Explainable Artificial Intelligence across Domains: Refinement of SHAP and Practical Applications

Veera Raghava Reddy Kovvuri

Submitted to Swansea University in fulfilment
of the requirements for the Degree of Doctor of Philosophy

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

March 2024

Declaration

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed [redacted] (candidate)

Date 25/06/2024

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed [redacted] (candidate)

Date 25/06/2024

I hereby give consent for my thesis, if accepted, to be available for electronic sharing.

Signed [redacted] (candidate)

Date 25/06/2024

The University's ethical procedures have been followed and, where appropriate, that ethical approval has been granted.

Signed [redacted] (candidate)

Date 25/06/2024

Abstract

Explainable Artificial Intelligence (XAI) has become a crucial area within AI, emphasizing the transparency and interpretability of complex models. In this context, this research meticulously examines diverse datasets from medical, financial, and socio-economic domains, applying existing XAI techniques to enhance understanding and clarity of the results. This work makes a notable contribution to XAI by introducing the Controllable fActor Feature Attribution (CAFA) approach, a novel method that categorizes dataset features into 'controllable' and 'uncontrollable' groups. This categorization enables a more nuanced and actionable analysis of feature importance. Furthermore, the research proposes an extension to CAFA, the Uncertainty-based Controllable fActor Feature Attribution (UCAFA) method, which incorporates a Variational Autoencoder (VAE) to ensure that perturbations remain within the expected data distribution, thereby enhancing the reliability of feature attributions. The effectiveness and versatility of CAFA are showcased through its application in two distinct domains: medical and socio-economic. In the medical domain, a case study is conducted on the efficacy of COVID-19 non-pharmaceutical control measures, providing valuable insights into the impact and effectiveness of different strategies employed to control the pandemic. Additionally, UCAFA is applied to the medical domain, demonstrating its ability to improve the reliability of feature attributions by considering uncertainty. The socio-economic domain is investigated by applying CAFA to several datasets, yielding insights into income prediction, credit risk assessment, and recidivism prediction. In the financial domain, the analysis focuses on global equity funds using established XAI methodologies, particularly the integration of the XGBoost model with Shapley values. This analysis provides critical insights into fund performance and diversification strategies across G10 countries. This thesis highlights the potential of CAFA and UCAFA as promising directions in the domain of XAI, setting the stage for advanced research and applications.

Acknowledgements

I am profoundly grateful to my supervisor, Dr. Monika Seisenberger, for her steadfast support, guidance, and mentorship throughout my PhD journey. The opportunities for growth and development she provided, along with her invaluable insights and encouragement, have been pivotal in my evolution both as a researcher and as a person. I am deeply appreciative of the knowledge and wisdom she has shared, which will undoubtedly influence my future endeavors.

My sincere thanks also go to my co-supervisor, Dr. Xiuyi Fan, for his unwavering support and guidance, particularly during the critical initial two years of my PhD. His engagement in various events and activities has offered me enriching experiences that have profoundly contributed to my personal and professional growth. I am especially thankful for her meticulous attention to detail and constructive feedback, which have significantly enhanced the quality of my work. I am indebted to Dr. Hsuan Fu for hosting me in Canada within the Department of Finance and Real Estate, thereby broadening my domain knowledge in Finance.

To my family—my late father, Venkata Reddy Kovvuri; my mother, Bhavani Kovvuri; my wife, Anusha Pothamsetti; my brother, Sudhakar Reddy Kovvuri and his wife, Poojitha Surya Kala; my father-in-law, S N V S N Reddy Pothamsetti; my mother-in-law, Dhanalakshmi Pothamsetti; and my sister-in-law, Hema Sri Pothamsetti—my heart is filled with gratitude for your unconditional love, sacrifice, and support. Your unwavering faith in me has always been my source of motivation and inspiration. This thesis is dedicated to you as a modest expression of my boundless gratitude for all that you have done for me.

I extend my heartfelt appreciation to my colleagues— Dr. Jamie Duell and Harry Bryant—for their companionship and unwavering support. Special thanks are due to Dr. Siyuan Liu and Dr. Berndt Müller for their invaluable contributions to my research. Thanks to Kalyan Chakravathi Gogula for his continuous support since my Bachelor's. My gratitude extends to my friends—Dr. Jamie Duell, Akhil Baby, Abhijith Jiji, Aditya Reddy Gudimetla, Reshma Reddy Palli and Kumari Bireddy—for their unwavering support and encouragement.

Lastly, I wish to acknowledge Prof. Wai Lok Woo and Dr. Sachi Nandan Mohanty, my external examiners, whose invaluable feedback on the manuscripts that constitute the core of this thesis has been instrumental in its refinement and improvement. Their constructive criticism and insights have been crucial in elevating the quality of my work. I am deeply thankful for their time and expertise.

Contents

List of Figures

List of Tables

List of Symbols and Acronyms

Acronyms

AI Artificial Intelligence

AR(1) First-order Autocorrelation

AUC Area Under the Curve

BEL Belgium

CAD Canadian Dollar

CAFA Controllable fActor Feature Attribution

CAN Canada

CDF Cumulative Distribution Function

CHE Switzerland

CHF Swiss Franc

CM Control Measures

CNN Convolutional Neural Network

COMPAS Correctional Offender Management Profiling for Alternative Sanctions

COVID-19 Coronavirus Disease 2019

CR Cafes and Restaurants

D-LIME Deterministic-LIME

DEU Germany

DNN Deep Neural Network

DT Domestic Travel

ELBO Evidence Lower Bound

ER Exchange Rates

EURO Euro

FN False Negative

FPR False Positive Rate

FP False Positive

FRA France

G10 Group of Ten

GBP British Pound

GBR United Kingdom

GPU Graphics Processing Unit

HHI Herfindahl Hirschman Index

HV Hospitals/Care and Nursing Home Visits

H Humidity

IR Interest Rates

ITA Italy

IT International Travel

JPN Japan

JPY Japanese Yen

KL Kullback-Leibler

LIME Local Interpretable Model-Agnostic Explanations

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

MInd Meeting Indoors

MLE Maximum Likelihood Estimation

ML Machine Learning

MOut Meeting Outdoors

MSE Mean Squared Error

NAV Net Asset Value

NCRAS National Cancer Registration and Analysis Service

NI Northern Ireland

NLD Netherlands

NS Non-essential Shops

PB Pubs and Bars

PPrfm Past Performance

RMSE Root Mean Squared Error

RNN Recurrent Neural Network

ROC Receiver Operating Characteristic

SC School Closures

SEK Swedish Krona

SHAP SHapley Additive exPlanations

SL Sports and Leisure

ST Stock Market

SVM Support Vector Machines

SWE Sweden

TNM Tumor, Node, Metastasis

TN True Negative

TPR True Positive Rate

TP True Positive

T Temperature

UCAFA Uncertainty-based Controllable Factor Feature Attribution

UCI University of California, Irvine

UK United Kingdom

US United States

VAE Variational Autoencoder

WHO World Health Organization

XAI eXplainable Artificial Intelligence

XGBoost eXtreme Gradient Boosting

Symbols

$\bar{y}$ Mean of the actual values

β Coefficients

$\beta_0, \beta_1, \ldots, \beta_p$ Regression coefficients

β_i Coefficient associated with feature i

$\boldsymbol{\beta}$ Vector of regression coefficients

δ Uncertainty threshold

$\gamma(\mathbf{x})$ Perturbation function on $\mathbf{x}$

Γ Family of distributions

$\hat{y}_i$ Predicted value

κ Discrete time index

$\mathbb{E}$ Expected value

$\mathbf{c}$ Latent space

$\mathbf{x}$ Instance in the original dataset X

$\mathbf{y}'$ Vector of predicted values

$\mathbf{z}$ Instance in the perturbed samples $\mathcal{Z}$

$\mathbf{z}'$ Instance in the filtered neighborhood $\mathcal{Z}'$

$\mathcal{F}$ Set of potential black-box models

$\mathcal{L}$ Loss function

$\mathcal{R}t$ Rate of infection

$\mathcal{Z}$ Perturbed samples

$\mathcal{Z}'$ Filtered neighborhood

ω_i Weight of feature i

VAELoss($\mathbf{x}$) VAE loss function

$\Phi(\cdot)$	CDF of the standard normal distribution
Φ	Feature importance values
$\Phi^j(\mathbf{x})$	Linear SHAP value for feature x^j
Φ_i	Explanation for instance i
ϕ_i	Shapley value for feature i
$\pi_{\mathbf{x}}$	Proximity threshold
$\pi_x(x')$	Proximity measure between instance x and sample x'
ψ	Optimal parameters for the decoder
$\sum_i s_i$	Sum of portfolio holdings across countries
Var	Variance
θ	Optimal parameters for the encoder
$c\tau$	Number of new infections on day τ
C_t	Confirmed cases on day t
c_t	Number of new infections on day t
d	Distance metric
$D_{\mathbf{x}}$	Dataset generated for data point $\mathbf{x}$
D_{KL}	KL divergence
F	Black-box model
f	Prediction model
F_c	Set of controllable features
F_t	Filtered confirmed cases on day t
F_u	Set of uncontrollable features
$f_x(S)$	Prediction of the model using the features in set S
g	Strong prediction model
$g_{t-\tau}$	Value of the Gamma distribution at time $t - \tau$
K	Sample class size
m	Number of features

N	Set of all features
n	Number of points in the dataset
P	Original data distribution
p	p-value
Q	Perturbed data distribution
R^2	R-squared (coefficient of determination)
R_t	Daily rate of infection
S	Subset of features excluding feature i
s_i	Portfolio holding in country i
$s_{\text{non-US}}$	Portfolio holding in non-US countries
s_{US}	Portfolio holding in the United States
v_j	Value of feature j
v_j^i	Value of feature j in instance i
$w(S)$	Weight assigned to the subset S
X	Original dataset
$x_1, \ldots, x_p$	Predictor variables
x_i	Feature value
y	Binary outcome variable

Chapter 1

Introduction

Contents

1.1 Motivation

The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) over the past decade has led to transformative breakthroughs in capabilities. AI systems can now surpass human performance on complex cognitive tasks like strategic gameplay [SSS+17], image recognition [HZRS15], and language translation [WSC+16]. They also power critical real-world technologies like medical diagnosis, autonomous vehicles, facial analysis, and financial algorithmic trading [JM15]. However, as AI continues to expand into these sensitive societal domains, concerns around fairness, accountability, and transparency have rapidly emerged as well [Rud19]. Most state-of-the-art AI involves complex data-driven systems like deep neural networks that operate as inscrutable "black boxes". Though very performant, these opaque models offer no transparency and interpretability into how they arrive at different predictions or decisions. This lack of explainability severely limits appropriate trust and adoption of AI across areas like healthcare, finance, and law.

Thus, explainable AI (XAI) has become a crucial research discipline focused on clarifying, demystifying and providing post-hoc explanations of model decisions and inferences [GMR+18]. The goal is to enable human users to comprehend model rationale and evaluate systemic biases, especially when AI is used for impactful tasks like diagnosing illness, granting loans, or assessing risk. This thesis specifically delves into the pressing challenge of advancing explainability, interpretability and trust in AI systems

operating on real-world critical datasets across medical, financial and socio-economic domains.

The subsequent sections delve into the evolution of Artificial Intelligence and Machine Learning, shedding light on the current paradigms in Explainable AI (XAI), including various methods and applications. Additionally, they present a comprehensive analysis of the challenges inherent in existing XAI algorithms.

1.1.1 Evolution of Artificial Intelligence and Machine Learning

The origins of artificial intelligence can be traced back to the Dartmouth Workshop in 1956, where the term "artificial intelligence" was coined and the field was defined as the effort to automate intellectual tasks normally performed by humans [MMRS55]. In the first few decades, research in AI focused largely on symbolic reasoning, knowledge representation, and search algorithms to solve problems. For instance, Newell and Simon's Logic Theorist program managed to prove mathematical theorems using heuristic search [NS56]. However, these early AI systems were limited by the computation power at the time and lacked enough data to learn effectively. As a result, progress stalled after initial optimism. The 1970s saw expert systems and knowledge bases for specific domains, but these relied extensively on hand-coded rules crafted by human experts. The inability of systems to automatically learn held back more broad advancement of AI [Buc05].

A major breakthrough arrived in the late 1990s, when machine learning emerged to the forefront as a subfield of AI. In contrast to manually coded rules, ML algorithms are designed to automatically learn patterns and insights from data. Especially since 2010, the exponential increase in available training data converges with immense computing advances from GPUs. This enables more complex statistical and neural network models to be efficiently trained on much larger datasets [JM15].

Modern AI is now primarily fueled by data-intensive machine learning techniques, especially deep neural networks. State-of-the-art systems have achieved remarkable predictive breakthroughs in areas like computer vision, speech recognition, game playing agents, and language translation [LBH15]. However, as much of contemporary AI has focused on predictive accuracy, the complex models powering state-of-the-art systems remain mostly "black-box" with little transparency into their decision-making processes. This lack of explainability gives rise to risks around trust and accountability as AI gets deployed in real-world scenarios. The pressing need for explainable and interpretable AI provides the key motivation behind this thesis exploring XAI methodologies across critical domains.

1.1.2 Current Paradigms in Explainable AI: Methods and Applications

As AI accelerates across critical domains like healthcare, justice, finance, and transportation, concerns around ethics, fairness, transparency and accountability have rapidly

emerged [ABC+19]. This underscores the crucial need for eXplainable AI (XAI) to clarify the internal logic and decision-making behind complex AI systems.

Broadly, XAI techniques fall under two paradigms: ad-hoc and post-hoc [ADRDS+20]. Ad-hoc explainability refers to techniques that are inherently interpretable and transparent by design. These include simple linear/logistic regression models, decision trees, rule-based systems like expert systems, and generalized additive models. Owing to their simplicity, ad-hoc methods provide straight-forward explanations about the relationship between input features and outputs. However, their accuracy lags behind complex models [Mol23]. Post-hoc explanation techniques are specifically focused on deciphering the inner workings of low-interpretability 'black-box' models like deep neural networks, support vector machines and ensemble methods [RvGH18]. Strategies encompass developing intrinsically interpretable proxy models to approximate the behavior of black-boxes or utilizing feature attribution to highlight input variables that influenced certain predictions [LL17].

While ad-hoc and post-hoc techniques have dominated, hybrid XAI approaches are also emerging to combine strengths of both paradigms [WYAL19]. For example, an intrinsically interpretable decision tree can be used to approximate and explain a neural network's behaviors. The transparency of the decision tree then supplements the accuracy of the original complex model. Explanations can be broadly categorized based on their scope - global explanations provide an overview of the model's overall behaviors while local explanations analyze individual predictions [TK19]. A global explanations is useful for purposes like model comparison and debugging biases. Meanwhile, local explanations enable case-specific understandings - like justifying loan decisions for applicants or clarifying diagnoses for patients.

Interactive XAI is an evolving paradigm focused on incorporating humans in the loop for exploratory explanations [AVW+18]. Here, stakeholder input guides the explanation process to customize and refine presented information based on user needs and domain constraints. This facilitates iterative trust building between humans and AI systems. Initial research has explored interactive visual interfaces, but substantial potential exists for conversational approaches as well. Domain applications are rapidly emerging across many areas [ABC+19]: In criminal justice, XAI can audit recidivism risk calculators; in healthcare, show doctors the reasons behind diagnostic predictions before deciding treatments; in finance, demystify credit scores for applicants; and in hiring, ensure fairness and mitigate unconscious biases. For safety-critical autonomous systems, explainability is also key to debug failures transparently and build public trust through accountability [ADRDS+20].

1.1.3 Challenges in Existing XAI Algorithms

While active research has led to promising developments in explainable AI (XAI), several crucial challenges remain open:

- **Accuracy vs. Explainability Tradeoffs**: State-of-the-art machine learning models that deliver highest predictive accuracy like deep neural networks and

ensemble techniques also tend to be complex black boxes with little transparency. Simpler linear models or decision trees are interpretable but far less accurate. Finding the right tradeoff between performance and explainability or generating faithful explanations for accurate complex models remains an active challenge [LL17].

- **Rigorous Human-Centric Evaluations**: Most current XAI techniques rely on proxy metrics and simulated user experiments for evaluation. However, quantitativesimilarity measures do not properly align with human understanding and simulated settings lack realism. Developing rigorous human-grounded evaluation frameworks with end-user studies for standardized and ethical testing is thus critical [DVK17].

- **Scalability & Generalizability**: Many state-of-the-art explanation methods perform well on small datasets but struggle to scale effectively to large high-dimensional modern datasets with computational efficiency. Additionally, they are often tailored to specific model types like neural networks. Enabling useful explanations for immense datasets across different model families like boosting and graphical models remains an open challenge [GMR+18].

- **Interactive Explanations**: Existing XAI approaches focus primarily on static explanation outputs. However, interactive paradigms that support users exploring and guiding explanations based on their needs can build better trust. outside strict laboratory contexts [WYAL19]. Realizing such flexible interactive interfaces is an emerging imperative.

- **Security Against Attacks**: Studies reveal possibilities of adversaries manipulating explanations deliberately to distort model behavior and erode user trust [SHJ+20]. Creating rigorous testing standards and defense mechanisms to ensure explanation robustness against such vulnerability exploits is now a growing research priority.

1.1.4 Feature Attribution in Contemporary AI Research

Feature attribution refers to techniques focused on identifying the relative influence or importance of input variables towards model outputs and predictions [AB18]. As complex machine learning models like Deep Neural Networks (DNNs) gain mainstream adoption across application areas such as computer vision, natural language processing, and healthcare, feature-attribution methods have emerged as a vital component of providing post-hoc explanations about model behaviors and decisions. Common approaches for feature-attribution analysis include:

- Sensitivity analysis: Evaluating output variance to systematic changes in an input feature's value [ZGMO22]. This quantifies the marginal effect of features.

- Gradient-based attribution: Using gradient information flowing into neural networks to assign contextual importance scores to features for a given prediction [ACG18].

- Perturbation-based attribution: Systematically masking or altering features to quantify resultant impacts on outputs compared to original model [FFR20].

- Surrogate models: Simpler intrinsically interpretable models trained to approximate attribution insights from complex black-box models.

Key applications of contemporary feature attribution are towards debugging model behaviors by flagging problematic correlations, auditing algorithms to ensure fairness and mitigate biases, and providing local explanations to build appropriate trust with end-users [Mol23]. Importance scores can trace dependencies in model logic and highlight actionable variables for recourse. However, recent studies have also exposed possible vulnerability of explanations to adversarial attacks intended to deliberately mislead feature attribution [SHJ+20]. Rigorously evaluating faithfulness of explanations and developing defense mechanisms are thus also rising priorities. Explanation techniques are also being integrated earlier into model development workflows for intrinsically interpretable designs [CRZ+22]. As attribution analysis sees greater adoption, maintaining standards on transparency and accountability will be crucial [ADRDS+20].

1.2 Contributions

This thesis examines diverse real-world datasets spanning medical, socio-economic and financial domains, applying XAI methodologies to enhance understanding and clarify model behaviors. The research contributions are fourfold:

- **Novel CAFA Methodology**: This research puts forward a new explainable AI (XAI) technique called Controllable fActor Feature Attribution (CAFA) [KLS+22] to selectively compute feature importance for controllable factors. CAFA addresses a limitation of existing feature attribution algorithms that treat all input features homogenously. By distinguishing between controllable features, which can be actively altered or adjusted by stakeholders to impact the outcome, and uncontrollable features, which are inherent or predetermined, CAFA excludes the influence of uncontrollable features when explaining individual predictions. Specifically, it generates a dataset by perturbing only controllable features while fixing uncontrollable features, and interprets the global feature importances from a strong predictor fitted on this dataset as the local explanation.

