higher performance metrics often tended to have

more complex architectures and were consequently less interpretable. On the other hand, models with lower performance metrics generally had simpler structures, making their explanations more understandable such as DT based models. This trade-off between performance and explainability underscores the need for a balanced approach when designing machine learning systems. While high-performance models are desirable for many applications, their lack of explainability can limit their real-world adoption, especially in domains where interpretability is crucial.

Therefore, the explainability assessments conducted within the PEF framework highlighted the importance of transparent and understandable machine learning models. These assessments revealed variations in the extent of explainability among models, with implications on trust, transparency, and fairness. Furthermore, our findings emphasized the trade-off between model performance and explainability, urging researchers and practitioners to strike a balance that aligns with the specific requirements of their applications.

6.4 Fairness Evaluation

In this section, we focus on the fairness evaluation aspect of the Performance-Explainability-Fairness Framework. We assess the fairness of ML models using the defined fairness metrics within the PEF framework, explore disparities and biases in model predictions, analyze the implications of fairness considerations on model selection and deployment, and discuss strategies for enhancing fairness in ML models.

Evaluation of Fairness Using PEF Framework Metrics: Within the PEF framework, fairness is evaluated using use case appropriate fairness metrics that may encompass different dimensions of fairness, including demographic parity, equalized odds etc. These metrics enable a quantitative assessment of how ML models treat different groups within a dataset or user population. Our fairness evaluation revealed variations in model performance across various fairness metrics. Some models demonstrated a higher degree of fairness by consistently delivering equitable outcomes for different demographic groups, while others exhibited disparities that may have real-world consequences. For example, with the finance domain default credit dataset, GBM provided far superior fairness through equalized odds difference than the ER, LR, DT and other classifier-based models. However, it’s crucial to emphasize that the fairness assessment’s outcomes can be context-dependent, as different fairness metrics may

favor distinct models. In this context, the SVM classifier-based model, when assessed through demographic parity difference, emerged as the most suitable option. These fairness metrics provide a robust and comprehensive basis for evaluating fairness in machine learning, enabling a more informed and ethical deployment of models in diverse real-world applications. However, the framework allows for the flexibility to utilize any fairness metrics that are suitable for the specific use case, without mandating a particular one.

Exploration of Disparities and Biases: The exploration of disparities and biases in model predictions revealed important insights. Disparities were often observed in situations where models had insufficient representation of certain groups in the training data. This led to underperformance for those groups, indicating the presence of bias. Additionally, bias in data labeling, either due to historical imbalances or societal prejudices, could manifest as biases in model predictions. Biases and disparities can have significant ethical and societal implications. In domains like criminal justice or lending, biased predictions can perpetuate discrimination and exacerbate societal inequalities. Therefore, it is crucial to identify and rectify these issues during the model evaluation phase.

Impact of Fairness Considerations on Model Selection and Deployment: The inclusion of fairness considerations in our benchmarking process had a substantial impact on model selection and deployment decisions. Models that demonstrated fairness across a range of metrics were prioritized, especially in sensitive domains where equitable outcomes are paramount. However, this sometimes came at the expense of pure performance metrics, highlighting the trade-off between performance and fairness. For example, in case of the law school admission dataset in the education domain, even though ADB and ET based models had better performance than RF, we choose RF based model as the best model for the dataset given it had a far superior fairness metric value. Additionally, with the criminology domain using COMPAS recidivism dataset, we had both GBM and LR emerged as best performant models, but LR had inferior fairness than GBM, so we selected GBM as the best suitable model for the use case. Additionally, in some cases, the fairness evaluation led to model fine-tuning or retraining to mitigate bias and improve fairness. This iterative process underscored the importance of continuous monitoring and improvement in ensuring fair AI systems.

Strategies for Improving Fairness in ML Models: Our research highlights several strategies for enhancing fairness in ML models.

- **Data Augmentation:** Augmenting underrepresented data can help mitigate biases caused by skewed training data. By creating additional training examples, effectively expanding the size of the dataset, a larger dataset can help improve the generalization and robustness of a machine learning model.
- **Algorithmic Fairness Techniques:** Implementing algorithmic fairness techniques, such as re-weighting or re-sampling, can rectify disparities in predictions. For example, for the health care domain with the diabetic dataset, we used resampling techniques.
- **Bias Detection and Mitigation:** Regularly assessing and mitigating bias during model development and deployment is crucial. This may involve identifying problematic features, data or decision-making processes and adjusting them to reduce bias. For example, for the criminology domain with COMPAS recidivism dataset, we remove several features and data that are not useful for the model as certain cases may have had alternative reasons for being charged.
- **Fairness-Aware Hyper-parameter Tuning:** Incorporating fairness-aware hyper-parameter tuning techniques during model training can promote equitable outcomes. For example, we used Fairlearn's GridSearch as a hyperparameter tuning technique that helps find the best combination of hyperparameters for a given model. It focuses on optimizing model performance with respect to fairness constraints. It searches through a range of hyperparameters to identify the best configuration that achieves a balance between performance and fairness.
- **Explainable AI:** Utilizing explainable AI techniques can help pinpoint the sources of bias in model predictions, enabling more targeted fairness interventions. For example, by integrating such techniques into our framework, we empower practitioners to not only identify fairness-related challenges but also to understand why and how these challenges arise. This enhanced transparency and interpretability foster more effective strategies for mitigating bias and promoting fairness in machine learning models across various domains, including finance.