  The key novelty of CAFA lies in enabling explanations that focus exclusively on controllable features, without having to subset the data fed into the prediction model. This preserves model performance while granting users actionable insights - such as gauging the effectiveness of medical interventions or policy controls. Experiments on a lung cancer dataset, breast cancer data and in analyzing

COVID-19 control measures showcase CAFA's reliability and usefulness. The consistent results between CAFA and benchmark methods like SHAP on the controllable subsets validate that it inherits the desirable properties of standard feature attribution algorithms. Overall, CAFA puts forth a simple yet powerful approach to handle feature heterogeneity for practical interpretability. The selective perturbation concept may spur further research towards controlling or directing the explanations derived from black-box AI systems.

- **COVID-19 Policy Explainability**: A major contribution of this work is applying the proposed CAFA method to gain insights into COVID-19 control policies and their effectiveness in containing virus transmission. Specifically, CAFA is used to assess the impact of various non-pharmaceutical interventions on the reproduction rate Rt as an indicator of epidemic spread. By filtering out the influence of uncontrollable factors, CAFA provides a clear picture of the most effective government measures. The top policies identified include restrictions on cafes, restaurants, pubs and bars - aligning with WHO guidelines on limiting crowded and confined spaces.

  From a methods perspective, this novel case study highlights the benefits of CAFA in explaining policy impacts. Training a random forest classifier to predict high/low transmission from policy actions, CAFA reveals the key control measures while overriding the strong signals from uncontrollable but non-actionable features like daily infections. Such selective explainability prevents misleading conclusions on policy relevance. The findings showcase CAFA's reliability in not only maintaining predictive accuracy but also directing explanations towards controllable levers - granting users actionable insights.

  Overall, the COVID-19 analysis provides a valuable demonstration of using CAFA for targeted explainability in policy making. By controlling the set of factors explanations are based on, CAFA exceeds conventional XAI methods in highlighting actionable and impactful policy options. The case study sets a precedent for deploying CAFA in other domains with heterogeneous features, where decision-makers need to isolate influences of interest. This can pave the way for optimized, transparency-aware policy planning and governance.

- **UCAFA: Uncertainty-based CAFA**: Building upon the CAFA approach, this research introduces the Uncertainty-based Controllable Factor Feature Attribution (UCAFA) method. UCAFA extends CAFA by leveraging a Variational Autoencoder (VAE) to ensure perturbations remain within the expected data distribution, addressing the issue of out-of-distribution samples that can skew explanations. By maintaining the focus on controllable factors and enforcing an uncertainty threshold, UCAFA significantly improves the reliability of feature attributions.

  Experiments on three healthcare datasets (lung cancer, breast cancer, and COVID-19) demonstrate UCAFA's superior performance compared to existing methods like LIME, SHAP, and CAFA. UCAFA exhibits faster convergence to baseline

probabilities, lower perturbation sensitivity, and reduced error rates. These findings underscore the importance of accounting for uncertainty when generating perturbations for model explanations. By focusing on in-distribution perturbations, UCAFA provides more reliable and interpretable feature attributions.

The enhanced interpretability and reliability of machine learning models in healthcare, as demonstrated by UCAFA, have significant medical implications. By providing more accurate and trustworthy explanations, UCAFA empowers healthcare professionals to make better-informed decisions regarding diagnosis, treatment, and resource allocation, potentially improving patient outcomes and healthcare efficiency. As such, UCAFA contributes to the growing field of explainable AI in healthcare, paving the way for more transparent and reliable clinical decision support systems.

- **Financial Performance Diagnostics**:

  This research makes several notable contributions to the literature on fund performance evaluation. First, it is the first study to apply explainable artificial intelligence (XAI) techniques, specifically XGBoost and SHAP, to examine the complex nonlinear relationships between various macro-financial and fund-level factors and fund performance. Leveraging the predictive power of machine learning and the interpretability of XAI, we uncovered novel insights into the diversification implications for country portfolios - finding that both over- and under-diversification can hurt performance, while a moderate level of diversification is optimal.

  Additionally, this research establishes the reliability and consistency of using XAI in financial applications. The signs and significance of relationships from the SHAP analysis align with the benchmark linear regression, and the findings are robust across countries and time periods. I therefore showcase the potential of XAI to supplement domain knowledge and provide richer implications, tackling open research questions. Finally, through examining subsamples based on diversification levels, the study reveals previously ambiguous effects of the Herfindahl-Hirschman Index on performance. we highlight the need to account for nonlinear effects when assessing portfolio concentrations. Overall, this research puts forth XAI as an impactful tool for gaining nuanced insights from complex financial data [KFFS23].

Through innovations in explainability methodology plus interdisciplinary demonstrations, this thesis opens new directions towards increasing transparency and trust in AI systems applied across highly consequential real-world contexts.

1.3 Published Works

The key peer-reviewed published research contributions as first author and co-author that form the backbone for this dissertation are:

1.3.1 Paper 1: On understanding the influence of controllable factors with a feature attribution algorithm: a medical case study

- **Kovvuri V.R.R., Liu S., Seisenberger M., Fan X., Muller B., Fu H. (2022).** On understanding the influence of controllable factors with a feature attribution algorithm: a medical case study. *2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, IEEE. This paper introduces a new XAI technique called Controllable fActor Feature Attribution (CAFA) which categorizes input features into controllable and uncontrollable groups. Quantitative analysis on medical datasets demonstrates CAFA's ability to filter out influences of uncontrollable variables in explanations while maintaining predictive accuracy. Qualitative assessment applies CAFA to evaluating the effectiveness of various COVID-19 policy interventions.

- **Kovvuri V.R.R. (First author)** independently developed the new CAFA technique for XAI, conducted literature review, as well as designed and led quantitative analysis. Kovvuri drafted full manuscript and revisions.

- **Liu S., Fan X., Seisenberger M., Muller B., and Fu H. (Co-Authors)** provided guidance on methodology, contributed narrowly to the implementation process and feedback on analysis and writing.

1.3.2 Paper 2: Fund performance evaluation with explainable artificial intelligence

- **Kovvuri V.R.R., Fu H., Fan X., Seisenberger M. (2023).** Fund performance evaluation with explainable artificial intelligence. *Finance Research Letters*, Elsevier. This article demonstrates the integration of machine learning and explainable AI to uncover drivers of equity fund growth across G10 economies. Analysis leverages the XGBoost model and SHAP for feature attribution. Explanations provide novel insights into the role of international diversification in determining portfolio performance. Results highlight the potential benefits of moderate diversification along with risks of over- and under-diversification.

- **Kovvuri V.R.R. (First author)** independently led literature review, data collection, model development, analysis of results, drafting and revising of the manuscript. Kovvuri spearheaded the integration of machine learning and XAI to evaluate drivers of equity fund growth.

- **Fu H., Seisenberger M., and Fan X. (Co-Authors)** provided guidance on methodology, and feedback on analysis and writing.

1.3.3 Paper 3: UCAFA: Uncertainty-based Controllable Factor Feature Attribution for Medical Records

- **Kovvuri V.R.R., Duell J., Fu H., Seisenberger M., Fan X. (2023).** UCAFA: Uncertainty-based Controllable Factor Feature Attribution for Medical Records. Submitted to the *22nd International Conference on Artificial Intelligence in Medicine (AIME 2024)*, 05/04/2024. This paper introduces the Uncertainty-based Controllable Factor Feature Attribution (UCAFA) method, an extension of the CAFA approach that leverages a Variational Autoencoder (VAE) to ensure perturbations remain within the expected data distribution. By maintaining the focus on controllable factors and enforcing an uncertainty threshold, UCAFA significantly improves the reliability of feature attributions. Experiments on three healthcare datasets demonstrate UCAFA's superior performance compared to existing methods like LIME, SHAP, and CAFA.

- **Kovvuri V.R.R. (First author)** independently developed the UCAFA method, conducted literature review, designed and led the experiments, and drafted the full manuscript and revisions.

- **Duell J., Fu H., Seisenberger M., and Fan X. (Co-Authors)** provided guidance on methodology, contributed to the implementation process, and provided feedback on analysis and writing.

1.3.4 Paper 4: An Initial Study of Machine Learning Underspecification Using Feature Attribution Explainable AI Algorithms: A COVID-19 Virus Transmission Case Study

- **Hinns J., Fan X., Liu S., Kovvuri V.R.R., Yalcin M., Roggenbach M. (2021).** An Initial Study of Machine Learning Underspecification Using Feature Attribution Explainable AI Algorithms: A COVID-19 Virus Transmission Case Study. *Lecture Notes in Computer Science*, Springer. This paper develops the concept of using feature attribution algorithms to identify machine learning model underspecification. The lead authors design and execute literature analysis, methods, experiments, results analysis, and authoring of the paper. As a co-author, Kovvuri contributes to this paper by assisting with relevant data collection and preprocessing. Kovvuri also provides feedback on draft versions of the manuscript.

- **Kovvuri V.R.R. (Co-Author)** contributed to this paper as a co-author by assisting with data collection and preprocessing for the COVID-19 virus transmission case study. Kovvuri also reviewed draft versions of the manuscript and provided feedback to the lead authors prior to publication.

1.3.5 Conference and Workshops Talks

1.3.5.1 Conference Talks

- Kovvuri, V.R.R. (2022). On understanding the influence of controllable factors with a feature attribution algorithm: A medical case study. In *2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, Biarritz, France. (Conference talk, delivered online via Zoom on August 9, 2022)

  Abstract: Feature attribution explainable AI (XAI) algorithms enable users to gain insight into the underlying patterns of large datasets through feature importance calculation. However, existing feature attribution algorithms treat all features in a dataset homogeneously, which may lead to misinterpretation of the consequences of changing feature values. In this work, we propose partitioning features into controllable and uncontrollable parts and introduce the Controllable fActor Feature Attribution (CAFA) approach to compute the relative importance of controllable features. We conducted experiments applying CAFA to two existing datasets and a novel COVID-19 non-pharmaceutical control measures dataset. The results demonstrate that CAFA can exclude the influences of uncontrollable features in the explanations while maintaining the full dataset for prediction.

1.3.5.2 Workshop Talks

- Kovvuri V.R.R. (2022). Controllable Actor Feature Attribution (CAFA) Report on BCTCS 2022.*Bulletin of EATCS*, **137**(2). This article was part of the XAI Special Session at the 2022 British Colloquium for Theoretical Computer Science (BCTCS), held at Swansea University from April 11-13, 2022. The 30-minute presentation introduced the Controllable Actor Feature Attribution (CAFA) approach for creating explainable AI models that can selectively compute feature importance between controllable and uncontrollable variables. Quantitative experiments on medical datasets demonstrated CAFA's ability to generate explanations focused solely on controllable features, eliminating interference from uncontrollable ones. This enables more reliable interpretations of how changes in controllable factors impact outcomes. The novel method and its analysis were well-received by an audience of over 60 theoretical computer scientists, generating insightful discussions on its applications in healthcare and other domains, such as finance and policy modeling.

- Kovvuri, V.R.R. (2022). Case study: COVID-19 non-pharmaceutical control measures dataset. In *School of Mathematics and Computer Science Research Day*, Swansea University, UK. (Research talk, delivered on May 27, 2022) **Abstract:** In this case study, we applied the Controllable fActor Feature Attribution (CAFA) approach to understand the effectiveness of COVID-19 non-pharmaceutical control measures. By analyzing the feature importance of various interventions, we found that restricting access to cafes, restaurants, pubs, and bars were the most

effective measures in containing the disease, as indicated by achieving an effective reproduction number (Rt) smaller than 1. This research demonstrates the potential of CAFA in providing actionable insights for policymakers in managing public health crises.

- Kovvuri V.R.R. (2022). AISB Workshop on Explainability and Transparency in AI (XTAI 2022) Investigating Global Open-Ended Funds diversification among G 11 countries through XAI Open-Ended funds are run by asset managers to diversify the funds pooled through the investment. In doing so, the specific risks associated with pooled funds can be mitigated. In this research, we use an eXplainable Artificial Intelligence (XAI) feature attribution algorithm to quantify the strategy of diversification based on its effect on the corresponding Net Asset Value (NAV). To do this, we collected data from the Morning Star Direct software database containing 313,737 unique funds and their fund allocation from December 2000 to November 2021 with a total of 21 Years as month frequency across G11 countries. The preliminary results using the funds originating from the USA, UK and Canada across G11 countries show that the important features using Shapley Additive eXplanation (SHAP) are "Stock Index" and "Funds Performance" with respect to previous quarters have a high influence towards the dynamics of NAV.

- Invited Speaker and Lead Facilitator, Full-Day Workshop on Explainable AI (XAI) in Finance, University Laval, March 20, 2023. Presented a 1-hour seminar providing an introduction to the SHAP explainable AI technique, including the intuition behind it and hands-on examples of its application within machine learning pipelines in Python. Demonstrated how SHAP can be leveraged for feature attribution to interpret model predictions and quantify feature importance. The talk set the foundation for a follow-up 2-hour hands-on coding workshop focused entirely on SHAP implementation. During the workshop, conducted interactive analysis of financial data using Jupyter notebooks, allowing attendees to gain practical experience leveraging SHAP for feature attribution and producing model explanations.

1.4 Thesis Overview

This thesis examines the application of explainable artificial intelligence (XAI) techniques across diverse datasets in medical, financial, and socio-economic domains areas to promote transparency and elucidate behaviors of complex predictive models on diverse datasets.

Motivation As artificial intelligence proliferates across high-stakes domains like healthcare and finance, model interpretability and explanations grow indispensable for trust, auditability and human-centered design (Chapter 1). Core questions tackled include: What key factors drive a model's predictions? How can we selectively explain parts of a model? Can we control explanations to focus only on actionable insights?

Background Methodology Spanning tabular datasets from said domains (Chapter 2), core techniques utilized entail: (i) Random forest and XGBoost for prediction (ii) SHAP and LIME for post-hoc explanation via feature attribution. By attributing relevance to input variables behind outcomes, these XAI techniques crucially clarify model mechanisms and relationships.

Novel Contribution 1 - CAFA Method A flagship contribution is the introduced Controllable fActor Feature Attribution (CAFA) technique (Chapter 5) that distinguishes between controllable and uncontrollable input variables. Via selective perturbation and global-for-local interpretations, CAFA generates explanations focused exclusively on controllable features. This prevents interference from unactionable ones in directing model transparency towards informed decision-making. Validated quantitatively on medical data (Chapter 3) applying CAFA on electronic health records data reveals tailored insights into how modifying treatments or lifestyle factors (controllable features) can impact risks, while overriding unchangeable factors like genetics or past conditions. Analyzing COVID-19 control measures shows CAFA highlights the most prudent government interventions aligned with WHO guidance, by filtering out noise from uncontrollable features like base transmission rates.

Novel Contribution 2 - COVID-19 Policy Explainability A major contribution involves formulating the effectiveness of various non-pharmaceutical interventions against the COVID-19 pandemic as an XAI modeling problem (Chapter 4). By applying the novel CAFA method to filter out uncontrollable factors, the analysis quantitatively assesses the impact of strategies like lockdowns, closures and mobility restrictions on managing virus transmission. The findings, which align with WHO guidance, showcase CAFA's ability to highlight the most effective government measures. This offers data-driven guidance into policy decisions.

Novel Contribution 3 - UCAFA Method Building upon CAFA, this research introduces the Uncertainty-based Controllable Factor Feature Attribution (UCAFA) method (Chapter 6). UCAFA extends CAFA by leveraging a Variational Autoencoder (VAE) to ensure perturbations remain within the expected data distribution. By maintaining the focus on controllable factors and enforcing an uncertainty threshold, UCAFA significantly improves the reliability of feature attributions. Experiments on three healthcare datasets demonstrate UCAFA's superior performance compared to existing methods like LIME, SHAP, and CAFA. The enhanced interpretability and reliability provided by UCAFA have significant implications for supporting clinical decision-making.

Novel Contribution 4 - Financial Insights Additionally, predictive modeling integrated with SHAP uncovers non-intuitive, previously ambiguous results across disciplines (Chapter 7), including how equity mutual fund growth trajectories are affected by macro-economic trends as well as intrinsic portfolio structures. Demonstrated

consistency with statistical relevance techniques (Chapter 8) exhibits viability for financial applications.

Socio-economic Applications Beyond healthcare and finance, analysis on socio-economic datasets reveals new linkages between public health indicators and social determinants like lifestyle behaviors or built infrastructure access (Chapter 9). The connections hint at potential risk factors to address systemically.

Impact Together, these cross-disciplinary demonstrations reveal how selectively explaining parts of the model or features augments transparency and trust in AI systems applied, while preserving predictive accuracy essential for adoption (Chapter 10). Methodological milestones expand the XAI toolkit for precision medicine, financial risk, policy decisions and related areas where algorithmic explainability is indispensable.

Limitations and Future Work Looking ahead, CAFA's technical refinements like handling data heterogeneity and categorical variables as well as tailored application spanning sectors with stakeholder needs represent fruitful directions (Chapter 10). Moreover, emerging XAI innovations around contrastive, counterfactual and interactive explanations can further amplify the promise and utility of interpretable ML.

In summary, this thesis strengthens the foundations for building reliable, transparent AI through contributions in XAI techniques plus cross-domain evidence. The dividends over the longer term remain enhanced accountability, acceptability and democratization of transformative technologies.

1.5 Thesis Structure

This thesis is organized into four parts. The first three parts are each structured around a specific application area: Medical, Finance, and Socio-Economic domains. The fourth part consists of a general discussion and a conclusion. All application areas utilize XAI techniques, specifically SHAP or its refinements. This is preceded by introductory chapters that provide the necessary background and context for the research.

Chapter 1 presents the motivation behind the research, highlighting the importance of XAI techniques in promoting transparency and elucidating the behaviors of complex predictive models across diverse domains. It also outlines the main contributions of the thesis, lists the published works, and provides an overview of the thesis structure.

Chapter 2 offers a comprehensive overview of the key concepts and techniques involved in the research. It discusses machine learning models, including linear regression, decision trees, random forests, and XGBoost, and introduces explainable artificial intelligence algorithms such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). The chapter also covers evaluation metrics for classification and regression tasks.

Part I focuses on XAI in the medical domain. Chapter 3 presents the criteria and rationale for dataset selection, an overview of the chosen datasets (Lung Cancer

and UCI Breast Cancer), and initial observations and analysis. Chapter 4 provides a comprehensive case study on the COVID-19 Non-Pharmaceutical Control Measures Dataset, highlighting the significance and background of the dataset, the implementation of XAI techniques, and the insights and conclusions drawn from the analysis. Chapter 5 introduces the novel Controllable fActor Feature Attribution (CAFA) approach, its algorithm, and its application to medical datasets and the COVID-19 dataset. Chapter 6 presents the Uncertainty-based Controllable Factor Feature Attribution (UCAFA) method, an extension of CAFA that incorporates uncertainty quantification using Variational Autoencoders (VAEs).

Part II concentrates on XAI in the finance domain. Chapter 7 provides an introduction to the domain, background information, aims, and contributions. It also describes the datasets used in the analysis, including data collection and the macro-finance and fund-level variables considered. Chapter 8 presents a comprehensive analysis of global equity funds, comparing Probit Regression with the XGBoost model, discussing XAI results based on input features, examining the influence of international diversification, and conducting robustness tests.

Part III explores XAI in the socio-economic domain. Chapter 9 presents the criteria and rationale for dataset selection, an overview of the chosen datasets (UCI Adult Income, German Credit, and ProPublica's COMPAS), and initial observations and analysis. It also demonstrates the application of CAFA to these socio-economic datasets, providing interpretations and insights.

Part IV concludes the thesis with a discussion and conclusion. Chapter 10 offers an interpretative analysis of the results across various datasets, discusses the real-world implications and potential of CAFA, and addresses the constraints and assumptions of the CAFA model. It recapitulates the principal discoveries and explores the impending applications and influence of CAFA in different domains.

Chapter 2

Background

Contents

This chapter establishes the key concepts and techniques that form the foundation of this thesis. We begin by discussing inherently interpretable machine learning models and their limitations in capturing complex relationships. We then introduce black-box models, which are essential for understanding the complex models explored later in the thesis. To bridge the gap between the predictive power of black-box models and the need for interpretability, we review popular explainable artificial intelligence (XAI) algorithms. Furthermore, we outline the evaluation metrics used to assess the performance and effectiveness of machine learning models and XAI algorithms in both classification and regression tasks.

2.1 Machine Learning Models

Machine learning models can be broadly categorized into two types: inherently interpretable models and black-box models. Popular inherently interpretable models, such as linear regression and decision trees, are often referred to as white-box models due to their transparent nature and the ease with which their decision-making processes can be understood. On the other hand, popular black-box models, such as random forests and XGBoost, are complex and opaque, making it challenging to interpret their internal workings [Mol23].

2.1.1 Linear Regression

Linear regression [MPV21] is a fundamental and inherently interpretable machine learning algorithm used for predicting continuous numerical outcomes. It models the

relationship between the input features and the target variable as a linear combination of the feature values.

Mathematical Framework

Given a dataset with n instances and p predictor variables, a linear regression model can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \tag{2.1}$$

where:

- y is the target variable
- $x_1, \ldots, x_p$ are the predictor variables
- $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients
- ε is the error term representing the unexplained variance

The goal of linear regression is to find the values of the coefficients that minimize the sum of squared residuals between the predicted and actual target values.

Interpretability

Linear regression is considered an inherently interpretable model due to its simplicity and the direct relationship between the input features and the target variable. The regression coefficients β_i represent the change in the target variable for a one-unit change in the corresponding predictor variable x_i, holding all other variables constant. This allows for a straightforward interpretation of the impact of each feature on the predicted outcome. Additionally, the significance of each predictor variable can be assessed using statistical tests, such as t-tests or F-tests, to determine whether the variable has a significant effect on the target variable. This further enhances the interpretability of the model.

Limitations

Despite its interpretability, linear regression has some limitations:

- It assumes a linear relationship between the predictor variables and the target variable, which may not always hold in real-world scenarios.
- It is sensitive to outliers, as they can heavily influence the regression coefficients.
- It may not capture complex non-linear relationships or interactions between variables.

In cases where the assumptions of linearity are violated or more complex relationships exist, other models such as decision trees or ensemble methods may be more appropriate.

2.1.2 Decision Trees

Decision trees [Qui86] are another class of interpretable machine learning models used for both classification and regression tasks. They recursively partition the feature space into subsets based on the most informative features, creating a tree-like structure of decision rules.

Mathematical Framework

A decision tree consists of internal nodes, branches, and leaf nodes. Each internal node represents a feature, and each branch emanating from a node corresponds to a possible value or range of values for that feature. The leaf nodes represent the predicted class or numerical value for the target variable. The construction of a decision tree involves selecting the best feature and split point at each node based on a criterion such as information gain, Gini impurity, or mean squared error. The process is repeated recursively until a stopping criterion is met, such as reaching a maximum depth or a minimum number of instances in a leaf node.

Interpretability

Decision trees are highly interpretable due to their rule-based nature. The path from the root node to a leaf node represents a series of decision rules based on the feature values. These rules can be easily understood and communicated to stakeholders, making decision trees a popular choice for domains where interpretability is crucial. Moreover, the feature importance can be derived from a decision tree by aggregating the reduction in impurity or error achieved by each feature across all the nodes where it is used. This provides insights into the relative significance of each feature in the decision-making process.

Limitations

While decision trees offer interpretability, they also have some drawbacks:

- They can be prone to overfitting, especially when the tree becomes deep and complex.
- They may struggle with capturing complex non-linear relationships or interactions between features.
- Small changes in the training data can lead to significant changes in the tree structure, making them unstable.

To address these limitations, ensemble methods such as random forests and gradient boosting, which combine multiple decision trees, are often used to improve predictive performance while maintaining some level of interpretability.

2.1.3 Random Forest

Random Forest [Bre01] is a versatile machine learning algorithm that can be used for both classification and regression tasks. It is an ensemble learning method, where the combined predictions of several base estimators usually decision trees lead to a more accurate and stable model.

Mathematical Framework

A Random Forest model consists of a collection of decision tree predictors $\{h(x, \Theta_k), k = 1, ..., K\}$ where x is the input vector, Θ_k are independently and identically distributed random vectors, and each tree casts a unit vote for the most popular class at input x in classification or average prediction in regression.

The ensemble prediction for a classification or regression problem is represented as:

$$H(x) = \frac{1}{K} \sum_{k=1}^{K} h(x, \Theta_k) \tag{2.2}$$

In this ensemble, K represents the number of trees, $h(x, \Theta_k)$ is the prediction of the k-th tree, and $H(x)$ is the final output of the Random Forest algorithm.

Random Forest introduces randomness in two ways: by bootstrapping the sample and by selecting a random subset of features at each split. This randomness helps in creating a diverse set of trees and is crucial for the robustness of the algorithm.

Training Procedure

During training, Random Forest creates each tree from a different sample of the data. This process, known as bootstrap aggregating or bagging, involves selecting a random subset of the training set with replacement. Each tree is grown to the largest extent possible without pruning, which means that the individual trees are deep and can capture complex structures in the data. For tree k, the variable Θ_k represents the randomness in the tree construction process and is used to generate the bootstrap sample and the random feature selection.

Algorithmic Enhancements

Random Forest algorithm has several key features that distinguish it from other learning algorithms:

- It can handle a large number of input variables without variable deletion.
- It provides an internal unbiased estimate of the generalization error as the forest building progresses.
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

- It offers methods for balancing error in class population unbalanced data sets.
- The computations can be performed in parallel to speed up the training process.

Random Forests tend to avoid overfitting problems that can occur with decision trees, making them a more reliable and robust algorithm for many applications. The diversity among the individual trees in the ensemble makes the Random Forest model less sensitive to the noise in the training data, and the bootstrapping method helps in reducing variance and retaining the bias.

The Random Forest algorithm is inherently suited for multiclass problems and can be applied to large datasets efficiently. Its ability to provide feature importance scores inherently is another reason for its widespread popularity in practical applications.

2.1.4 eXtreme Gradient Boosting (XGBoost)

XGBoost [CG16], standing for eXtreme Gradient Boosting, represents an advanced and efficient implementation of gradient boosting algorithms. It is extensively utilized in a variety of machine learning challenges, recognized for its superior performance and operational efficiency.

Mathematical Framework

XGBoost builds upon an ensemble of decision trees, formulated in an additive fashion. Given a dataset containing n samples with m features, denoted as $\{(x_i, y_i)\}_{i=1}^{n}$, where x_i is the feature vector and y_i is the corresponding target value, the predictive model for an individual sample is represented as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{2.3}$$

In this expression, $\hat{y}_i$ denotes the predicted outcome for the i-th instance, f_k symbolizes the k-th decision tree, K signifies the total number of trees, and $\mathcal{F}$ encompasses the space of all potential decision trees. The objective function that XGBoost endeavors to minimize integrates a loss component L and a regularization aspect Ω, articulated as:

$$\text{Obj}(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.4}$$

Here, Θ encapsulates the model's parameters, l is a convex loss function that quantifies the discrepancy between the predicted and actual values, and Ω inflicts a penalty on model complexity.

The regularization term is specifically delineated as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2.5}$$

with T indicating the leaf count within the tree, w_j representing the score on the j-th leaf, γ embodying the complexity cost per leaf, and λ being the L2 regularization on the leaf scores.

Training Procedure

The XGBoost algorithm employs a greedy strategy for tree construction and utilizes a quantile sketch approach to manage sparse data during tree learning. Trees are sequentially constructed with each new tree targeting the correction of residuals produced by preceding trees.

Notably, XGBoost is equipped to adeptly process missing data. During the learning phase, data instances missing values are allotted a default direction at each node, a decision derived from the data itself.

Algorithmic Enhancements

The XGBoost algorithm introduces numerous advancements over traditional gradient boosting techniques:

- Inclusion of a regularization term to mitigate overfitting, thereby enhancing model generalization.
- Adoption of a block structure for the data matrix, which promotes cache-aware access patterns and bolsters computational performance.
- Parallelized construction of trees to fully leverage the capabilities of multi-core processing architectures.
- Intelligent management of missing data through an automatic learning mechanism that determines the most favorable direction for missing values at each tree split.
- An effective tree-pruning strategy that employs a depth-first approach and excises branches with minimal contributions to predictive outcomes.

The robustness of XGBoost can be ascribed to its scalability, enabling it to manage vast datasets, and its capacity to unravel intricate nonlinear relationships within the data. Its adaptability and accuracy have cemented its status as a favored algorithm in the machine learning domain, especially for applications where precision in prediction is essential.

2.1.5 Question of interpretability

While inherently interpretable models like linear regression and decision trees offer transparency and ease of understanding, they may not always capture complex relationships and interactions in the data. On the other hand, black-box models like random forests and XGBoost can achieve high predictive performance by leveraging ensembles

of decision trees and advanced optimization techniques. However, their complexity and lack of transparency make it challenging to interpret their decision-making process. To bridge the gap between the predictive power of black-box models and the need for interpretability, XAI algorithms have emerged. These algorithms aim to demystify the inner workings of complex models and provide insights into how they arrive at their predictions. In the following section, we will explore two popular XAI algorithms, LIME and SHAP, which can be applied to black-box models to enhance their interpretability.

2.2 Explainable Artificial Intelligence Algorithms

2.2.1 Local Interpretable Model-agnostic Explanations (LIME)

In the domain of Explainable Artificial Intelligence (XAI), **Local Interpretable Model-agnostic Explanations (LIME)** [**RSG16**] is a pioneering approach that facilitates a detailed understanding of the decision-making process behind individual predictions made by complex machine learning models. This section elucidates the fundamental mechanics of LIME, shedding light on its capability to demystify the operations of high-dimensional and often non-transparent models.

Principles of Local Approximation

The foundational principle of LIME is that it is possible to locally approximate the decision surface of a complex model, which may be too intricate to understand in its entirety. Focusing on a constrained neighborhood around a point of interest, LIME constructs a simple model that mimics the complex model's behavior in that specific region. This surrogate model is more interpretable and provides insight into the reasoning of the complex model for a given prediction.

Local vs. Global Interpretability

While global interpretability entails comprehending a model's decision-making process across all inputs, LIME specializes in *local interpretability*, which is concentrated on explaining individual predictions. This is particularly crucial as numerous advanced machine learning models, such as those based on deep learning, are too convoluted to be globally interpreted. LIME overcomes this by homing in on an individual data point and elucidating the prediction it generates.

Simpler Models for Interpretation

To facilitate interpretation, LIME utilizes simpler models like linear regressions or decision trees to approximate the decision boundary of the complex model within the local scope. These models are deemed interpretable as their decision-making rules are easily comprehensible. For example, in a linear model, each feature is attributed a coefficient indicating its impact on the outcome. Positive coefficients signify a feature's propensity to influence the model's prediction toward one classification, while negative

coefficients suggest an influence in the opposite direction. This straightforwardness enables users to discern the most pivotal features for a specific prediction.

The Process of Generating Local Explanations

1. **Perturbation:** LIME initiates the process by perturbing the input data, generating a multitude of similar yet varied instances. These form a dataset that embodies the local feature space surrounding the focal instance.

2. **Model Predictions:** The predictions of the complex model on this newly created perturbed dataset are procured, offering a glimpse into the model's behavior in proximity to the instance under examination.

3. **Weight Assignment:** Weights are assigned to each perturbed instance contingent on their similarity to the original instance, generally by leveraging a kernel function. Proximal instances are given higher weights, rendering them more significant in the construction of the local explanation.

4. **Local Model Training:** A simple model is trained on the weighted perturbed data to emulate the complex model's decision boundary. This training prioritizes the accurate reflection of the complex model's local behavior.

5. **Interpretation:** The interpretative model's coefficients serve as a decipherable representation of the complex model's decision process for the particular instance. These coefficients are often presented as a list elucidating the contribution of each feature to the prediction.

The figure below visually delineates the LIME process and is a representative illustration of the various stages involved:

1. The green dot marks the original instance within the feature space, around which the model's behavior is to be investigated.

2. The blue to red points represent perturbed samples generated in the vicinity of the original instance. The intensity of their color signifies their weight, with red denoting higher weight (indicating closeness to the original instance) and blue indicating lower weight (signifying distance from the original instance). These weights are computed based on the proximity of each perturbed sample to the original instance, typically utilizing a kernel function.

3. The **black line** symbolizes the decision boundary of the local interpretable model trained by LIME. This boundary is an abstract representation of the complex model's decision boundary within the local confines surrounding the original instance. The direction in which the original instance lies relative to this line reflects the predicted category for the original instance by the local model.

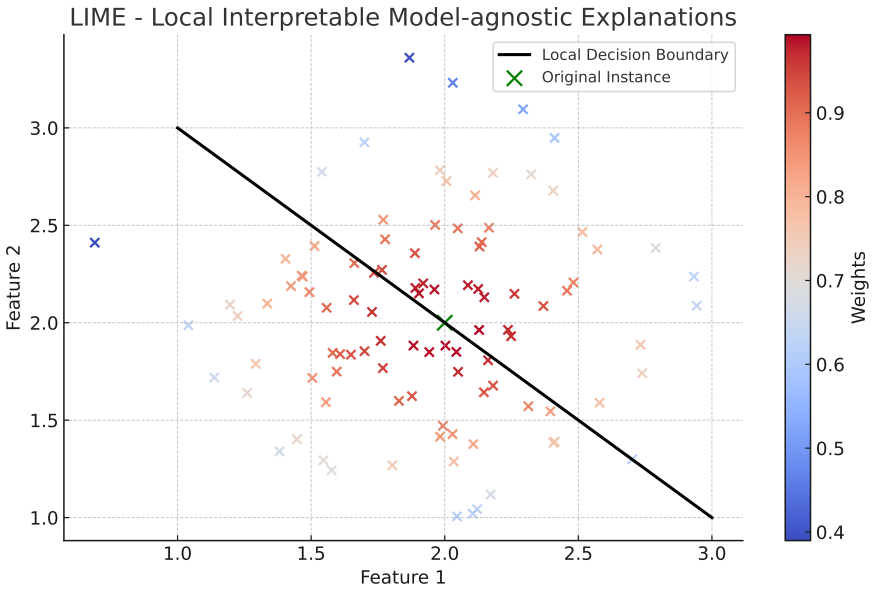


Figure 2.1: Illustration of the LIME process.

The illustration above provides a visual overview of the LIME methodology. To further understand the process, we will now examine the mathematical framework that underlies this approach.

The Mathematical Framework of LIME

The principal concept of LIME is to locally approximate the prediction function f of a complex model around the vicinity of a particular instance x that needs to be explained. The approach involves the following steps:

1. **Sampling**: Generate a new dataset composed of perturbed samples around x.
2. **Weighting**: Assign weights to these new samples based on their proximity to x.
3. **Model Fitting**: Fit an interpretable model g on the dataset, considering the weights assigned.
4. **Explanation**: Utilize the interpretable model g to elucidate the prediction at the instance x.

Sampling and Weighting

Upon selection of an instance x, LIME generates a collection of perturbed samples $\{x'_1, x'_2, \ldots, x'_n\}$ and acquires the corresponding predictions $f(x')$ from the complex

model. To gauge the locality, LIME calculates a proximity measure $\pi_x(x')$ which assigns more significant weights to samples in closer proximity to x.

The proximity measure is defined through an exponential kernel as follows:

$$\pi_x(x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \tag{2.6}$$

where σ is a hyperparameter that dictates the kernel's width, and $\|x - x'\|$ is a metric for the distance between the instance x and the sample x'.

Model Fitting

A transparent model g is then constructed to fit the dataset with the goal of minimizing the loss function delineated below, which quantifies how well g approximates f within the locality defined by π_x:

$$\mathcal{L}(f, g, \pi_x) = \sum_{x', y'} \pi_x(x')(f(x') - g(x'))^2 \tag{2.7}$$

Herein, g is commonly a linear model for tabular or textual data, and a decision tree for image data. The selection of g is pivotal as it embodies the balance between simplicity, thereby interpretability, and fidelity to f.

Explanation

After the training of g, it serves to explicate the prediction at x. In the context of linear models, this explanation is provided in terms of feature importance:

$$g(x) = w_0 + \sum w_i x_i \tag{2.8}$$

where w_0 represents the intercept, and w_i denotes the coefficients linked with features x_i of the instance x. The magnitude and sign of w_i indicate the contribution of each feature towards the prediction of x.

In summary, LIME serves as an instrumental technique for elucidating the inner workings of complex predictive models. It achieves this by constructing interpretable approximations for individual predictions, thus rendering the model's decisions more transparent and comprehensible. This attribute of LIME is particularly valuable when the stakes are high and the need for trust and clarity in machine learning outputs is paramount. As the landscape of machine learning continues to expand, the relevance of methods like LIME is only set to increase, bridging the gap between algorithmic performance and human interpretability.

2.2.2 SHapley Additive exPlanations (SHAP)

In recent years, the deployment of machine learning models has seen a rapid expansion across a variety of sectors. These models have been integral in driving decisions

ranging from personalized medical treatments to financial forecasting. However, as the complexity of these models increases, so does the difficulty in comprehending their decision-making processes. This presents a challenge in situations where understanding the rationale behind a model's predictions is crucial for ethical, legal, and practical reasons.

SHAP [LL17] method designed to bring transparency to the predictions of machine learning models. SHAP utilizes the Shapley value—a concept from the cooperative game theory—to fairly attribute the output of a model to its input features. Each feature's contribution is measured, considering its interaction with other features, providing a comprehensive view of its influence on the model's predictions. This explanation technique is particularly important for ensuring that machine learning models are used responsibly. By clarifying how and why decisions are made, SHAP helps users to trust and effectively manage the outputs of complex algorithms. This is essential in high-stakes domains where the consequences of decisions can have profound implications.

Additionally, SHAP is model-agnostic, meaning it can be applied to any machine learning model, from linear regression to deep neural networks. This versatility makes SHAP an invaluable tool in the machine learning toolkit for researchers and practitioners alike, seeking to build models that are not only powerful but also interpretable. As the demand for interpretable machine learning continues to grow, SHAP provides a key to unlocking the "black box", paving the way for more accountable and understandable AI systems. The next section will delve into the mathematical foundations of SHAP, elucidating how it quantitatively attributes significance to model features and thus serves as a cornerstone for interpretability in machine learning.

Mathematical Framework for SHAP

The mathematical foundation of SHAP is rooted in the concept of the Shapley value [Sha53] from cooperative game theory. The Shapley value is a solution concept that offers one way to distribute the total gains achieved by the coalition of all players (features in the context of SHAP) fairly among them.

Formally, the Shapley value of feature i in a coalition is defined as the average marginal contribution of feature i across all possible permutations of the features. Mathematically, it is expressed as:

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left(f_x(S \cup \{i\}) - f_x(S)\right) \tag{2.9}$$

where:

- N is the set of all features.
- S is a subset of features excluding feature i.
- $\phi_i(f, x)$ is the Shapley value for feature i.
- $f_x(S)$ is the prediction of the model using the features in set S.

- $|S|$ is the number of features in subset S.
- $|N|$ is the total number of features.
- The sum is taken over all subsets S that do not include feature i.