Therefore, the fairness evaluation within the PEF framework provides a structured approach to assess and address biases and disparities in ML models. It emphasizes the significance of fairness in model selection and deployment, especially in sensitive domains, and

offers strategies for improving fairness, contributing to the development of more equitable AI systems.

6.5 Integration of Performance, Explainability, and Fairness

In this section, we focus on the pivotal aspect of integrating performance, explainability, and fairness within the framework. As AI and ML continue to shape diverse domains, this integration becomes paramount. We explore the synergies and challenges of harmonizing these dimensions, ultimately forging a path toward responsible and equitable AI and ML deployment.

Interplay between Performance, Explainability, and Fairness: The inclusion of performance, explainability, and fairness within the benchmarking framework marks a significant transformation in the assessment of machine learning models. These dimensions are inextricably linked, with each influencing and shaping the others. The study has illuminated the intricate interplay between these aspects, revealing that enhancing one dimension often entails trade-offs or synergies with the others. As an illustration, when dealing with a healthcare dataset, we encountered a trade-off situation. We had to carefully weigh our options and opted for the ADB classifier-based model, which demonstrated superior fairness considerations, even though the GBM classifier-based model delivered the best performance.

Impact of Improvements in One Dimension: Improvements in one dimension, such as fairness, have a ripple effect across the entire framework. For instance, when efforts are made to enhance fairness, model performance may experience variations. Achieving fairness may require the introduction of constraints or modifications that influence the model's predictive accuracy. Therefore, it is imperative to strike a balance between these dimensions, ensuring that enhancements in one do not come at the expense of the others.

Case Studies and Illustrative Examples: The research incorporates case studies and examples that vividly illustrate the trade-offs and synergies between these dimensions. For instance, when assessing fairness in the context of the "default of credit card client" dataset, it became evident that optimizing fairness metrics led to variations in model performance. Conversely, prioritizing predictive performance often came at the cost of fairness. These real-world instances underscore the practical challenges faced by practitioners and highlight the need for a holistic evaluation framework.

Therefore, the integration of performance, explainability, and fairness within the benchmarking framework marks an important advancement in the field of ML model assessment. This discussion underscores the intricate dynamics between these dimensions, emphasizing the need for a balanced approach that optimizes each while considering their inherent trade-offs. Real-world case studies serve as practical demonstrations of these principles, shedding light on the complex decision-making processes when deploying ML models in contexts where fairness, explainability, and performance is paramount.

6.6 Comparison with Existing Benchmarking Approaches

In this section, we compare the Performance-Explainability-Fairness framework with existing benchmarking methodologies, assessing its comprehensiveness and effectiveness in addressing the challenges of evaluating machine learning models.

Comprehensiveness of PEF Framework: The PEF framework offers a holistic approach to benchmarking ML models that sets it apart from many existing methodologies. Traditional benchmarking often focuses solely on performance metrics such as accuracy, precision, or recall. While these metrics are undeniably essential, they provide an incomplete view of a model's suitability for real-world deployment. In contrast, the PEF framework integrates not only performance metrics but also explainability and fairness assessments. This comprehensive approach ensures that a broader set of factors is considered when evaluating a model's fitness for a particular task or application. This is particularly relevant in fields like healthcare and finance, where ethical and transparency considerations are paramount.

Effectiveness of PEF Framework: The effectiveness of the PEF framework is evident in its ability to provide a more nuanced evaluation of ML models. By including explainability and fairness assessments, it addresses the limitations of traditional benchmarking, which often fails to capture the interpretability and equity aspects of models. This makes the PEF framework more adaptable to a wide range of applications and contexts, where model explanations and fairness considerations are becoming increasingly important. Moreover, by considering the trade-offs between performance, explainability, and fairness, the PEF framework helps stakeholders make more informed decisions about model deployment.

Advantages of the PEF Framework:

- **Holistic Evaluation:** The PEF framework offers a well-rounded evaluation of ML models, considering not only predictive performance but also their ability to provide interpretable explanations and ensure fairness. This holistic perspective aligns better with the real-world demands of AI systems.
- **Bias Mitigation:** The incorporation of fairness assessments within the PEF framework enables the identification and mitigation of bias and discrimination in ML models. This is essential for ensuring equitable outcomes in various domains.
- **Transparency and Trust:** By evaluating explainability alongside performance, the PEF framework enhances model transparency and trustworthiness, addressing concerns regarding the "black box" nature of complex models.

Limitations of the PEF Framework:

- **Complexity:** The PEF framework introduces complexity into the benchmarking process, as it requires expertise in both performance evaluation and explainability/fairness assessments. This may deter some practitioners, especially those with limited resources or expertise.
- **Resource Intensive:** Conducting thorough explainability and fairness assessments can be resource-intensive, particularly for large datasets and complex models. This may limit the scalability of the PEF framework in certain scenarios.
- **Subjectivity:** Assessing model explanations and fairness can involve subjective judgments, potentially introducing bias. Standardizing these assessments remains a challenge.

In conclusion, the PEF framework offers a comprehensive and effective approach to benchmarking ML models, addressing the limitations of existing methodologies. Its advantages include a more holistic evaluation, bias mitigation, and improved transparency and trust. However, its complexity, resource requirements, and potential subjectivity in assessments should be carefully considered when applying the framework in practice.

6.7 Implications and Applications

In this section, we delve into the practical implications of the research findings. We examine the influence of these findings on the wider machine learning field, investigate

potential real-world uses of the PEF framework, and contemplate its significance in decision-making, especially within sensitive domains.

Practical Implications for the Broader Field of ML: The research findings presented in this dissertation have several practical implications for the broader field of machine learning:

- **Enhanced Model Development:** The integration of explainability and fairness assessments into benchmarking offers valuable insights for model development. Researchers and practitioners can use these assessments to identify areas for improvement in their models, leading to more robust and ethically sound AI systems.
- **Improved Model Trust:** The dissertation emphasizes the importance of model transparency and trustworthiness. As AI applications become more widespread, building trust among users and stakeholders is paramount. By adopting the PEF framework, organizations and developers can ensure their models are not only accurate but also understandable and fair, enhancing user confidence.
- **Ethical Considerations:** The PEF framework highlights the ethical considerations that must be addressed in machine learning. As AI systems impact various aspects of society, including healthcare, finance, and criminal justice, the framework provides guidance on how to evaluate and mitigate bias, thereby promoting fairness and equity.

Potential Applications of the PEF Framework in Real-World Scenarios: The PEF framework has significant potential for various real-world applications:

- **Healthcare:** In the healthcare domain, the PEF framework can be employed to assess diagnostic models. Models that not only exhibit accuracy but also offer understandable explanations for their predictions can aid healthcare professionals in making well-informed decisions, thereby enhancing the quality of patient care.
- **Finance:** In the financial sector, where transparency and fairness are crucial, the PEF framework can be used to evaluate risk assessment models. Ensuring these models are both accurate and fair can help prevent discriminatory lending practices and promote financial equity.
- **Criminal Justice:** When evaluating predictive criminal justice models, the PEF framework can aid in identifying and mitigating biases, reducing the potential for unfair or discriminatory law enforcement practices.

Relevance of the Framework in Sensitive Decision-Making Processes: The PEF framework's relevance in decision-making processes, especially in sensitive domains, cannot be overstated. In these contexts, decisions often have far-reaching consequences, and the framework can help ensuring the following:

- **Ensure Accountability:** By requiring models to provide explanations for their predictions and assessments of fairness, the PEF framework holds AI systems accountable for their decisions, making it clear who is responsible in case of errors or ethical violations.
- **Facilitate Compliance:** In sectors subject to regulatory oversight, such as healthcare and finance, the framework can assist organizations in complying with regulations that mandate fairness and transparency in AI systems.
- **Promote Ethical Decision-Making:** The PEF framework encourages organizations and developers to make ethical considerations an integral part of their decision-making processes. This is particularly important in domains where fairness and equity are paramount, such as criminal justice and healthcare.

In summary, the PEF framework's practical implications span the broader field of machine learning, offering valuable insights for model development, trust-building, and ethical considerations. Its potential real-world applications extend to various domains where fairness, transparency, and accountability are essential, with particular relevance in sensitive decision-making processes.

6.8 Chapter Summary

The key takeaways from this chapter encompass several vital facets. Our exploration of the PEF framework reveals that it offers a multifaceted approach to ML model evaluation. By encompassing performance, explainability, and fairness assessments, it provides a nuanced understanding of model behavior. Explainability and fairness assessments illuminate the inner workings of models, enabling transparency, trust, and bias identification. The PEF framework contributes to advancing the field of ML model benchmarking by bridging the gap between model performance and ethical considerations, promoting responsible AI development. This holistic approach addresses the shortcomings of traditional benchmarking methods by emphasizing the ethical implications of AI systems.