This formula computes the contribution of feature i by averaging its impact on the prediction over all possible combinations of features. The term $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ in the Shapley value formula 2.9 represents the weights assigned to each feature subset when calculating the SHAP values. These weights ensure that each feature's contribution is fairly distributed across all possible subsets of features, giving more importance to the subsets where the feature's presence or absence has a more significant impact on the model's prediction.

To demonstrate this concept, consider a simple example with three features: Age (A), Income (I), and Credit Score (C). In this case, the total number of features, N, is 3. To calculate the SHAP value for the Age (A) feature, we need to compare the model's predictions for the following pairs of feature subsets:

- $\{\}$ vs. $\{A\}$, where $S = \{\}$ and $|S| = 0$
- $\{I\}$ vs. $\{A, I\}$, where $S = \{I\}$ and $|S| = 1$
- $\{C\}$ vs. $\{A, C\}$, where $S = \{C\}$ and $|S| = 1$
- $\{I, C\}$ vs. $\{A, I, C\}$, where $S = \{I, C\}$ and $|S| = 2$

Suppose we have a model that predicts the probability of a person defaulting on a loan. Let's focus on the first pair of subsets, $\{\}$ vs. $\{A\}$, and assume the model makes the following predictions:

- $f(\{\}) = 0.5$
- $f(\{A\}) = 0.6$

The marginal contribution of the Age (A) feature for this pair of subsets is calculated as:

$$f(\{A\}) - f(\{\}) = 0.6 - 0.5 = 0.1$$

This marginal contribution is then weighted by $\frac{|S|!(|N|-|S|-1)!}{|N|!}$, which in this case is:

$$\frac{0!(3-0-1)!}{3!} = \frac{0!2!}{3!} = \frac{1}{3}$$

The SHAP value for Age (A) is calculated by taking the weighted average of the marginal contributions across all pairs of feature subsets. This process is repeated for the other features, Income (I) and Credit Score (C), to obtain their respective SHAP values. The resulting SHAP values indicate the influence of each feature on the model's

prediction for a given instance, taking into account the interactions between features. The SHAP values provide a detailed understanding of each feature's contribution to the model's output, allowing for a more interpretable and transparent explanation of the model's behavior.

In practice, this calculation can be computationally intensive for models with a large number of features, prompting the development of efficient algorithms for approximating SHAP values. The essence of SHAP lies in its ability to provide a detailed and fair attribution of the prediction output, which aligns with the properties of the Shapley value including efficiency, symmetry, null player, and additivity. These properties ensure that the contributions of the features sum up to the actual prediction, are independent of the order of the features, give no importance to features that do not change the prediction, and allow for the decomposition of the model prediction in a linear fashion, respectively.

Properties of SHAP

SHAP values are defined to satisfy four fundamental properties from game theory, which, in the context of machine learning, translate into desirable attributes for feature importance explanations.

1. **Efficiency**: The sum of the SHAP values for all features equals the difference between the prediction of the model and the average prediction across all data points.
$$\sum_{i=1}^{N} \phi_i = f(x) - \mathbb{E}[f(X)] \tag{2.10}$$

2. **Symmetry**: If two features contribute equally to all possible combinations of feature subsets, then their SHAP values are equal.
$$\text{If } f_x(S \cup \{i\}) = f_x(S \cup \{j\}) \text{ for all subsets } S \subseteq N \setminus \{i, j\}, \text{ then } \phi_i = \phi_j \tag{2.11}$$

3. **Dummy**: If a feature does not change the prediction for any possible combination of features, then its SHAP value is zero.
$$\text{If } f_x(S \cup \{i\}) = f_x(S) \text{ for all subsets } S \subseteq N, \text{ then } \phi_i = 0 \tag{2.12}$$

4. **Additivity**: For any two models f and g, the SHAP value for the combined model $h = f + g$ is the sum of the SHAP values for f and g.
$$\phi_i(h, x) = \phi_i(f, x) + \phi_i(g, x) \tag{2.13}$$

The Efficiency property guarantees that SHAP values account for all the contributions to the output, leaving no unexplained variance. Symmetry ensures that features with the same contribution receive the same attribution, preventing bias towards any particular

feature. The Dummy property ensures that irrelevant features, which do not affect the prediction, do not receive any undue credit. Lastly, Additivity allows for the decomposition of model explanations across additive components, which is particularly useful for models that are themselves additive or ensemble-based. Together, these properties ensure that the SHAP values provide a reliable and justifiable explanation of the model's predictions, reflecting the contribution of each feature to the output in a way that is consistent with the overall behavior of the model.

Variants of SHAP

The diversity of machine learning models, each with unique structural characteristics and computational requirements, necessitates different approaches to explanation. This is the primary reason for the development of various SHAP methodologies. While the core principle of attributing model output to individual features remains consistent, the method of computation of SHAP values must be adapted to the architecture of the model for efficiency and fidelity.

Kernel SHAP

Kernel SHAP is a model-agnostic method that uses a specially weighted local linear regression to estimate SHAP values for any model. Here is the formula that expresses how the SHAP value is approximated for feature i:

$$\phi_i \approx \sum_{S \subseteq N \setminus \{i\}} w(S) \left[f_x(S \cup \{i\}) - f_x(S) \right], \tag{2.14}$$

where $w(S)$ is the weight assigned to the subset S and is determined by the Kernel SHAP algorithm. Kernel SHAP is particularly useful when dealing with models for which no specialized SHAP computation method has been developed. It can be computationally intensive, so it's generally used when the number of features is not too large or when interpretability is prioritized over computational efficiency.

Tree SHAP

Tree SHAP is an optimized version of SHAP for tree-based models, such as decision trees, random forests, gradient boosting machines, and XGBoost. It exploits the tree structure to compute exact SHAP values efficiently. The general formula for SHAP values in tree-based models considers the conditional expectations as the model's predictions:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \left[f_x(S \cup \{i\}) - f_x(S) \right], \tag{2.15}$$

However, Tree SHAP optimizes this computation by using the tree structure to evaluate these differences without needing to enumerate all subsets S.
Tree SHAP is most appropriate when you are using a tree-based machine learning model and need to compute SHAP values quickly, especially when dealing with a large number of features or needing to explain many predictions.

Deep SHAP

Deep SHAP extends the ideas from SHAP to deep learning models. It is based on a combination of DeepLIFT (an existing method for attributing neural network predictions to inputs) and SHAP. It computes approximate SHAP values for deep learning architectures. It is designed to handle the complex architectures of deep neural networks, making it suitable when working with high-dimensional data and models such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs).

Linear SHAP

Linear SHAP is specifically designed for linear models. It is efficient because it takes advantage of the additive structure of linear models. The SHAP value for a feature in a linear model is simply the feature's value multiplied by its corresponding coefficient:

$$\phi_i = x_i \cdot \beta_i, \tag{2.16}$$

where x_i is the feature value and β_i is the coefficient associated with feature i.
Linear SHAP is best used with linear regression models, logistic regression, and any other model where the prediction is a linear combination of the input features.

In practice, the choice of SHAP variant is dictated by the model type. Kernel SHAP offers wide applicability, Tree SHAP brings efficiency to tree-based models, Deep SHAP caters to the nuanced architectures of deep learning, and Linear SHAP is ideal for linear relationships. Ultimately, the objective is to illuminate the model's decision-making process, enabling the development of more transparent and trustworthy AI systems.

SHAP Visualizations

Interpreting complex machine learning models can be a daunting task, especially when trying to understand the influence of individual features on model predictions. SHAP visualizations play a pivotal role in bridging the gap between high-dimensional model data and human-readable interpretations. These visualizations translate the SHAP values into various comprehensible formats, enabling stakeholders to quickly identify which features are most influential for a model's predictions. In this subsection, we explore different types of SHAP visualizations, each tailored to present the model's behavior in an informative and accessible manner. To illustrate the concepts discussed in this section, we generated a toy example using a simple XGBoost classifier trained on a synthetic dataset with five features: Feature 1, Feature 2, Feature 3, Feature 4, and Feature 5.

Summary Plot

The SHAP summary plot provides a global view of feature importance's and their effects on the model output. It aggregates the SHAP values across the entire dataset to identify patterns and outliers in feature attributions, offering a broad perspective on the model's decision-making process.

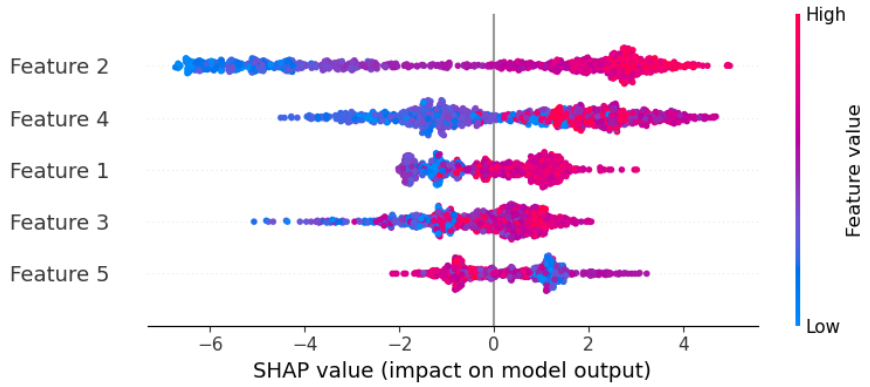


Figure 2.2: SHAP Summary Plot

In the summary plot (Figure 2.2), the x-axis represents the SHAP value, which indicates the impact of a feature on the model's output. Features are sorted along the y-axis based on their overall importance, with the most important features appearing at the top. Each point represents an individual data instance, and its color corresponds to the feature value, with red indicating high values and blue indicating low values.

Force Plot

The SHAP force plot breaks down the contribution of each feature for a single data instance, providing a granular understanding of individual predictions. It visually depicts how each feature's value pushes the model output from a base value, typically the mean model output over the dataset, towards the actual prediction.

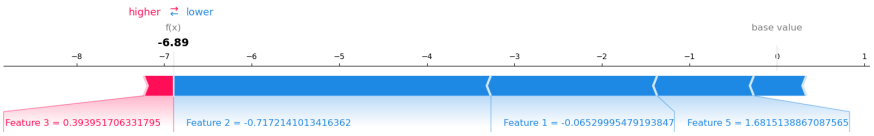


Figure 2.3: SHAP Force Plot

In the force plot (Figure 2.3), the x-axis represents the model output value, and the plot shows how each feature contributes to the final prediction. The base value (the mean model output) is shown as a gray line, and the colored bars represent the impact of each feature. Red bars indicate features that push the prediction higher, while blue bars indicate features that push the prediction lower. The width of each bar corresponds to the magnitude of the feature's contribution.

Dependence Plot

Dependence plots show the effect of a single feature across the whole dataset, revealing potential interactions between features. By plotting the SHAP value of a feature against its actual value for all instances, this visualization can suggest the presence of non-linear relationships.

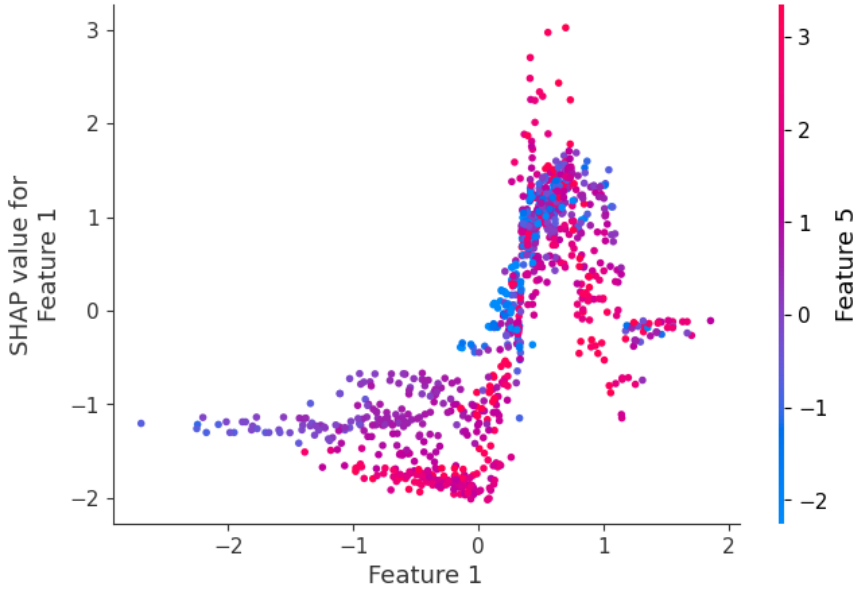


Figure 2.4: SHAP Dependence Plot

In the dependence plot (Figure 2.4), the x-axis represents the actual feature value, while the y-axis represents the SHAP value for that feature. Each point in the plot corresponds to an individual data instance. The plot shows how the model's output changes as the feature value varies, allowing us to identify non-linear relationships and potential interactions with other features.

Waterfall Plot

The waterfall plot sequentially shows the cumulative impact of each feature on an individual prediction. Starting from the base value, it adds or subtracts the effect of features in descending order of importance, culminating in the final prediction.

In the waterfall plot (Figure 2.5), the x-axis represents the model output value, and the y-axis lists the features in descending order of importance. The base value (the mean model output) is shown as the starting point, and each feature's contribution is added or subtracted from this value, leading to the final prediction. Red bars indicate features that increase the prediction, while blue bars indicate features that decrease the prediction. Through these visualizations, SHAP not only provides insights into the feature contributions but also facilitates a deeper understanding of the underlying

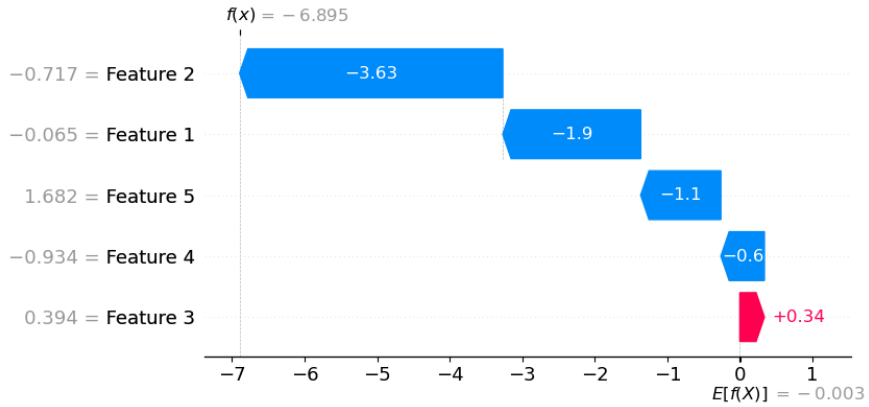


Figure 2.5: SHAP Waterfall Plot

model mechanics. Such clarity is indispensable for model validation, troubleshooting, and ensuring that the predictions align with real-world expectations.

In conclusion, SHAP offers a powerful and flexible framework for interpreting complex machine learning models. By grounding its approach in cooperative game theory, SHAP provides a rigorous methodology for feature importance attribution that is both fair and consistent across different types of models. The various SHAP methods, including Kernel SHAP, Tree SHAP, Deep SHAP, and Linear SHAP, cater to a wide range of models from simple linear regressions to complex tree-based ensembles and deep neural networks. This versatility ensures that practitioners can apply SHAP to virtually any machine learning problem, thereby demystifying model predictions and fostering greater trust and transparency in AI systems. Furthermore, the visualizations generated by SHAP, such as summary plots and force plots, serve as intuitive tools for both technical and non-technical stakeholders to grasp the reasoning behind model predictions. As the field of machine learning continues to advance, the interpretability provided by SHAP will remain invaluable, ensuring that our models remain comprehensible, accountable, and aligned with ethical standards.

2.3 Evaluation Metrics

2.3.1 Classification

In machine learning, classification tasks predict discrete outcomes by assigning data points to predefined categories. The quality of these predictions is typically measured using an array of metrics such as precision, recall, F1 score, accuracy, and the Receiver Operating Characteristic (ROC) curve. Let TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. Then the following are the definitions of precision, recall, F1 score, accuracy and ROC curve:

Precision

Precision measures the proportion of predicted positive instances that are actually positive. Mathematically, it is defined as the ratio of true positives to the total number of positive predictions

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2.17}$$

Recall

Recall measures the proportion of actual positive instances that are correctly predicted as positive. Mathematically, it is defined as the ratio of true positives to the total number of actual positive instances

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2.18}$$

F1 score

F1 score is the harmonic mean of precision and recall. It is a balanced measure that takes both precision and recall into account. Mathematically, it is defined as

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.19}$$

Accuracy

Accuracy measures the proportion of correct predictions among all predictions. Mathematically, it is defined as the ratio of the total number of correct predictions to the total number of predictions

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{2.20}$$

ROC curve

The Receiver Operating Characteristic (ROC) curve is a tool for evaluating binary classifiers, plotting the true positive rate (TPR) against the false positive rate (FPR) at various thresholds. The TPR, also known as sensitivity, and the FPR, referred to as 1-specificity, are calculated as follows

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2.21}$$

where TP is the number of true positives and FN is the number of false negatives. The FPR is defined as

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{2.22}$$

The area under the ROC curve (AUC) measures the model's discriminative power, with 1 indicating perfect classification and 0.5 representing a random guess.

2.3.2 Regression

Regression models are fundamental in the field of machine learning, predicting continuous outcomes based on the relationships learned from input features. Evaluating the performance of these models is crucial to ensuring their reliability and validity in practical applications. A variety of metrics are established to assess regression models, each focusing on different aspects of prediction accuracy. In this subsection, we discuss some of the most widely used regression evaluation techniques.

Mean Absolute Error (MAE)

The Mean Absolute Error is a measure of the average magnitude of errors in predictions. It is calculated by taking the mean of the absolute differences between the actual values and the predicted values

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\,, \tag{2.23}$$

where y_i represents the true value, $\hat{y}_i$ the predicted value, and n the number of samples.

Mean Squared Error (MSE)

Mean Squared Error measures the average of the squares of the errors, giving more weight to larger errors, which can be particularly useful when large errors are undesirable

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{2.24}$$

Root Mean Squared Error (RMSE)

The Root Mean Squared Error is the square root of the MSE, providing a metric in the same units as the response variable, often making it more interpretable than MSE

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{2.25}$$

R-squared (R^2)

R-squared, or the coefficient of determination, quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables. It is an indicator of the goodness of fit of the model

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{2.26}$$

where $\bar{y}$ is the mean of the actual values.

Adjusted R-squared

Adjusted R-squared is a modified version of R^2 that adjusts for the number of predictors in the model. It accounts for the fact that R^2 will always increase with the addition of more predictors

$$\text{Adjusted } R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}, \tag{2.27}$$

where p represents the number of predictors.

Explained Variance Score

The explained variance score quantifies the proportion of variance in the dataset that is accounted for by the model

$$\text{Explained Variance} = 1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)}. \tag{2.28}$$

Summary

This section provides a comprehensive overview of the key concepts and techniques involved in this thesis. It begins by discussing the different types of machine learning models, including inherently interpretable models like linear regression and decision trees, as well as black-box models such as random forests and XGBoost. The chapter highlights the trade-off between interpretability and predictive performance, emphasizing the need for XAI algorithms to bridge this gap. The chapter then delves into two popular XAI algorithms, LIME and SHAP, which aim to provide explanations for the predictions of black-box models. LIME focuses on local interpretability by approximating the decision surface of a complex model in a local neighborhood, while SHAP assigns importance values to each feature based on their contribution to the model's output using the concept of Shapley values from cooperative game theory.

Finally, the chapter introduces various evaluation metrics used to assess the performance of machine learning models in both classification and regression tasks. These metrics provide a quantitative way to measure the effectiveness of the models and compare different algorithms. Overall, the background section lays a solid foundation for understanding the key concepts, techniques, and evaluation methods used in this research, setting the stage for the novel contributions and insights presented in the subsequent chapters.

Part I

XAI in the Medical Domain

Chapter 3

Datasets and Preliminary Analysis

Contents

3.1 Criteria and Rationale for Dataset Selection

The criteria for dataset selection in this research are meticulously designed to ensure the development of robust and ethically sound AI models, particularly in the healthcare domain. Key considerations include the suitability of the datasets for binary classification tasks, enabling precise evaluation with standard machine learning metrics. The datasets must exhibit a blend of continuous and categorical features, both intrinsic and adjustable, to aptly demonstrate the capabilities of the models. Emphasis is placed on choosing datasets from high-impact domains like healthcare, where explainability is crucial for societal benefit. The size of the datasets is also a critical factor, being large enough for training complex models yet manageable for rapid prototyping. Accessibility is prioritized, with a preference for publicly available datasets to ensure reproducibility. Moreover, data quality, completeness, ethical compliance, and privacy considerations are paramount, alongside the need for the datasets to be representative of diverse populations. Preference is given to datasets with a history of use in research for benchmarking purposes and those that hold cross-disciplinary relevance, thus bridging the gap between AI development and practical healthcare applications.

Expanding upon this overview, the subsequent paragraph comprehensively delves into each criterion, emphasizing the specific factors influencing the dataset selection process. This detailed examination highlights the rigor and meticulousness in selecting datasets most appropriate for research in Explainable Artificial Intelligence (XAI) within

the medical domain, focusing on its applications and relevance.

- **Prediction Tasks**: Classification problems with binary target variables to allow precise quantitative evaluations using standard machine learning metrics.
- **Feature Diversity**: A mix of continuous and categorical features with both intrinsic and adjustable variables to demonstrate partitioning capabilities.
- **Domain Relevance**: Real-world datasets from high-impact domains like healthcare where explainability offers significant societal value.
- **Data Volume**: Number of instances suitably large for effectively training complex ML models but also small enough to rapidly prototype algorithms.
- **Accessibility**: Publicly available open datasets having little or no restrictions to maximize reproducibility.
- **Data Quality and Completeness**: Emphasis on datasets with comprehensive and accurate information, minimal missing values, and well-documented data collection methods [HMC21].
- **Ethical Compliance and Privacy Considerations**: Selection of datasets adhering to ethical guidelines and privacy laws, including patient consent and anonymization of personal identifiers [RHH+18, PON19].
- **Representative Diversity**: Datasets should represent diverse populations, covering various demographics to ensure equitable and wide applicability of AI models [RHH+18].
- **Previous Usage and Validation**: Preference for datasets previously used in research, allowing for benchmarking and contextual evaluation of AI model performance.
- **Cross-Disciplinary Relevance**: Datasets should be relevant not only to AI development but also to clinicians and healthcare practitioners, ensuring real-world applicability [GORB20, CRG17].

Guided by these criteria, two open medical datasets hosted trusted repositories were selected:

- Simulacrum Lung Cancer Data [PNA]
- UCI Breast Cancer Data [ZS88]

The rationale for choosing clinical oncology datasets include:

- Running initial evaluations on cancer prediction tasks underscores the eventual societal benefits of deploying more explainable AI in the healthcare domain.

- The variables cover a breadth of intrinsically uncontrollable demographic factors along with numerous adjustable diagnostic, treatment and outcome attributes. This range suits demonstrating capabilities of algorithmically partitioning features.

- Statistical generalization of results does not comprise the main priority at this exploratory stage. Instead, the focus lies in effectively illustrating functionality on apt and impactful datasets before subsequent expansion.

3.2 Overview of Chosen Datasets

3.2.1 Lung Cancer Dataset

This research utilizes the lung cancer data from the Simulacrum project by Health Data Insight CiC as an initial case study for evaluating explainable AI techniques. The Simulacrum dataset accurately reflects properties of real-world data from National Cancer Registration and Analysis Service (NCRAS) while protecting patient confidentiality[1].

The lung cancer data contains 2,242 instances mapped across 24 input features spanning demographics, diagnoses, treatments and outcomes Descriptions of the feature categories are:

- **Demographic Factors**: Includes 'age' in years at diagnosis, 'gender', and 'ethnic background' coded via high-level census categories

- **Cancer Staging**: Comprises TNM classification sub-stages, tumor grade descriptors, and morphological codes quantifying disease progression

- **Treatment Attributes**: Encodes surgery, chemotherapy, radiation and regimen specifics like drugs, cycles, delays, etc.

- **Outcomes**: Captures best response, treatment toxicity, protocol deviations, vital status after 12 months etc. as bins.

The prediction modeling task is formulated as inferring the 12-month vital status of patients as a binary classification problem. The target variable has classes encoding whether the patient was alive or deceased one year after diagnosis.

Overall, the granularity and span of real-world features around lung cancer coupled with the ability to predict survival outcomes makes this an ideal dataset for evaluating explainable AI techniques in a clinical healthcare setting. The richness and heterogeneity of variables can help demonstrate how algorithmic approaches can provide more targeted and actionable insights to medical practitioners.

[1]Description of features used in this dataset can be found at the Cancer Registration Data Dictionary and the SACT Data Dictionary, with links available at:
`https://simulacrum.healthdatainsight.org.uk/available-data/table-descriptions/`.

3.2.2 UCI Breast Cancer Dataset

The second dataset selected for this research is the Breast Cancer dataset from the UCI Machine Learning Repository. This dataset aligns with the criteria outlined in Section 3.1, offering a complementary perspective to the Simulacrum Lung Cancer dataset.

The UCI Breast Cancer dataset comprises 286 instances, each characterized by 9 features. The task is formulated as a binary classification problem to predict cancer recurrence events, with the target variable indicating either "recurrence-events" or "no-recurrence-events". The features capture a mix of intrinsic patient attributes and adjustable diagnostic measurements:

- **Uncontrollable Features**: age, menopause
- **Controllable Features**: tumor size, inv-nodes, node-caps, deg-malig, breast, breast-quad, irradiate

This compact dataset focuses on key prognostic indicators, providing a targeted case study to demonstrate the partitioning capabilities of explainable AI techniques, particularly the proposed CAFA approach introduced in Chapter 5. The clear distinction between intrinsic and adjustable features aligns well with CAFA's emphasis on classifying dataset features into 'controllable' and 'uncontrollable' categories for more actionable insights.

From an explainability standpoint, this dataset allows for a concise illustration of how algorithmic transparency can highlight the influence of specific biomarkers and treatment factors on predicting cancer recurrence events. Evaluating techniques on this binary classification task complements the analysis on the lung cancer dataset, underscoring the versatility of explainable AI in enhancing clinical decision-making across different oncological domains, as emphasized in the abstract.

Moreover, the UCI Breast Cancer dataset has been widely utilized in machine learning research [AMAMN16, SAZ12] aligning with the selection criterion of choosing benchmark datasets. This enables contextual evaluation of the developed techniques, including CAFA, against existing approaches. Overall, the inclusion of the UCI Breast Cancer dataset strengthens the robustness and generalizability of this research.

3.3 Initial Observations and Analysis

3.3.1 Lung Cancer Dataset

To gain initial insights into predicting the 12-month mortality outcome from the Lung Cancer dataset, we experimented with various machine learning techniques. These included Logistic Regression, XGBoost, Random Forest Classifier, Support Vector Machines (SVM), and Neural Network Classification. Table 3.1 presents the performance metrics of these algorithms.

The results demonstrate that the Random Forest classifier achieves the highest performance across all metrics, with an accuracy of 95%, precision of 96%, recall of 95%,

Algorithm	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.89	0.91	0.87	0.89	0.88
XGBoost	0.93	0.93	0.92	0.92	0.92
Random Forest	0.95	0.96	0.95	0.95	0.95
SVM	0.92	0.93	0.90	0.91	0.91
Neural Network	0.93	0.92	0.93	0.93	0.92

Table 3.1: Performance comparison of different algorithms for the Lung Cancer dataset

F1-score of 95%, and an area under the receiver operating characteristic curve (AUC) of 0.95. XGBoost and Neural Network Classification also exhibit strong performance, with accuracies of 93% and 93%, respectively. Logistic Regression and SVM, while still providing good results, have lower performance compared to the tree-based ensemble methods and neural networks. Logistic Regression achieves an accuracy of 89%, while SVM reaches an accuracy of 92%.

The superior performance of the Random Forest classifier can be attributed to its ability to effectively capture complex relationships and interactions among the features, handle both continuous and categorical variables, and its robustness to outliers. XGBoost, another tree-based ensemble method, also demonstrates strong predictive power by leveraging gradient boosting techniques. Neural Network Classification, with its ability to learn intricate patterns and representations from the data, also achieves high performance. However, the interpretability of neural networks can be challenging compared to tree-based methods like Random Forest, which provide inherent feature importance measures [GMR+18]. Logistic Regression and SVM, while widely used for binary classification tasks, may have limitations in capturing complex non-linear relationships present in the Lung Cancer dataset [CMC+19].

Based on these comparisons and considering the interpretability aspects, the Random Forest classifier remains a suitable choice for our analysis of the Lung Cancer dataset. Its high predictive performance, coupled with its ability to provide feature importance measures, aligns well with our goal of gaining insights into the factors influencing 12-month mortality predictions.

To optimize the performance of the Random Forest model, we conducted a grid search over various hyperparameters, including the number of trees, maximum depth, and minimum samples per leaf. The best configuration achieved an impressive accuracy of 97% on the test set, demonstrating the model's effectiveness in predicting 12-month mortality.

Table 3.2 presents the detailed performance metrics of the optimized Random Forest model, including precision, recall, and F1-score for the 12-month mortality prediction task. While the Random Forest model achieved high accuracy, it is crucial to understand the factors contributing to its predictions. To gain insights into the importance of each feature, we applied the Tree SHAP (SHapley Additive exPlanations) technique introduced in Chapter 2. SHAP is a unified approach to interpreting model predictions by assigning importance values to each feature based on their contribution to the model's

Metric	Precision	Recall	F1-score
12-month Mortality	0.98	0.96	0.97

Table 3.2: Performance metrics for the Lung Cancer dataset using Random Forest

output.

Figure 3.1 presents the SHAP global explanation for the Lung Cancer dataset as a violin plot. The plot visualizes the distribution of SHAP values for each feature, with features ranked in descending order of importance. Higher SHAP values indicate a greater influence on the model's predictions.

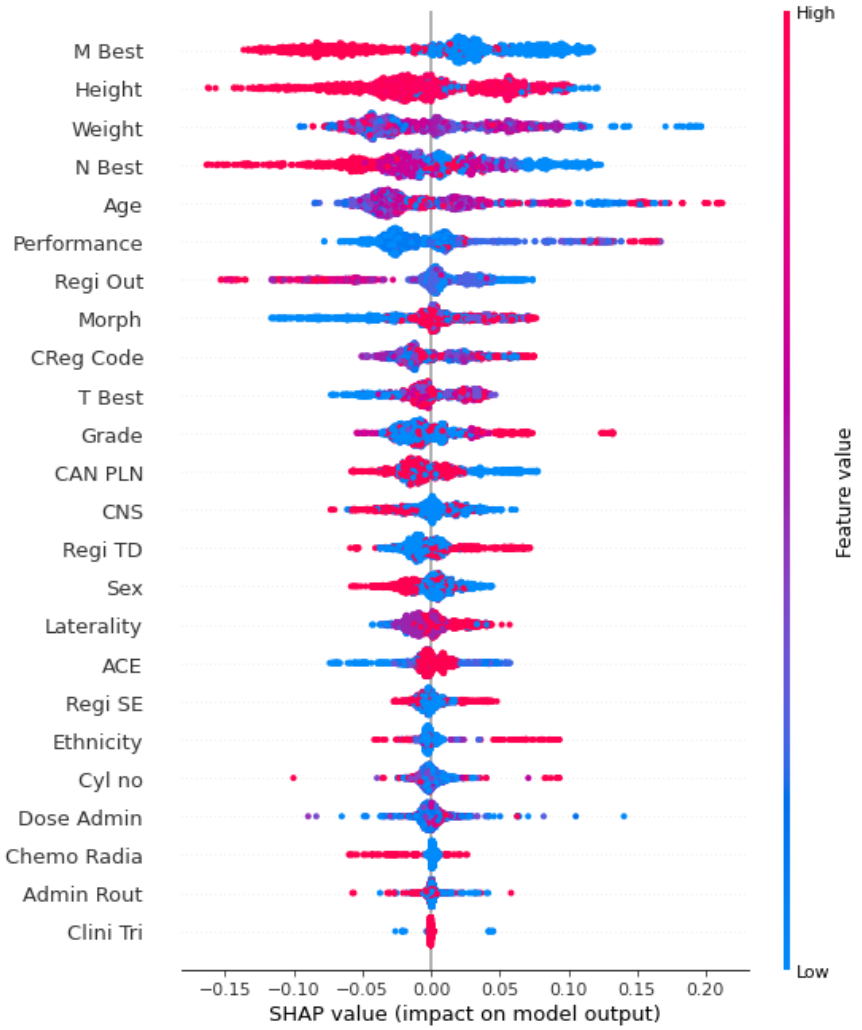


Figure 3.1: SHAP global explanation for the Lung Cancer dataset

From the SHAP global explanation, we observe that features such as TNM classi-

fication sub-stages, age, and tumor grade have a significant impact on the 12-month mortality predictions. These findings align with domain knowledge, as advanced cancer stages, older age, and higher tumor grades are known to be associated with poorer prognosis [GCC+16].

These initial observations and analyses provide a solid foundation for further exploration of the Lung Cancer dataset using explainable AI techniques. The proposed Controllable fActor Feature Attribution (CAFA) approach, introduced in Chapter 5, will build upon these insights to offer a more refined understanding of the impact of controllable and uncontrollable factors on lung cancer mortality predictions.

By focusing on features that can be directly influenced or modified, CAFA aims to provide actionable insights for clinical decision-making and targeted interventions. The application of CAFA to the Lung Cancer dataset will be discussed in detail in Section 5.3, showcasing its potential to advance explainable AI in the healthcare domain.

3.3.2 UCI Breast Cancer Dataset

To gain initial insights into the UCI Breast Cancer dataset, we employed various machine learning algorithms to predict breast cancer recurrence events. The algorithms considered include Logistic Regression, XGBoost, Random Forest, Support Vector Machines (SVM), and Neural Network Classification. Table 3.3 presents the performance metrics of these algorithms.

Algorithm	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.68	0.70	0.66	0.68	0.67
XGBoost	0.73	0.75	0.71	0.73	0.72
Random Forest	0.77	0.79	0.75	0.77	0.76
SVM	0.75	0.77	0.73	0.75	0.74
Neural Network	0.67	0.69	0.65	0.67	0.66

Table 3.3: Performance comparison of different algorithms for the UCI Breast Cancer dataset

The results demonstrate that the Random Forest classifier achieves the highest performance among the compared algorithms, with an accuracy of %, precision of 79%, recall of 75%, F1-score of 77%, and an AUC of 0.76. XGBoost and SVM also exhibit relatively good performance, with accuracies of 73% and 75%, respectively. Logistic Regression and Neural Network Classification provide lower performance, with accuracies of 68% and 67%, respectively.

The superior performance of the Random Forest classifier can be attributed to its ability to handle the mix of features, capture non-linear relationships, and its robustness to outliers. XGBoost, with its gradient boosting technique, demonstrates good performance on the dataset. SVM, known for its ability to find optimal decision boundaries, also provides relatively good results.

Logistic Regression and Neural Network Classification, while commonly used for binary classification tasks, exhibit lower performance compared to the other algorithms. This suggests that the relationships present in the UCI Breast Cancer dataset may be more complex and require more advanced algorithms to capture them effectively.

Based on these comparisons and considering the interpretability aspects, the Random Forest classifier is selected as the primary model for further analysis and interpretation of the UCI Breast Cancer dataset. Its high predictive performance and inherent feature importance measures align well with our goal of gaining insights into the factors influencing breast cancer recurrence predictions.

To further optimize the performance of the Random Forest classifier, we performed a grid search over hyperparameters. The grid search resulted in an optimized Random Forest model with an improved accuracy of 79%, precision of 81%, recall of 77%, F1-score of 79%, and an AUC of 0.78. This enhancement in performance highlights the importance of tuning the model's hyperparameters to better capture the underlying patterns in the dataset.

We applied the Tree SHAP technique to the trained Random Forest model to understand the importance of each feature in predicting cancer recurrence events. Figure 3.2 displays the SHAP global explanation as a violin plot, showcasing the distribution of SHAP values for each feature. The features are ranked in descending order of importance, with higher SHAP values indicating a greater influence on the model's predictions.

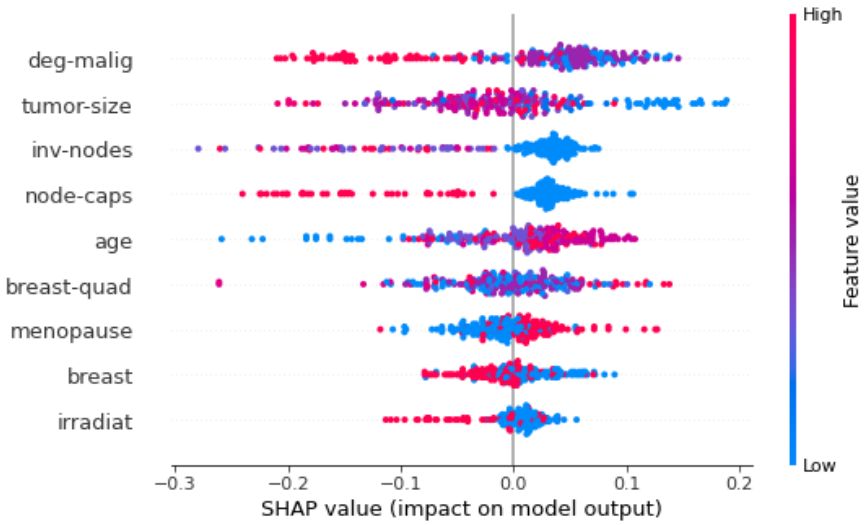


Figure 3.2: SHAP global explanation for the UCI Breast Cancer dataset

These initial observations and analyses provide a foundation for further exploration of the datasets using explainable AI techniques, such as the proposed Controllable fActor Feature Attribution (CAFA) approach, to gain more actionable insights into the factors influencing cancer outcomes.

Summary

The chapter focuses on the selection and preliminary analysis of two medical datasets for evaluating explainable AI techniques in healthcare: the Simulacrum Lung Cancer dataset and the UCI Breast Cancer dataset. The criteria for dataset selection are outlined, emphasizing factors such as prediction tasks, feature diversity, domain relevance, data volume, accessibility, quality, ethical compliance, representative diversity, previous usage, and cross-disciplinary relevance. The Lung Cancer dataset contains 2,242 instances with 24 features, while the Breast Cancer dataset comprises 286 instances with 9 features. Both datasets are used for binary classification tasks to predict cancer outcomes. Various machine learning algorithms, including Logistic Regression, XGBoost, Random Forest, SVM, and Neural Networks, are applied to the datasets. The Random Forest classifier achieves the highest performance on both datasets, with accuracies of 97% and 79% for the Lung Cancer and Breast Cancer datasets, respectively.

The chapter also introduces the use of the Tree SHAP technique to interpret the importance of each feature in the trained models. The SHAP global explanations reveal that features such as TNM classification sub-stages, age, and tumor grade have a significant impact on lung cancer mortality predictions, while specific features influence breast cancer recurrence predictions.