CHAPTER 7

CONCLUSIONS

This concluding chapter serves as a comprehensive summary of the research endeavor, encompassing practical implications, theoretical, practical, and methodological contributions, inherent limitations, and prospective avenues for future exploration. It provides a holistic perspective on the multifaceted dimensions of the study's impact. The practical implications underscore the real-world significance of the research findings, demonstrating how they can be applied in various domains to address pertinent issues. The theoretical and methodological contributions delineate the novel insights and methodologies devised during the course of this research, augmenting the existing body of knowledge. However, it is vital to acknowledge the limitations of the study, which guide future research endeavors toward addressing these constraints. Consequently, this chapter concludes by delineating the trajectory for future research directions, charting a course for continued scholarly exploration and innovation in the field.

7.1 Impact of the Artifact

The Performance-Explainability-Fairness framework stands as an innovative framework within the domain of benchmarking machine learning models, ushering in an evolution towards responsible and equitable AI implementations. Its impact is felt across various facets, reshaping the landscape of assessing and utilizing ML models. This innovative framework not only addresses the crucial need for enhanced model evaluation but also underscores the imperative of fairness and transparency in the deployment of AI systems.

At its core, the PEF framework serves as a robust solution to the escalating complexities arising from the widespread adoption of AI and ML systems in our modern technology landscape. These systems, renowned for their predictive capabilities, frequently grapple with issues of opacity, inscrutability, and unfairness. The PEF framework, incorporating the pivotal aspects of performance, explainability, and fairness, emerges as a guiding tool that directs both practitioners and scholars towards models that not only excel in predictive precision but also instill confidence and uphold equitable outcomes. In doing so, it addresses the critical need for

responsible and accountable AI deployment, thereby participating significantly to the evolving discourse surrounding the ethics and efficacy of AI system.

Moreover, the PEF framework's influence extends beyond the technical realm into the broader socio-economic landscape. Organizations adopting the framework are better positioned to mitigate risks associated with biased decisions, thereby safeguarding their reputation and credibility. By fostering transparency and accountability, the framework help generate public trust, which is pivotal in promoting the widespread acceptance and adoption of AI systems.

In essence, the PEF framework serves as a compass guiding the AI community toward models that transcend traditional performance metrics. Its impact is manifest in the emergence of AI systems that not only excel in predictive accuracy but also operate within the bounds of ethics and fairness. As we navigate an increasingly AI-driven world, the PEF framework stands as a testament to the commitment of researchers and practitioners to harness the power of AI for the greater good, ushering in an era where technology aligns seamlessly with human values and aspirations.

7.2 Contributions

This research stands as an original contribution in the ever-evolving landscape of explainable AI and ML. This research integrates the crucial dimensions of performance, explainability, and fairness, by addressing the multifaceted challenges associated with ML model assessment, it provides a comprehensive blueprint for evaluating ML models that not only excel in predictive accuracy but also operate transparently and equitably.

This dissertation's contributions extend beyond academia. It presents a practical framework applicable to various domains where fairness, transparency, and trust is paramount. By prioritizing ethical AI development, this research has the potential to shape the broader research community's practices, foster responsible AI deployment, and contribute to a more equitable and trustworthy AI-powered future. Below we discuss the contribution of this research from the research and practical utility perspective.

7.2.1 Contributions to Research

The research makes a profound contribution to the realm of AI and machine learning research. This innovative framework redefines the paradigm of ML model assessment by

harmoniously integrating three pivotal dimensions: performance, explainability, and fairness. In doing so, it not only advances our understanding of how to evaluate and select ML models effectively but also empowers practitioners and researchers to build AI systems that are not only highly accurate but also transparent and ethically sound. This contribution is particularly timely in an age where the ethical implications of AI are under intense scrutiny, as it equips the AI community with a robust methodology for addressing biases, enhancing trust, and ensuring responsible AI deployments. By pioneering this holistic approach, the research lays the foundation for a future where AI technologies align seamlessly with human values, promoting fairness, accountability, and societal well-being in an AI-driven world.

One of the most significant contributions of the PEF framework is its emphasis on fairness. In a world increasingly aware of the ethical implications of AI, fairness considerations are paramount. By evaluating ML models for fairness, the framework aids in uncovering biases, disparities, and inequities within the model's predictions. It enables organizations to rectify these issues and ensures that AI-driven decisions are just and equitable across diverse demographic groups. The impact here is not only societal but also legal, as regulatory bodies are increasingly requiring fairness assessments in AI deployments. Following are some key contributions:

- **Comprehensive Benchmarking:** The framework establishes a comprehensive approach to benchmarking ML models, incorporating not only performance metrics but also explainability and fairness aspects. This holistic assessment aids in identifying models that excel in various dimensions, promoting well-informed model selection.
- **Methodological Advancement:** The framework advances the methodology for evaluating ML models by incorporating multiple dimensions, thus setting a precedent for more comprehensive and robust benchmarking practices.
- **Interdisciplinary Bridge:** The framework bridges the interdisciplinary gap between AI, ethics, and fairness, fostering collaboration and dialogue among researchers from various fields.
- **Future Research Opportunities:** It identifies avenues for future research, particularly in the areas of data bias mitigation, the development of fairness-aware algorithms, and the refinement of benchmarking methodologies which are outlined in the later section of this chapter.

In sum, the PEF Framework extends the boundaries of ML model evaluation, offering a holistic and ethical approach that advances research in the field of AI.

7.2.1 Contributions to Practice

This research's significance extends far beyond academia, as it equips industries, organizations, and policymakers with the tools and methodologies needed to navigate the ethical and practical complexities of AI deployments. In a contemporary landscape marked by escalating dependence on AI and a heightened recognition of its profound societal consequences, this research emerges as a guiding force. It directs our trajectory towards a future in which technology harmoniously aligns with fundamental human values, thereby nurturing trust, fostering accountability, and upholding principles of justice in the AI-driven world. Following are the contribution to practice of this research:

- **Ethical AI Deployment:** By emphasizing fairness as a crucial component of benchmarking, the framework contributes to the responsible and ethical deployment of AI systems. It guides practitioners in evaluating models for potential bias and ensuring equitable outcomes.
- **Transparency and Explainability:** The inclusion of explainability metrics encourages the development of AI models that are more transparent and interpretable. This enhances user trust and facilitates model understanding, which is particularly vital in critical applications such as healthcare and finance.
- **Practical Applicability:** The framework's practical utility is demonstrated through a real-world case study, making it accessible and relevant for practitioners and researchers seeking to assess ML models across domains.
- **Guidance for Model Selection:** It offers guidance on model selection based on specific priorities, whether performance-centric, explainability-focused, or fairness-driven, enabling stakeholders to align model choices with their particular requirements.
- **Adaptability and Extensibility:** The framework's adaptability allows for customization to suit diverse use cases and domains, promoting its applicability in a wide range of AI applications.

- **Awareness of Biases:** By acknowledging the limitation of data bias, the framework raises awareness about potential biases in real-world datasets, prompting researchers and practitioners to address this critical issue.

In sum, the PEF Framework extends the boundaries of ML model evaluation, offering a holistic and ethical approach that advances practical utility in the field of AI.

7.3 Limitations

While the PEF Framework proposed in the dissertation offers a comprehensive approach for benchmarking machine learning models, it is essential to acknowledge its limitations as well to provide a balanced perspective. Some of the limitations of this framework include:

- **Complexity and Computational Resources:** Implementing the PEF framework may demand significant computational resources and time, particularly when evaluating numerous models and assessing their performance, explainability, and fairness comprehensively. This complexity could limit its applicability to researchers or organizations with limited resources.
- **Subjectivity in Fairness Metrics:** The fairness assessment within the framework relies on fairness metrics, which themselves may be subject to interpretation and debate. Defining what is "fair" in any given context can be challenging, and the choice of fairness metrics may influence the results.
- **Data Quality and Bias:** The framework doesn't directly involve in data quality assessment. The data analysis and preparation with pre-processing should be included in the model development process as, in practice, real-world datasets frequently harbor inherent biases and systematic errors that can permeate the entire modeling process. Thus, mitigating data bias presents a multifaceted challenge that may necessitate supplementary measures beyond the scope of the framework's provisions.
- **Resource-Intensive Fairness Mitigation:** The framework identifies fairness issues but does not provide explicit guidance on how to mitigate them. Implementing fairness interventions, especially in real-world applications, can be resource-intensive and challenging.
- **Interpretability vs. Complexity Trade-off:** The framework faces the ongoing challenge of balancing model interpretability with complexity and performance.

Striving for high performance and fairness may lead to more complex models that are harder to interpret.

In conclusion, while the PEF framework provides a valuable structure for evaluating and benchmarking machine learning models, researchers and practitioners should be aware of its limitations and consider them when applying the framework to specific research questions or real-world applications. Addressing these limitations is crucial for refining and expanding the framework's utility and ensuring its relevance in an ever-evolving field like AI and machine learning.

7.4 Future Directions

This research opens up exciting avenues for future research in the realm of machine learning and artificial intelligence. One promising direction is the development of advanced fairness mitigation strategies within the framework. Researchers can explore novel techniques and algorithms aimed at not only identifying fairness issues but also actively addressing them during the model training process. This includes devising methods that can adaptively balance fairness and performance based on the specific requirements of different applications.

Additionally, future research can investigate deeper into the explainability component of the framework. Efforts can be directed towards enhancing model interpretability, particularly for complex models like deep neural networks, by developing more intuitive and human-understandable explanations. Investigating the convergence of explainability and fairness can also yield productive insights, with the goal of providing clear and fair explanations for model decisions, especially in critical domains like healthcare and finance.