The preliminary analysis sets the stage for further exploration of the datasets using explainable AI techniques, particularly the proposed Controllable fActor Feature Attribution (CAFA) approach, which aims to provide more actionable insights by focusing on controllable and uncontrollable factors influencing cancer outcomes.

Chapter 4

Case Study: COVID-19 Non-Pharmaceutical Control Measures Dataset

Contents

4.1 Significance and Background of the Dataset

With the outbreak of the COVID-19 pandemic in December 2019, many countries have implemented some non-pharmaceutical control measures to contain the spread of the virus in the absence of effective vaccination and treatment. In this case study, we use CAFA to study the effectiveness of the non-pharmaceutical control measures implemented in the UK.

We formulate the effectiveness of control measures as an XAI modelling problem. We focus on studying the relationship between control measures and the daily reproduction rate R_t. R_t is one of the most important metrics used to measure the epidemic spread. A value greater than 1 suggests the epidemic being expanding; a value less than 1 indicates shrinking. We employ the approach presented in [FMG+20] for estimating R_t from daily infection cases. We then pose the following classification problem:

> *Given non-pharmaceutical control measures applied on a specific day, predict whether R_t is smaller or greater than 1 on that day.*

By solving this prediction problem with a classifier, we can identify the control measures that contribute most significantly to the prediction. Analyzing the behavior of the prediction model provides insights into the effectiveness of these control measures.

No.	Region	Total Data Points	Timeline
1	East Midlands	352	21/02/2020 – 06/02/2021
2	East of England	347	26/02/2020 – 06/02/2021
3	London	362	11/02/2020 – 06/02/2021
4	North East	342	02/03/2020 – 06/02/2021
5	North West	345	28/02/2020 – 06/02/2021
6	South East	368	05/02/2020 – 06/02/2021
7	South West	362	11/02/2020 – 06/02/2021
8	West Midlands	343	01/03/2020 – 06/02/2021
9	Yorkshire and Humber	374	30/01/2020 – 06/02/2021
10	Northern Ireland	370	03/02/2020 – 06/02/2021
11	Scotland	345	28/02/2020 – 06/02/2021
12	Wales	346	27/02/2020 – 06/02/2021

Table 4.1: Summary of Raw Data Set

UK Data Collection

Total number of daily and cumulative case, deaths and tests were collected from the website developed and published by the public health England on behalf of government of United Kingdom (UK)[1]. Data on Non-pharmaceutical Control Measures were obtained manually based on UK's policies on different entities like Meeting Indoors, Meeting Outdoors, Domestic Travel, International Travel, Cafes and Restaurants, Pubs and Bars, Sports and leisure, Hospitals/Care and Nursing Home Visits, Non-essential shops, School Closures from available website references by Wikipedia[2] and published news articles. Control Measures are coded based on level of severity (e.g., 'High', 'Moderate', 'Low') for all control measures excluding Non-essential shops and School closures and (e.g., 'Opened', 'Closed') for Non-essential shops and School closures, as represented in Table 4.2 and Table 4.3. In addition, data points for Temperature and Humidity were extracted from all weather data available from Reliable Prognosis website [Ltd20]

Notably, total 4256 data points were collected up to 06/02/2021 across different regions in UK for e.g., England (East Midlands, East of England, London, North East, North West, South East, South West, West Midlands, Yorkshire and Humber), Northern Ireland (NI), Scotland and Wales as shown in Table 4.1

Evaluate Rate of Infection (R_t) from Data set

In our research, we adapted the following research [FMG+20] where Rt is calculated based on daily number of confirmed cases and to model the time between the person tested positive and successive positive person is serial interval distribution. As presented

[1]COVID-19 Dashboard (UK): `https://coronavirus.data.gov.uk`

[2]For example, for Wales the control measure data has been collected from `https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic_in_Wales`

	Severity Level		
	High	**Moderate**	**Low**
Meeting Indoors	No outside households allowed (except extended)	Limited indoor meeting	No restrictions
Meeting Outdoors	Leave house only with valid reason (e.g., work, food)	Meet 1 person outside for exercise	No restrictions
Domestic Travel	Stay Local (within 5 miles)	Avoid unnecessary travel	No restrictions
International Travel	Non-essential travel banned	Travel allowed with quarantine (10-14 days)	No restrictions
Cafes and Restaurants	Closed (Takeaway allowed)	Time-restricted dine-in (Takeaway allowed)	Full services allowed
Pubs and Bars	Shutdown	Seating only, time / alcohol restrictions	No restrictions
Sports and Leisure	Outdoors closed, gyms shutdown	Limited outdoor sports, individual exercise	No restrictions
Hospitals/Care Visits	Shutdown	Seating only, alcohol restrictions	No restrictions

Table 4.2: Levels of Severity of each Non-pharmaceutical Control Measures

in [FMG+20] we used distribution be a Gamma distribution g with mean 7 and standard deviation 4.5 for all regions at all time.

The discrete convolution function for number of new infections c_t on a given day t is given by

$$c_t = R_t \sum_{\tau=0}^{t-1} c_\tau \; g_{t-\tau} \tag{4.1}$$

where c_τ is the number of new infection on day τ. From equation 4.1, we can evaluate value of R_t by

$$R_t = \frac{c_t}{\sum_{\tau=0}^{t-1} c_\tau \; g_{t-\tau}} \tag{4.2}$$

In equation 4.2, c_t and c_τ are available directly from the data set.For suppose $x = t, \tau \text{and} c_x$ can be calculated by subtracting the number of confirmed positive cases

	Non-Essential Shops	School Closures
Opened	Clothing, electrical, furniture and Beauty salons, tattooists, nail bars, spas, tanning shops are opened with respect to health and Safety Rules	Nurseries , Primary and Secondary Schools opened with respect to health and safety rules
Closed	All Non-essential shops Closed	Nurseries, Primary and Secondary Schools Closed except for vulnerable children, children's of key workers and special school children

Table 4.3: Coding for Non-essential shops and School Closures

on day x and the number confirmed positive cases on day $x-1$. We can able to obtain $g_t - \tau$ by integrating the Gamma distribution as

$$g_\kappa = \int_{\tau=\kappa-0.5}^{\kappa+0.5} g(\tau)d\tau$$

for $\kappa = 2, 3, \ldots$ and

$$g_1 = \int_{\tau=0}^{1.5} g(\tau)d\tau$$

We illustrated the evaluated R_t for different regions in the UK in Figures 4.1 and 4.2. Specifically, England and Wales are presented in Figure 4.1, while Scotland and Northern Ireland (NI) are shown in Figure 4.2

Data Pre-Processing

The raw dataset comprises entries from the dashboard published for public access [Eng20] and manually coded data obtained from government policies published in news articles, which required further processing. In our research, we employed a three-stage approach for preprocessing. Firstly, we converted all values of non-pharmaceutical control measures (CM) from strings to numerical representations. For example, regions in the UK were coded with numerical values such as 'East Midlands': 1, 'East of England': 2, 'London': 3, and so forth up to 'Wales': 12. Control Measures were categorized based on severity ('Low': 1, 'Moderate': 2, 'High': 3) as trinary, and closures ('Opened': 0, 'Closed': 1) as binary. For instance, Meeting Indoors (MInd), Meeting Outdoors (MOut), Domestic Travel (DT), International Travel (IT), Cafes and Restaurants (CR), Pubs and Bars (PB), Sports and Leisure (SL), Hospitals/Care and Nursing Home Visits (HV) were assigned trinary values, while Non-essential Shops (NS) and School Closures (SC) were assigned binary values. Additionally, we discretized the values of Temperature (T) into four intervals and Humidity (H) into three intervals as shown: T:$(-\infty, 0), [0, 10), [10, 20), [20, \infty)$, H:$[0, 40), [40, 80), [80, \infty)$.

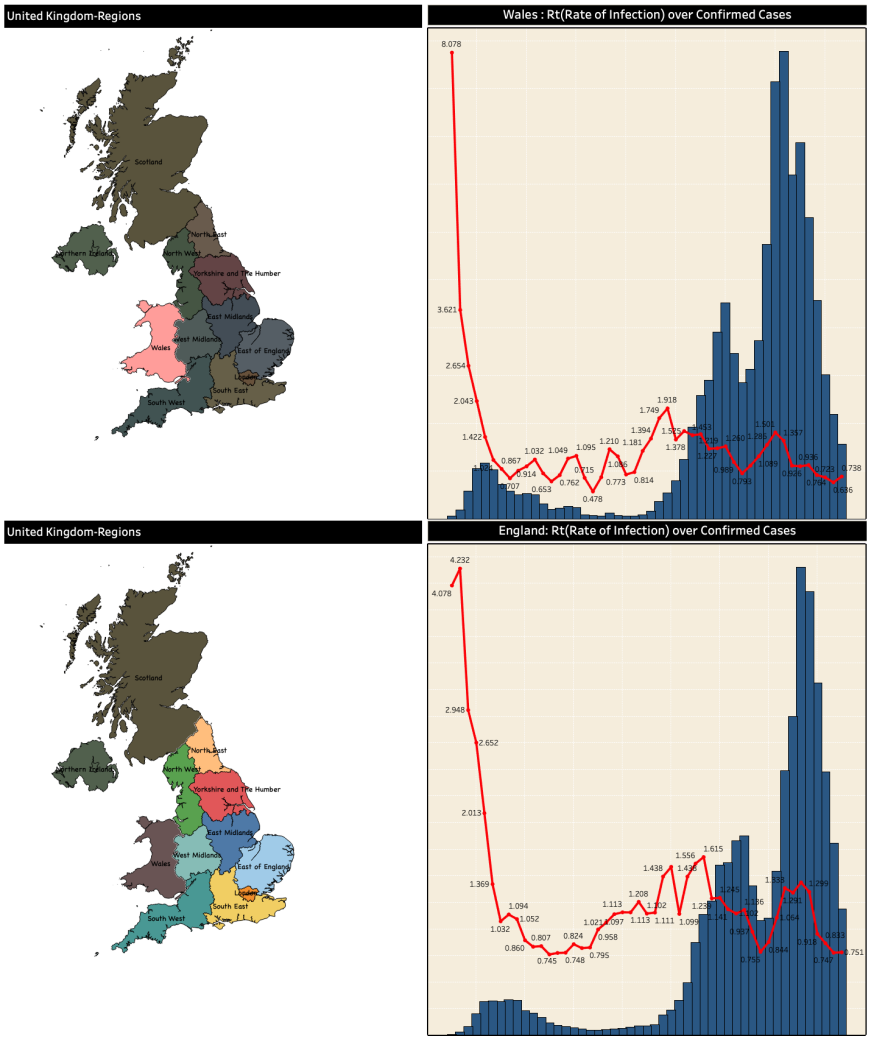


Figure 4.1: Rate of Infection(R_t) over Confirmed positive cases in Wales and England

Secondly, we processed the control measures by breaking down the trinary control measures into binary ones, such that each level of severity has its own column. For instance, in our dataset, we had a trinary control measure like MInd (Meeting Indoors), which accepts three different values: Low (L), Moderate (M), High (H). Through splitting, we converted this trinary measure into binary ones, resulting in individual control measures such as MInd_L, MInd_M, MInd_H. This approach was applied to all trinary control measures. Moreover, we discarded data points with confirmed positive cases fewer than the threshold of 20 in a region. As per equation 2, the evaluated R_t value assumes a reasonably large t; otherwise, both $c\tau$ and $g_{t-\tau}$ would be too small, resulting in an unnaturally large R_t. Furthermore, we replaced the existing entries in control measures (processed in Stages 1 and 2) with a count that increases sequentially until there is a change in policy regarding a specific control measure, at which point the count

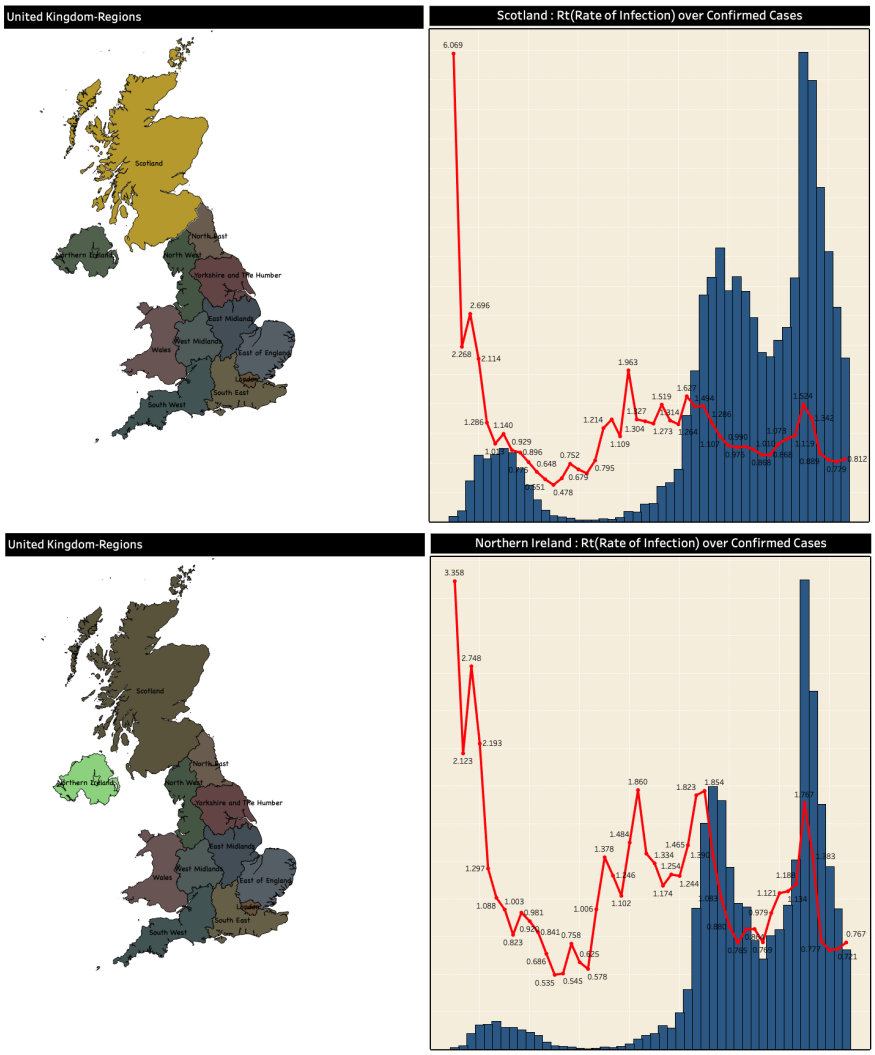


Figure 4.2: Rate of Infection(R_t) over Confirmed positive cases in Scotland and NI

resets. For example, MI(H)=12 means that 12 days have passed since implementing high restrictions on meeting indoors. Next, we evaluated the R_t value, which is sensitive to noise. For example, we identified an anomaly in the dataset, where fewer cases were reported on weekends compared to weekdays for unknown reasons. To address this, we used a sliding-window mean filter with a radius of 1 to mitigate the noise in the data. That is, for a day t, if its confirmed cases are C_t, then the filtered confirmed cases F_t would be calculated as $(Ct-1+C_t+Ct+1)/3$. The evaluated R_t value, after removing noise, is then added to each row in the dataset to be used as the target for prediction.

In the final stage, to obtain interpretable qualitative results, we further discretized the "confirmed cases", "deaths", "tests", and R_t, which were modified, appended, and

evaluated in the preprocessing stages. For features other than R_t, we used the K-means discretization technique, as the number of intervals is not straightforward to estimate. R_t was discretized into the following intervals: $\{[0, 1), [1, \infty)\}$. For instance, if $R_t = 0.83$, since $0 \leq 0.83 < 1$ and $[0, 1)$ is the first interval for discretizing R_t, $R_t = 0.83$ is mapped to 0. Given another entry, $R_t = 1.23$, since $1 \leq 1.23 < \infty$ and $[1, \infty)$ is the second interval, $R_t = 1.23$ is mapped to 1.

The data collection, pre-processing, and evaluation of the R_t described in this section lay the foundation for the subsequent application XAI techniques. By ensuring the dataset is properly structured, with relevant features and an accurately computed target variable, we can effectively apply XAI methods to gain insights into the effectiveness of non-pharmaceutical control measures. The following sections will build upon this foundation, using XAI to uncover the complex relationships between control measures and the spread of COVID-19 in the UK.

4.2 Implementing XAI on the COVID-19 Dataset

To gain insights into the effectiveness of non-pharmaceutical control measures implemented in the UK during the COVID-19 pandemic, we applied various machine learning algorithms to predict whether the daily reproduction rate R_t is smaller or greater than 1 on a specific day, given the control measures in place.

We compared the performance of several algorithms, including Logistic Regression, XGBoost, Random Forest, Support Vector Machines (SVM), and Neural Network Classification. Table 4.4 presents the accuracy scores of these algorithms on the COVID-19 dataset.

Algorithm	Accuracy
Logistic Regression	0.87
XGBoost	0.92
Random Forest	0.94
SVM	0.89
Neural Network	0.91

Table 4.4: Accuracy comparison of different algorithms for the COVID-19 dataset

The results demonstrate that the Random Forest classifier achieves the highest accuracy of 0.94, outperforming other algorithms. The superior performance of the Random Forest can be attributed to its ability to handle a mix of categorical and numerical features, capture non-linear relationships, and its robustness to outliers. Based on these results, we selected the Random Forest classifier as the primary model for further analysis and interpretation of the COVID-19 dataset. Its high predictive accuracy and inherent feature importance measures align well with our goal of understanding the impact of control measures on the daily reproduction rate.

To gain deeper insights into the contribution of each control measure to the model's predictions, we applied the Tree SHAP. SHAP assigns importance values to each feature based on their contribution to the model's output, providing a global explanation of the model's behavior.

Figure 4.3 presents the SHAP global explanation plot for the COVID-19 dataset. Each point in the plot represents an instance from the dataset, with the color indicating the feature value (blue for low values and red for high values). The x-axis shows the impact of each feature on the model's prediction, with points on the right side (positive SHAP values) indicating that the feature pushes the classifier towards predicting a higher probability of the desired outcome ($R_t < 1$), while points on the left side (negative SHAP values) indicate the opposite. The summary plot provides valuable insights into the relationship between different features and the model's predictions. Features with low values (represented by blue points) that appear on the right side of the plot are positively correlated with the desired outcome of keeping the rate of infection (R_t) less than 1. Conversely, features with high values (represented by red points) on the left side of the plot are negatively correlated with the desired outcome.

The SHAP summary plot effectively captures the complex relationships between features and the model's predictions, highlighting the factors that contribute to keeping R_t below the critical threshold of 1.

From the SHAP global explanation, we observe that control measures such as Cafes and Restaurants high restrictions, High restrictions on pubs and bars, and high number cases have a significant impact on the model's predictions.

These insights from the SHAP global explanations provide a foundation for understanding the effectiveness of non-pharmaceutical control measures in controlling the spread of COVID-19. In the next section, we will apply the proposed Controllable fActor Feature Attribution (CAFA) approach to further investigate the impact of controllable and uncontrollable factors on the model's predictions.

4.3 Drawn Insights and Conclusions

The analysis of the COVID-19 dataset using explainable AI techniques, particularly the SHAP global explanations, provides valuable insights into the effectiveness of non-pharmaceutical control measures in controlling the spread of the virus in the UK. The Random Forest classifier, with its high predictive accuracy, serves as a reliable model for understanding the impact of various control measures on the daily reproduction rate R_t. The SHAP global explanation reveals that control measures such as high restrictions on cafes and restaurants, high restrictions on pubs and bars, and a high number of cases have a significant influence on the model's predictions. This suggests that these measures play a crucial role in determining whether the daily reproduction rate R_t is smaller or greater than 1 on a specific day.

The insights gained from the SHAP global explanation align with the real-world understanding of the effectiveness of non-pharmaceutical interventions. Restrictions on social gatherings in indoor settings, such as cafes, restaurants, pubs, and bars, have been

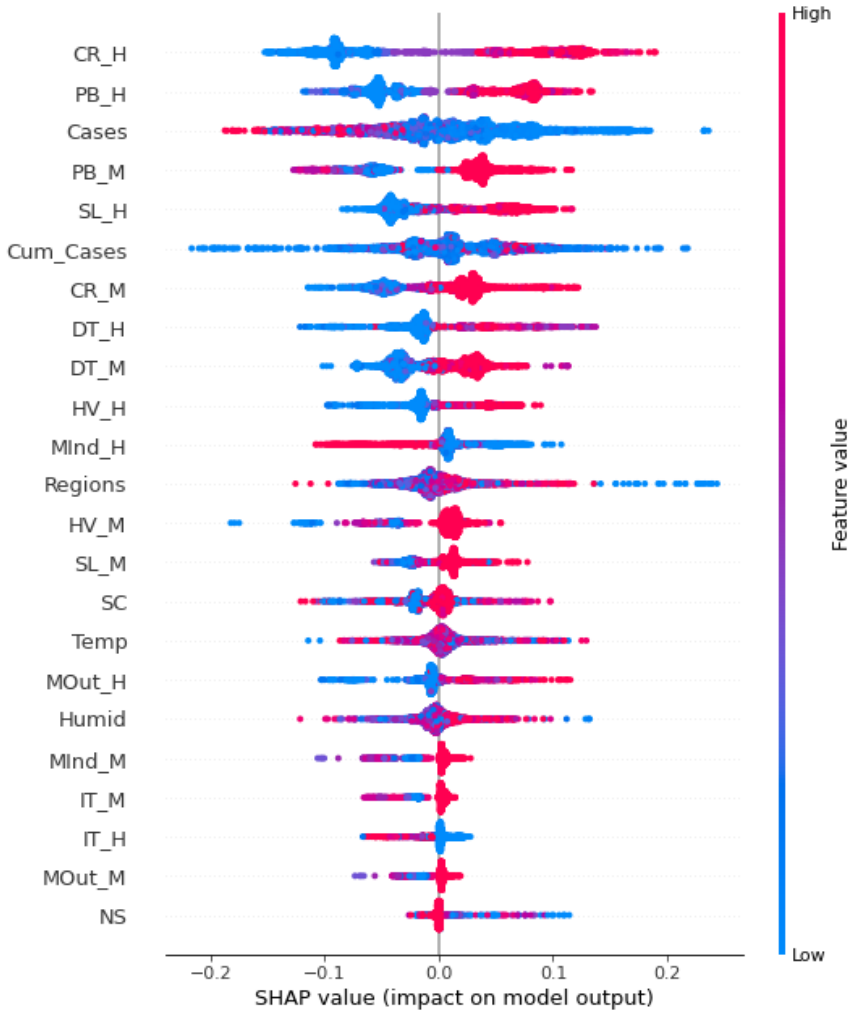


Figure 4.3: SHAP global explanation for the COVID-19 dataset

widely recognized as important measures in reducing the transmission of COVID-19 [FMG+20]. These settings often involve close contact among individuals, increasing the risk of virus spread. By imposing high restrictions on these venues, the UK government aimed to limit social interactions and curb the spread of the virus.

Moreover, the significance of the high number of cases in the model's predictions highlights the importance of monitoring and responding to the current state of the pandemic. A high number of cases indicates a greater prevalence of the virus in the community, which can lead to increased transmission and a higher reproduction rate. This insight emphasizes the need for timely and effective control measures to be implemented based on the current epidemiological situation. However, it is important to note that the SHAP global explanation provides an overall view of feature importance across the entire dataset. To gain a more granular understanding of the impact of control

measures on specific instances, local explanations can be employed. Local explanations, such as SHAP local explanations, can reveal the contribution of each control measure for individual predictions, allowing for a more nuanced analysis of their effectiveness in different scenarios.

Furthermore, the proposed Controllable fActor Feature Attribution (CAFA) approach can be applied to the COVID-19 dataset to distinguish between controllable and uncontrollable factors. By focusing on the controllable factors, such as the implementation of specific control measures, policymakers and public health officials can make informed decisions about which interventions to prioritize and how to allocate resources effectively.

In conclusion, the analysis of the COVID-19 dataset using explainable AI techniques provides valuable insights into the effectiveness of non-pharmaceutical control measures in controlling the spread of the virus in the UK. The SHAP global explanation highlights the importance of restrictions on social gatherings in indoor settings and the significance of monitoring the current state of the pandemic. These insights can guide policymakers in making data-driven decisions and implementing targeted interventions to mitigate the impact of COVID-19. Further analysis using local explanations and the CAFA approach can provide a more comprehensive understanding of the complex interactions between control measures and their impact on the daily reproduction rate.

Chapter 5

Controllable fActor Feature Attribution (CAFA)

Contents

5.1 Introduction

Feature attribution algorithms [LL17] are a popular class of Explainable AI (XAI) algorithms. Given a prediction instance, they tell the relative "importance" of each feature in the instance. In addition to "explaining" the prediction model, importance measures also reveal insight about the instance being explained, e.g., [ADG$^+$21] shows that XAI can help "generating the hypothesis about causality" in developing decision support systems. In this sense, feature attribution algorithms are considered as a data mining tool for extracting and discovering patterns in large datasets. For instance,[DFB$^+$21] uses feature attribution algorithms to understand important factors affecting cancer patient survivability; [F$^+$20] employs feature attribution algorithms to study factors affecting the transmission of SARS-CoV-2; and [L$^+$21] uses feature attribution to analyse factors affecting foreign exchange markets. However, existing feature attribution algorithms (see e.g., [AB18, Mol23, TG20] for overviews) treat all features homogeneously when computing their relative importances. Such homogeneity may not always give desirable interpretations when feature attribution algorithms are used for data mining purposes. Consider the following hypothetical example.

> Suppose we want to estimate the chance for some individual having breast cancer, with features like *age*, *gender*, *weight*, *alcohol intake*, *smoking habits*, *family history*, etc. A predictive model estimates the likelihood of the person having breast cancer; and a feature attribution algorithm gives attributions like *age: 0.3*, *gender: 0.13*, *weight: 0.27*, *alcohol intake: 0.15*, *smoking: 0.3*, *family history: 0.36*, etc.

From these calculated values, we notice that certain features, such as *age*, *gender* and *family history*, while being influential to the prediction, are *uncontrollable risk factors* [Dra06]. Knowing the relative importance of these features makes little contribution to clinical decision making. On the other hand, features representing *controllable risk factors* such as *weight*, *alcohol intake* and *smoking habits* are vital to clinical interventions [Dra06]. Thus, from an intervention perspective, it is necessary to distinguish these two classes of factors and compute their influences accordingly. We raise the question:

What are the influences of controllable factors used in a prediction?

To answer this question, a naive approach would be to build another predictive model, which only considers controllable factors, and apply feature attribution algorithms to that model. However, as explained in [KL21] and [ŽC16], dropping features from models can negatively impact the model performance as we will show in Table 5.1 in experimental study. Thus, instead of building models with fewer features, we suggest creating algorithms that are able to treat controllable factors differently from uncontrollable ones.

In this thesis, we present *Controllable fActor Feature Attribution (CAFA)*. Through *selective perturbation* and *global-for-local interpretation*, CAFA computes the relative importance of controllable factors for individual instances using prediction models built from all features. We apply CAFA on lung cancer data in Simulacrum[1] and on the UCI breast cancer dataset[2] to study the influence of controllable factors on survival time or recurrence. In a second experiment (see Section 5.5), we apply CAFA to data from a COVID-19 transmission case study (cf. also [HFL+21, KED+21]) to explore the effectiveness of non-pharmaceutical control measures.

5.2 CAFA Algorithm

CAFA computes feature importances for controllable factors through *selective perturbation* and *global-for-local interpretation*. Conceptually, CAFA is inspired by LIME such that a set of perturbed samples is generated to compute the feature importance. However, there are two main differences. Firstly, unlike LIME where the perturbation is carried out uniformly throughout all features, CAFA selectively perturbs features representing controllable factors. Secondly, with the dataset generated, instead of fitting

[1] Simulacrum is a dataset "developed by Health Data Insight CiC derived from anonymous cancer data provided by the National Cancer Registration and Analysis Service, which is part of Public Health England", `https://simulacrum.healthdatainsight.org.uk/publications/acknowledging-the-simulacrum/` .

[2] `http://archive.ics.uci.edu/ml/datasets/Breast+Cancer`

a weak interpretable model for computing explanations, a strong model is chosen to fit the dataset. We then determine the feature importance of controllable factors by using an explainer to compute the global explanation on the dataset. Fig. 5.1 illustrates CAFA's selective perturbation strategy.

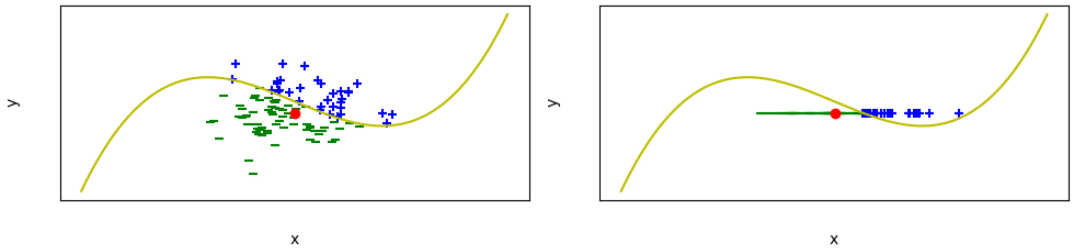


Figure 5.1: Selective Perturbation in CAFA. The point of interest (explanation point) and the generated dataset are shown in the figures. The red dot denotes the point of interest in a 2D space. The yellow curve is the decision boundary. Blue "+" and green "-" denote generated positive and negative samples, respectively. The figure on the left illustrates the standard perturbation (LIME), where both features x and y are perturbed; the figure on the right illustrates the selective perturbation (CAFA), where only the x axis, representing the controllable factor, is perturbed.

Given a prediction model f, for a data point $\mathbf{x}$ with m features partitioned into two sets F_c (*controllable*) and F_u (*uncontrollable*) such that $F_c \cap F_u = \{\}$, to compute feature importance for F_c, we construct a data set with n points

$$D_{\mathbf{x}} = \{(\mathbf{x}_1, f(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, f(\mathbf{x}_n))\}$$

such that for all $(\mathbf{x}_i, f(\mathbf{x}_i)) \in D_{\mathbf{x}}$, the following two conditions hold:

nolistsep $\delta(\mathbf{x}, \mathbf{x}_i) \leq \pi_{\mathbf{x}}$, where δ is a distance function and $\pi_{\mathbf{x}}$ is some proximity threshold, and

noliistsep for $\mathbf{x} = \langle v_1, \ldots, v_m \rangle$, and $\mathbf{x}_i = \langle v_1^i, \ldots, v_m^i \rangle$, for all j $(1 \leq j \leq m)$, it is the case that if feature j is in F_u, then $v_j = v_j^i$.

For two instances $\mathbf{x}_1 = \langle v_1^1, \ldots, v_m^1 \rangle$ and $\mathbf{x}_2 = \langle v_1^2, \ldots, v_m^2 \rangle$, the distance function $\delta(\mathbf{x}_1, \mathbf{x}_2)$ is

$$\delta(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{i=1}^{m} \omega_i d(v_i^1, v_i^2)}{\sum_{i=1}^{m} \omega_i}, \tag{5.1}$$

where ω_i is the weight of feature i and $d(v_i^1, v_i^2)$ is defined by[3]:

- if feature i is categorical, then

$$d(v_i^1, v_i^2) = \begin{cases} 0 & \text{if } v_i^1 = v_i^2, \\ 1 & \text{otherwise;} \end{cases} \tag{5.2}$$

[3]Note that we assume some standard normalization / scaling pre-processing is performed on the dataset so all continuous features take values in the range [0,1].

- if feature i is continuous, then

$$d(v_i^1, v_i^2) = |v_i^1 - v_i^2|. \tag{5.3}$$

We then build a strong prediction model g from $D_{\mathbf{x}}$ and calculate the global explanation $g(D_{\mathbf{x}})$ using SHAP by first computing local explanations for all instances in $D_{\mathbf{x}}$ and then averaging the results. Overall, for an instance $\mathbf{x}$ and explanations Φ_i computed over $D_{\mathbf{x}}$,

$$\text{CAFA}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \Phi_i. \tag{5.4}$$

Thus, we use the *global* explanation computed with a strong predictor on $D_{\mathbf{x}}$ as the *local* explanation for $\mathbf{x}$. This *global-for-local interpretation* is superior to LIME's local surrogate approach, as it has been shown that SHAP is more robust than LIME [Hon18, LEC+20, SHJ+20, YFL21].

Algorithm 1 describes the process in detail. Since all points in $D_{\mathbf{x}}$ have the same values for their uncontrollable features, these features have no correlation to class labels of points in $D_{\mathbf{x}}$. Thus, their feature importance will be assigned to 0, as they make no contribution to the prediction. By setting that each class contains K samples (Line 7), we ensure that $D_{\mathbf{x}}$ is balanced.

Algorithm 1 Selective Perturbation and Global-for-Local Interpretation.

Input: Data point $\mathbf{x}$, Prediction model f, Proximity threshold $\pi_{\mathbf{x}}$, Distance Function δ, Controllable features F_c, Sample class size K **O**utput: Feature Importance Φ

1: Let $D'_{\mathbf{x}} = []$;
2: **do**
3: Randomly generate a data point $\mathbf{x}'$ such that for all features $v \in F_u$, $\mathbf{x}'$ contains the same value as $\mathbf{x}$ in v and $\delta(\mathbf{x}, \mathbf{x}') \leq \pi_{\mathbf{x}}$;
4: Append $(\mathbf{x}', f(\mathbf{x}'))$ to $D'_{\mathbf{x}}$;
5: Let r be the size of the smallest class in $D'_{\mathbf{x}}$;
6: **while** $r < K$;
7: Construct $D_{\mathbf{x}}$ from $D'_{\mathbf{x}}$ by sampling K elements from each class in $D'_{\mathbf{x}}$;
8: Let Φ be the global explanation for $g(D_{\mathbf{x}})$ with a strong predictor g;
9: **return** Φ;

5.3 Categorization of Features into Controllable and Uncontrollable Groups

The principle of categorizing features into controllable and uncontrollable groups is a fundamental aspect of the CAFA approach. This categorization is based on the idea that some features can be directly influenced or modified by decision-makers, while others are inherent or fixed characteristics that cannot be easily changed. Controllable features

are those that can be actively manipulated or adjusted through interventions, policies, or individual actions. These features are of particular interest to decision-makers as they represent actionable factors that can be targeted to influence outcomes. Examples of controllable features in the medical domain include treatment options, medication dosages, lifestyle choices, and adherence to medical guidelines.

On the other hand, uncontrollable features are those that are intrinsic or immutable characteristics of individuals or the environment. These features cannot be easily modified and are often determined by factors beyond the control of decision-makers. Examples of uncontrollable features include age, gender, genetic predisposition, and certain environmental factors like climate or geography.

The distinction between controllable and uncontrollable features is crucial for generating actionable insights and making informed decisions. By focusing on controllable features, CAFA enables users to identify the factors that can be targeted for interventions and policy changes. This information can guide the development of targeted strategies and the allocation of resources to areas where they can have the greatest impact. It is important to note that the categorization of features into controllable and uncontrollable groups may vary depending on the specific domain and the context of the problem. What may be considered controllable in one setting may be uncontrollable in another. Therefore, domain expertise and careful consideration of the specific characteristics of the features are necessary when applying the CAFA approach.

5.4 Application of CAFA to Medical Datasets

As an experiment, we apply CAFA to the lung cancer data in Simulacrum and the UCI breast cancer dataset. We predict 12-months survival on the lung cancer dataset, which contains 2,242 instances specified by 24 features:

- Four uncontrollable features: age, ethnicity, sex and height;
- 20 controllable features: morph, weight, dose administration, regimen outcome description, administration route, clinical trial, cycle number, regimen time delay, cancer plan, T best, N best, grade, CReg code, laterality, ACE, CNS, performance, chemo radiation, regimen stopped early, and M Best.[4]

The breast cancer dataset comprises 286 data instances, predicting cancer recurrence, each containing 9 features, which are:

- Two uncontrollable features: age and menopause;
- Seven numerical controllable features: tumor size, inv-nodes, node-caps, deg-malig, breast, breast-quad, and irradiate.

[4]Description of features used in this dataset can be found at the Cancer Registration Data Dictionary and the SACT Data Dictionary, with links available at:
`https://simulacrum.healthdatainsight.org.uk/available-data/table-descriptions/`.

Random forest classifiers are used in both cases.

Firstly, we illustrate that simply dropping uncontrollable features will negatively impact the prediction accuracy. As shown in Table 5.1, the accuracy drops across the three datasets, i.e., lung cancers, breast cancer, and covid19 (we will introduce covid19 dataset in the next section), suggesting features importances achieved from models from fewer features may be different from the ones achieved from using the original dataset.

	Original	Controllable features only
Lung Cancer	0.97	0.85
Breast Cancer	0.79	0.76
COVID19	0.94	0.88

Table 5.1: The prediction for lung cancer, breast cancer, and COVID19 dataset by using the original dataset and the dataset with controllable features only.

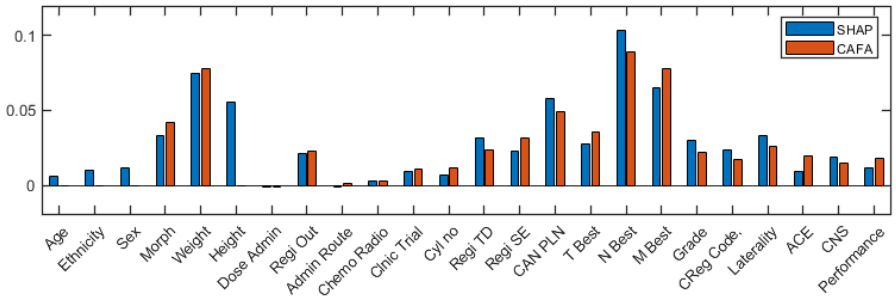
We then explore the influence of controllable features on prediction results on individual instances (local explanations). To this end, we randomly sample an instance from each dataset, as follows:

- *Lung Cancer: age 71; ethnicity 5; sex 0; morph 8140; weight 49.8; height 1.83; dose administration 8; regimen outcome 1; administration route 1; clinical trial 2; cycle number 1; regimen time delay 0; cancer plan 0; T Best 3; N Best 0; grade 3; CReg Code 401; laterality 2; ACE 9; CNS 99; performance 0; chemo radiation 0; regimen stopped early 1; M Best 0.*
- *Breast Cancer: age 40; menopause 0; tumor-size 6; inv-nodes 0; node-caps 1; deg-malig 3; breast 0; breast-quad 3; irradiate 0.*

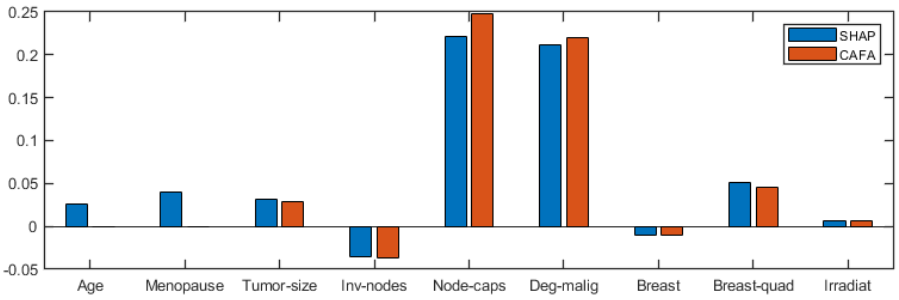
For each instance $\mathbf{x}$, we generate $D_{\mathbf{x}}$ containing 1,000 perturbed instances (binary classification, $K = 500$) and carry out the CAFA calculation as shown in Algorithm 1. We let $\pi_{\mathbf{x}}$ be the average distance between points and feature weights $\omega_i = 1$. Results from SHAP and CAFA are shown in Fig. 5.2. In this figure, the x-axis shows the features; y-axis shows feature importance. For each feature, the left (blue) bar shows the SHAP result of the feature, and the right (red) bar shows the importance calculated with CAFA. We observe that:

1. For uncontrollable features, i.e., "age", "ethnicity", "sex", and "height" from the lung cancer dataset as well as "age" and "menopause" from the breast cancer dataset, the assigned importance value is 0, as expected;
2. For controllable features, there is a strong correlation, 0.96 for lung cancer and 0.99 for breast cancer, between values represented by the blue and the red bars, suggesting that CAFA is agreeable with SHAP.