Moreover, the scalability and efficiency of the framework can be optimized to accommodate large-scale datasets and real-time decision-making scenarios. This involves the development of distributed computing techniques and scalable algorithms for assessing and enhancing fairness and explainability in ML models.

Lastly, exploring interdisciplinary collaborations with experts in ethics, law, and social sciences can help refine the ethical and legal implications of fairness in AI systems. This interdisciplinary approach can aid in developing comprehensive guidelines and policies for deploying ML models that adhere to societal norms and regulations while ensuring equitable outcomes.

In conclusion, the future research direction for the dissertation involves refining and extending the Performance-Explainability-Fairness framework to address emerging challenges in AI and machine learning, with a focus on fairness mitigation, improved model interpretability, scalability, and interdisciplinary collaborations to foster responsible and ethical AI development.

REFERENCES

Barredo Arrieta, A., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Baskerville, R. L., Curtin University, Kaul, M., Storey, V. C., & Georgia State University. (2015). Genres of Inquiry in Design-Science Research: Justification and Evaluation of Knowledge Production. *MIS Quarterly*, *39*(3), 541–564. https://doi.org/10.25300/MISQ/2015/39.3.02

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). *Fairlearn: A toolkit for assessing and improving fairness in AI* (MSR-TR-2020-32). Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

Chakrobartty, S., & El-Gayar, O. (2021). Explainable Artificial Intelligence in the Medical Domain: A Systematic Review. *AMCIS 2021 Proceedings. 1*, 11.

Chakrobartty, S., & El-Gayar, O. (2022). Towards a Performance-explainability-fairness Framework for Benchmarking ML Models. *AMCIS*, 9.

Chakrobartty, S., & El-Gayar, O. F. (2023). Fairness Challenges in Artificial Intelligence. In *Encyclopedia of Data Science and Machine Learning* (pp. 1685–1702). IGI Global.

Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2018). *Learning to Explain: An Information-Theoretic Perspective on Model Interpretation*.

Clore, John, Cios, Krzysztof, DeShazo, Jon, & Strack, B. (2014). *Diabetes 130-US hospitals for years 1999-2008*.

Davis, J., & Goadrich, M. (2006). The Relationship between Precision-Recall and ROC Curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233–240. https://doi.org/10.1145/1143844.1143874

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*, 214–226. https://doi.org/10.1145/2090236.2090255

fairlearn.metrics package—Fairlearn 0.7.0 documentation. (2021). https://fairlearn.org/v0.7.0/api_reference/fairlearn.metrics.html

Fauvel, K., Masson, V., & Fromont, É. (2020). A Performance-Explainability Framework to Benchmark Machine Learning Methods: Application to Multivariate Time Series Classifiers. *arXiv:2005.14501 [Cs, Stat]*. http://arxiv.org/abs/2005.14501

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, *51*(5), Article 5. https://doi.org/10.1145/3236009

Gunning, D. (2017). *Explainable Artificial Intelligence (XAI)*. DARPA. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence Program. *AI Magazine*, *40*(2), Article 2. https://doi.org/10.1609/aimag.v40i2.2850

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the Class Imbalance Problem. *2008 Fourth International Conference on Natural Computation*, 192–201. https://doi.org/10.1109/ICNC.2008.871

Hall, M., Harborne, D., Tomsett, R., Galetic, V., Quintana-Amate, S., Nottle, A., & Preece, A. (2019). *A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems*. 2.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *arXiv:1610.02413 [Cs]*. http://arxiv.org/abs/1610.02413

Hasib, K. Md., Iqbal, Md. S., Shah, F. M., Al Mahmud, J., Popel, M. H., Showrov, Md. I. H., Ahmed, S., & Rahman, O. (2020). A Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem. *Journal of Computer Science*, *16*(11), 1546–1557. https://doi.org/10.3844/jcssp.2020.1546.1557

Heidari, H., Loi, M., Gummadi, K. P., & Krause, A. (2019). A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 181–190. https://doi.org/10.1145/3287560.3287584

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). *Design Science in Information Systems Research*. *28*(1), 75–105.

Holzinger, A. (2018). From Machine Learning to Explainable AI. *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, 55–66. https://doi.org/10.1109/DISA.2018.8490530

Holzinger, A., Plass, M., Holzinger, K., Crisan, G. C., Pintea, C.-M., & Palade, V. (2017). A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop. *CoRR, abs/1708.01104*. http://arxiv.org/abs/1708.01104

Kamishima, T., Akaho, S., & Sakuma, J. (2011). Fairness-aware Learning through Regularization Approach. *2011 IEEE 11th International Conference on Data Mining Workshops*, 643–650. https://doi.org/10.1109/ICDMW.2011.83

Kanda, E., Epureanu, B. I., Adachi, T., Tsuruta, Y., Kikuchi, K., Kashihara, N., Abe, M., Masakane, I., & Nitta, K. (2020). Application of explainable ensemble artificial intelligence model to categorization of hemodialysis-patient and treatment using nationwide-real-world data in Japan. *PloS One*, *15*(5), Article 5.

Kearns, M., Neel, S., Roth, A., & Wu, Z. (2017). *Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness*.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2019). An Empirical Study of Rich Subgroup Fairness for Machine Learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 100–109. https://doi.org/10.1145/3287560.3287592

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). *How We Analyzed the COMPAS Recidivism Algorithm—ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, *49*(1), Article 1. https://doi.org/10.1002/hast.973

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *CoRR, abs/1705.07874*. http://arxiv.org/abs/1705.07874

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021). On the Applicability of Machine Learning Fairness Notions. *ACM SIGKDD Explorations Newsletter*, *23*(1), Article 1. https://doi.org/10.1145/3468507.3468511

Mattson, P., Reddi, V. J., Cheng, C., Coleman, C., Diamos, G., Kanter, D., Micikevicius, P., Patterson, D., Schmuelling, G., Tang, H., Wei, G.-Y., & Wu, C.-J. (2020). MLPerf: An Industry Standard Benchmark Suite for Machine Learning Performance. *IEEE Micro*, *40*(2), 8–16. https://doi.org/10.1109/MM.2020.2974843

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), Article 6. https://doi.org/10.1145/3457607

Mohseni, S., Zarei, N., & Ragan, E. D. (2018). A Survey of Evaluation Methods and Measures for Interpretable Machine Learning. *CoRR*, *abs/1811.11839*. http://arxiv.org/abs/1811.11839

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Naylor, M., French, C., Terker, S., & Kamath, U. (2021). Quantifying Explainability in NLP and Analyzing Algorithms for Performance-Explainability Tradeoff. *arXiv:2107.05693 [Cs]*. http://arxiv.org/abs/2107.05693

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Powers, D. M. W. (2020). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation*.

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

Reddi, V. J., Cheng, C., Kanter, D., Mattson, P., Schmuelling, G., & Wu, C.-J. (2021). The Vision Behind MLPerf: Understanding AI Inference Performance. *IEEE Micro*, *41*(3), 10–18. https://doi.org/10.1109/MM.2021.3066343

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR, abs/1602.04938*. http://arxiv.org/abs/1602.04938

Shreffler, J., & Huecker, M. R. (2022). *Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios*. StatPearls Publishing, Treasure Island (FL). http://europepmc.org/books/NBK557491

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2239–2248. https://doi.org/10.1145/3219819.3220046

Tjoa, E., & Guan, C. (2020). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21. https://doi.org/10.1109/TNNLS.2020.3027314

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*, 1–7. https://doi.org/10.1145/3194770.3194776

Wightman, L. F. (1998). *LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series.*

Yeh, I.-C. (2016, January 26). *UCI Machine Learning Repository: Default of credit card clients Data Set*. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Yeh, I.-C., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473–2480.

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *Artificial Intelligence and Statistics*, 962–970.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. https://doi.org/10.1145/3278721.3278779

APPENDICES

APPENDIX A: DATASET DESCRIPTION

Finance Domain - Credit card client dataset

Table 15. Default of credit card client's dataset (Yeh & Lien, 2009)

Attribute Name	Description
ID	User ID
X1 (LIMIT_BAL)	Amount of the given credit in NT$ that includes both the individual consumer and supplementary credit.
X2 (SEX)	Gender (1 = male; 2 = female)
X3 (EDUCATION)	Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
X4 (MARRIAGE)	Marital status (1 = married; 2 = single; 3 = others)
X5 (AGE)	Age (year)
X6 – X11 (PAY_0 – PAY_6)	X6 – X11 represents correspondingly the repayment status in September 2005 to April 2005. The payback status is measured using the following scale: -1 = pay on time; The values 1..9 represents payment delay of 1..9 months and above.
X12 – X17 (BILL_AMT1 – BILL_AMT6)	The amount of the bill statement in NT$. X12 – X17 represents bill statement amount in September 2005 to April 2005.
X18 – X23 (PAY_AMT1 – PAY_AMT6)	These represents previous payment amount in NT$. X18 – X23 represents the amount paid in September 2005 to April 2005.
Y (default payment next month)	The binary variable Y is a response column, representing the default payment (1: Yes, 0: No).

Healthcare Domain - Diabetes dataset

Table 16. Diabetes dataset (Clore & Strack, 2014)

Attribute Name	Description
RACE	Race or ethnicity of the patient.
GENDER	Gender of the patient.
AGE	Age of the patient.
DISCHARGE_DISPOSITION_ID	Disposition status upon discharge from the hospital.
ADMISSION_SOURCE_ID	Source of admission to the hospital.
TIME_IN_HOSPITAL	Number of days the patient spent in the hospital.
MEDICAL_SPECIALTY	Medical specialty of the attending physician or department.
NUM_LAB_PROCEDURES	Number of laboratory procedures the patient underwent.
NUM_PROCEDURES	Number of non-laboratory procedures the patient underwent.
NUM_MEDICATIONS	Number of distinct medications prescribed to the patient.
PRIMARY_DIAGNOSIS	Primary diagnosis or medical condition for which the patient was admitted.
NUMBER_DIAGNOSES	Number of diagnoses entered to describe the patient's condition.
MAX_GLU_SERUM	Results of the maximum glucose serum test.
A1CRESULT	Results of the A1C test, a measure of long-term blood glucose control.
INSULIN	Indicating insulin is up, down or steady
CHANGE	Indicator of whether there was a change in diabetes medication.
DIABETESMED	Indicator of whether the patient was prescribed diabetes medication.
MEDICARE	Indicator of whether the patient is covered by Medicare.
MEDICAID	Indicator of whether the patient is covered by Medicaid.
HAD_EMERGENCY	Indicator of whether the patient had an emergency visit.
HAD_INPATIENT_DAYS	Indicator of whether the patient had inpatient hospitalization days.
HAD_OUTPATIENT_DAYS	Indicator of whether the patient had outpatient hospitalization days.
READMITTED	Indicator of whether the patient was readmitted to the hospital.
READMIT_BINARY	Binary indicator (0 or 1) of hospital readmission.
READMIT_30_DAYS	Binary indicator (0 or 1) of hospital readmission within 30 days.

Criminology Domain - COMPAS recidivism

Table 17. COMPAS recidivism dataset (Larson et al., 2016)

Attribute Name	Description
SEX	Gender of the individual (e.g., Male, Female).
AGE	Age of the individual at the time of assessment.
AGE_CAT	Categorized age group (e.g., Less than 25, 25-45, Over 45).
RACE	Race or ethnicity of the individual.
DECILE_SCORE	Risk assessment score assigned by the COMPAS system.
PRIORS_COUNT	Count of prior criminal convictions.
C_DAYS_FROM_COMPAS	Number of days from the COMPAS assessment to the current case.
C_CHARGE_DEGREE	Degree or severity of the current charge.
C_CHARGE_DESC	Description of the current criminal charge.
IS_RECID	Binary indicator (0 or 1) whether the individual is recidivating.
IS_VIOLENT_RECID	Binary indicator (0 or 1) whether the individual committed a violent offense.
SCORE_TEXT	Textual representation of the risk score (e.g., Low, Medium, High).
V_DECILE_SCORE	Risk assessment score for violent recidivism.
V_SCORE_TEXT	Textual representation of the violent recidivism risk score.
TWO_YEAR_RECID	Binary indicator (0 or 1) whether the individual was rearrested or reincarcerated within two years of the assessment.

Education Domain - Law school admissions dataset

Table 18. Law school admission dataset (Wightman, 1998)

Attribute Name	Description
DECILE1B	Academic decile rank based on first-year law school grades.
DECILE3	Academic decile rank based on third-year law school grades.
LSAT	Law School Admission Test (LSAT) score.
UGPA	Undergraduate Grade Point Average (UGPA).
ZFYGPA	Z-Score of the first-year law school grades.
ZGPA	Z-Score of overall law school grades.
FULLTIME	Indicator of whether the student attended law school full-time.
FAM_INC	Family income or socioeconomic status of the student.
MALE	Gender of the student (Male or Female).
TIER	Tier or ranking of the law school attended by the student.
RACE	Race or ethnicity of the student.
PASS_BAR	Indicator of whether the student passed the bar examination.

APPENDIX B: THE SURVEY INSTRUMENT

Task to be Completed

Use the eight-dimensional framework to capture the trade-offs among potential ML models in terms of their performance, explainability and fairness.

Questions

Yes/No Questions

1. Is the “Performance” characteristic of the framework relevant and clear?
2. Is the “Comprehensibility” characteristic of the framework relevant and clear?
3. Is the “Granularity” characteristic of the framework relevant and clear?
4. Is the “Information type” characteristic of the framework relevant and clear?
5. Is the “Faithfulness” characteristic of the framework relevant and clear?
6. Is the “User category” characteristic of the framework relevant and clear?
7. Is the “Fairness Context” characteristic of the framework relevant and clear?
8. Is the “Fairness” characteristic of the framework relevant and clear?
9. Does the framework currently miss a dimension that is important?

Rating Questions

Questions with rating answers on a scale of: Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree.

10. The fairness characteristics of the framework are important for the models my team develops and deploys.
11. The performance-explainability-fairness model benchmarking framework is useful.
12. The performance-explainability-fairness model benchmarking process is easy to follow.
13. I am inclined to use this framework in my project.

Open Ended Questions

14. What are the strengths of the framework?
15. How can the proposed framework be improved?