This suggests that CAFA successfully excludes influences of uncontrollable features with its calculation, while maintaining properties of standard feature attribution algorithms such as SHAP.



(i) A lung cancer instance randomly selected from the Simulacrum dataset. Uncontrollable features are: *Age*, *Ethnicity*, *Sex*, and *Height*.

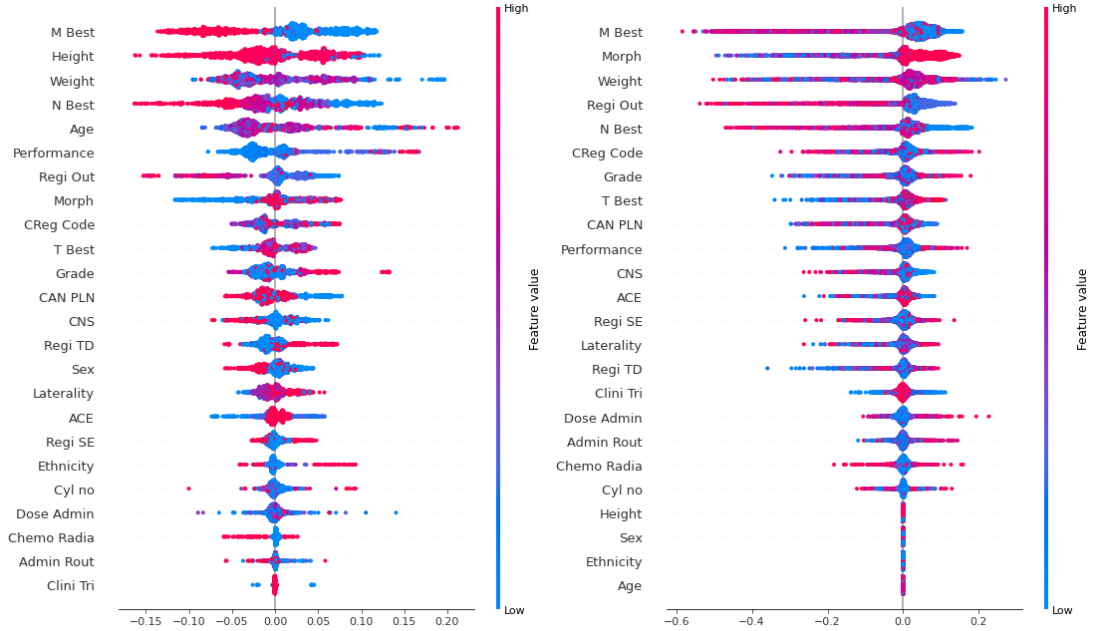


(ii) A breast cancer instance randomly selected from the UCI breast cancer dataset. Uncontrollable features are *Age* and *Menopause*.

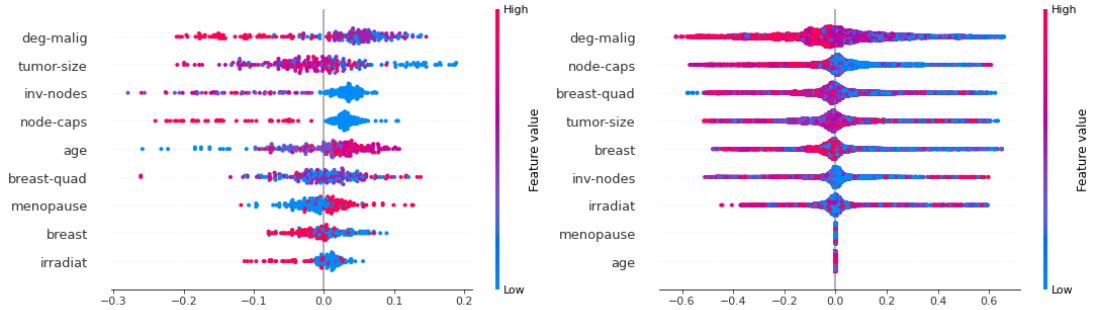
Figure 5.2: Illustration of CAFA vs. SHAP on two explanation instances selected from two medical datasets. We observe that (1) with CAFA, all uncontrollable features are assigned importance 0; (2) for controllable features, CAFA produces results that are agreeable with the ones given by SHAP.

We further study the influence of uncontrollable features with CAFA for global explanations. We randomly sample 100 instances from each dataset and compute global explanations with SHAP and CAFA. We produce "violin plots" using the summary plot function from the SHAP library. Fig. 5.3 (a) and (b) illustrate global explanations for the lung and breast cancer datasets, respectively. There, the x-axis is the feature importance and the y-axis is the features. Color (red to blue) represents the value of a feature.

For Fig. 5.3 (a) and (b), the left-hand side figures show results from SHAP; and at the right-hand side figures show results from CAFA. We can see that: (1) as seen in local explanation cases (Fig. 5.2), all uncontrollable features are assigned an importance value 0; (2) similar patterns to SHAP on controllable features can be seen from CAFA, i.e., similar color patterns for a specific feature; (3) the orders of feature importance differ from SHAP to CAFA. We conclude that, for global explanations, CAFA precludes uncontrollable features from contributing to explanations, and CAFA produces distinct explanations to SHAP even if uncontrollable features are excluded.



(i) Global views of lung cancer cases in the Simulacrum (left: SHAP; right: CAFA). Uncontrollable features are: *Age*, *Ethnicity*, *Sex*, and *Height*.



(ii) Global views of the UCI Breast Cancer dataset (left: SHAP; right: CAFA). Uncontrollable features are: *Age*, and *Menopause*.

Figure 5.3: Global explanations calculated using SHAP and CAFA on the Simulacurm Lung Cancer dataset and the Breast Cancer dataset. Same as Fig. 5.2, we see that uncontrollable features in both datasets have importance 0; and CAFA produces similar results to SHAP for controllable features.

5.5 Application of CAFA to the COVID-19 Dataset

With the outbreak of the COVID-19 pandemic in December 2019, many countries have implemented some non-pharmaceutical control measures to contain the spread of the virus in the absence of effective vaccination and treatment. In this case study, we use CAFA to study the effectiveness of the non-pharmaceutical control measures implemented in the UK.

We formulate the effectiveness of control measures as an XAI modelling problem. We focus on studying the relationship between control measures and the daily reproduction rate R_t. R_t is one of the most important metrics used to measure the epidemic spread. A value greater than 1 suggests the epidemic being expanding; a value less than 1 indicates shrinking. We employ the approach presented in [FMG+20] for estimating R_t from daily infection cases. We then pose the following classification problem:

> *Given non-pharmaceutical control measures applied on a specific day, predict whether R_t is smaller or greater than 1 on that day.*

Having this prediction problem solved by a classifier, we use CAFA to identify control measures that make the greatest contribution to the prediction. Thus, by analysing the behaviour of the prediction model, we gain insight into the effectiveness of control measures.

We have collected a dataset containing daily infection numbers and control measures from 04/January/2020 to 06/September/2021. Each instance consists of uncontrollable features (i.e., daily number of infections, cumulative cases, daily number of deaths and tests performed, temperature and humidity) and controllable features (i.e., implemented control measures). The numbers of daily cases, cumulative cases, deaths, and tests performed are collected from the Public Health England website[5]. Control measure information is retrieved from Wikipedia[6] and various news articles.

We have considered control measures *school closures (SC)*, restrictions on *meeting friends and family indoors (MInd), meeting friends and family outdoors (MOut), domestic travel (DT), international travel (IT), hospitals and nursing home visits (HV), opening of cafes and restaurants (CR), accessing pubs and bars (PB), sports and leisure venues (SL)*, and *non-essential shops (NS)*. The values for control measures are binary, e.g, for "school closure", the values are "open" and "closed"; for "restrictions on meeting indoors" the values are "High" (H) or "Moderate" (M). To accommodate the temporal effect of control measures, each feature is represented categorically. For instance, if they are open, then the "school closure" feature takes value 0; if the schools are closed for 0-5 days, then it takes value 1; etc.

In total, we have collected 4,256 data points across 12 UK regions: East Midlands, East of England, London, North East, North West, South East, South West, West Midlands, Yorkshire and Humber, Northern Ireland, Scotland and Wales. To remove noise and achieve a more accurate R_t estimation, we drop data points with cumulative cases less than 20 for each region and keep 3,936 instances. A sliding-window mean filter of size 3 has been used to filter noise in daily cases.

We split the dataset as 70% for training and 30% for testing, and use a random forest classifier. We achieve a high prediction accuracy of 94.4%. Since we aim to obtain a bird's-eye view of how control measures are affecting the disease, we focus on calculating global explanations. To this end, for each instance $\mathbf{x}$, we generate $D_{\mathbf{x}}$ with

[5]COVID-19 Dashboard (UK): `https://coronavirus.data.gov.uk`

[6]For example, for Wales the control measure data has been collected from `https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic_in_Wales`

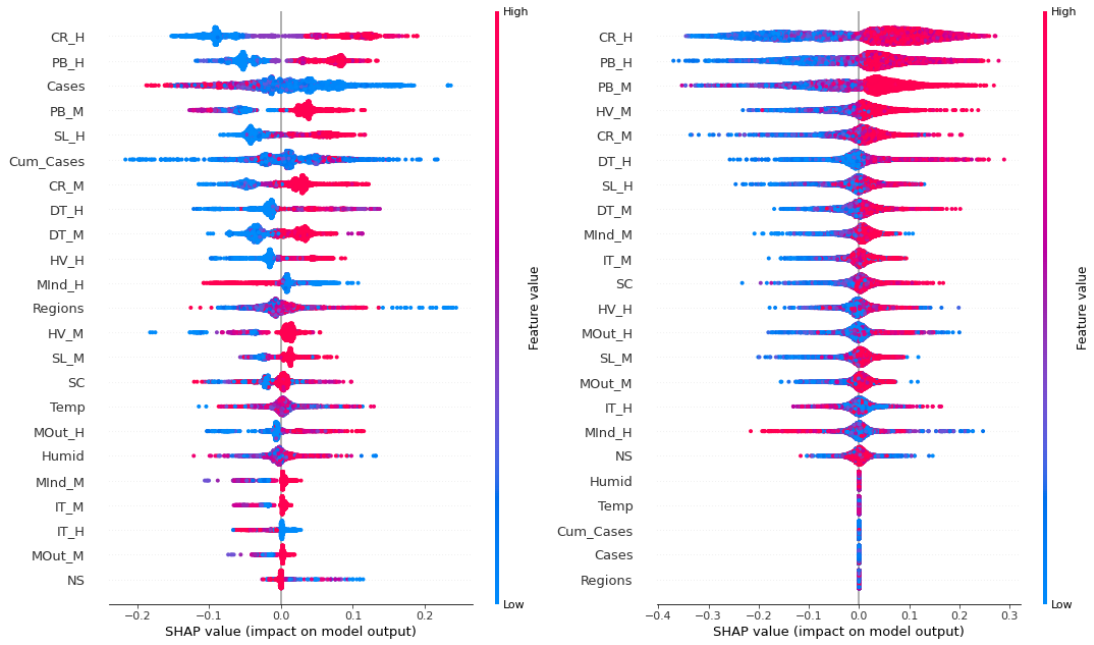


Figure 5.4: Global views of the COVID dataset (SHAP Left; CAFA Right). Uncontrollable features are: *Humidity (Humid)*, *Temperature (Temp)*, *Cumulative Cases (Cum_cases)*, *Daily Infections (Cases)* and *Regions*.

$K = 500$. π_x is the average distance between any two instances; $\omega_i = 1$. By following Algorithm 1, we obtain feature importance using CAFA. The global explanations are shown in Fig. 5.4, right-hand side, with SHAP results shown on the left.

The SHAP results at the left demonstrate that the number of daily cases and cumulative cases both have strong impact in predicting R_t. However, as both are uncontrollable, knowing that they have strong influence to the prediction does not help us understand the effectiveness of control measures. With CAFA (Fig. 5.4 right-hand side), the importance of all uncontrollable features are assigned to 0. Overall, we observe that:

- SHAP considers *High Restriction on Cafes and Restaurants Access (CR_H)*, *High Restriction on Pubs and Bars Access (PB_H)*, *Number of Daily Infections (Cases)*, *Number of Daily Infections (Cases)*, *Medium Restriction on Pubs and Bars Access (PB_M)*, and *High Restriction Sport and Leisure Facilities (SL_H)* as the top five effective control measures; whereas

- CAFA considers *CR_H, PB_H, PB_M*, *Medium Restriction on Hospital and Nursing Home Visits (HV_M)* and *Medium Restriction on Cafes and Restaurants Access (CR_M)* as the top five effective control measures.

CAFA's results are in alignment with WHO's COVID-19 guideline stating the "Three C's" rule that the virus is more transmissible with (1) *Crowded places*; (2) *Close-contact*

settings; and (3) *Confined and enclosed spaces with poor ventilation.*[7] Focusing on restricting access to cafes and restaurants as well as pubs and bars seem to be a very reasonable strategy in reducing the virus transmission, for the reason that these are the most prominent locations meeting the Three C's for most of the population.

5.6 Summary

In this chapter, we introduced the Controllable fActor Feature Attribution (CAFA) approach, a novel XAI technique that addresses the challenge of estimating the importance of controllable features in prediction models. CAFA selectively perturbs controllable factors while leaving uncontrollable ones unchanged, generating a dataset of perturbed instances. By computing global explanations on this dataset, CAFA provides local explanations for each prediction instance, focusing on the impact of controllable features.

We applied CAFA to medical datasets, including lung cancer and breast cancer data, demonstrating its ability to exclude the influence of uncontrollable features and provide meaningful explanations. Additionally, we conducted a real-world case study on the effectiveness of COVID-19 non-pharmaceutical control measures in the UK, showcasing CAFA's potential for identifying the most influential measures in containing the disease.

The chapter also discussed the principle of categorizing features into controllable and uncontrollable groups, highlighting the importance of this distinction for generating actionable insights. The successful application of CAFA to both existing medical datasets and the COVID-19 case study demonstrates its potential as a valuable tool for understanding the impact of controllable factors in various domains. By providing accurate and interpretable explanations, CAFA enables users to make informed decisions and design targeted interventions based on the identified controllable factors.

However, it is important to acknowledge that CAFA, like any other XAI approach, has its limitations and assumptions. The quality of the explanations generated by CAFA depends on the quality and representativeness of the dataset, as well as the choice of the perturbation strategy and the underlying prediction model. Future research could explore alternative perturbation techniques and investigate the robustness of CAFA across different datasets and prediction models. In conclusion, CAFA presents a promising approach for enhancing the interpretability and actionability of machine learning models, enabling users to focus on the controllable aspects that drive predictions and outcomes. As XAI continues to gain importance across various domains, techniques like CAFA will play a crucial role in facilitating informed decision-making and targeted interventions based on the insights derived from complex models.

[7] `https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted`

Chapter 6

Uncertainty-based Controllable Factor Feature Attribution (UCAFA)

Contents

6.1 Introduction

The application of Machine Learning (ML) methods for Health Data is common [Tom21, Raj18]. Unfortunately, many of the predictive ML methods used are not inherently interpretable thus many efforts have gone into the interpretability of ML in application to health data [LEC+20, DFFS23, KED+21]. A common approach for increasing the interpretability of ML is through eXplainable Artificial Intelligence (XAI) approaches [AR23]. The umbrella of XAI approaches are discretized into two focal points; to explain a specific model (*model-specific*), and to provide explanations agnostic to the model choice (*model-agnostic*) [LFM+23]. Due to the innate complexity of developing a model-agnostic approach to XAI, there is often the use of perturbation techniques to provide information locally around an instance, which is then followed by an approximation of the ML model with an inherently interpretable model; this is seen with popular approaches such as Local Interpretable Model-Agnostic Explanations (LIME) [RSG16].

Despite the success of XAI methods that use perturbation techniques such as LIME, the use of perturbation techniques is prone to out-of-distribution samples in the perturbed neighborhood [QYC+21]. Many approaches in literature that extend

on the LIME method are focused on defining the neighborhood. For example, the Deterministic-LIME (D-LIME) [ZK21] approach utilised hierarchical clustering in an attempt to define the neighbourhood from existing instances, therefore this approach is prone to instances being far away from the instance to explain and thus limiting the size of the neighbourhood. In the medical context, in distribution perturbation are crucial because patient data can often exhibit intricate patterns and variations specific to different conditions and medical histories. By maintaining the distribution of the original patient data, the perturbed samples accurately reflect the diverse range of potential scenarios that a patient might encounter.

On the other hand, Controllable Factor Feature Attribution (CAFA) which is introduced in Chapter 5, inherently offers a strategic alternative that aims to mitigate the issue of out-of-distribution perturbations. By focusing on controllable factors, CAFA selectively perturbs only these features, maintaining the fixed values for uncontrollable features. This selective perturbation naturally diminishes the chance of creating non-representative samples, as alterations are confined to a more constrained feature space–one that is designed to embody actionable insights within the data. However, CAFA is not without its limitations. The method relies on an arbitrary distance metric to determine whether an instance is within the distribution. Even with selective perturbation, CAFA may still generate instances that are not entirely representative of the original data distribution, leading to potential inaccuracies in the feature attributions.

To address this, we propose the Uncertainty-based Controllable Factor Feature Attribution (UCAFA) method, to provide more certain and reliable explanations for controllable features. Autoencoders have been used for estimate prediction uncertainties [WHO+24, WHF23], thus the UCAFA leverages a Variational Autoencoder (VAE) to ensure perturbations remain within distribution, effectively addressing the issue of out-of-distribution samples that can skew explanations. This approach not only maintains the focus on controllable factors, as seen in CAFA, but also significantly improves the reliability of attributions by enforcing an uncertainty threshold. The reliability of feature attribution methods can be evaluated through the insertion (insertion score) of important features used in various studies [HMO23, LDC+23]. For example, consider the Simulacrum Lung Cancer dataset used in [DFB+21, DFSF24], figure 6.1 elucidates how the average prediction probability evolves as we incrementally insert features according to their corresponding feature attribution values, starting with the most influential and proceeding in descending order of importance. The baseline probability–indicated by a dashed line across the graphs–serves as a reference point, which represents the complete set of features influence on the model predictions. Thus the intuition is that as more features are added, the prediction probabilities gradually converge towards the baseline probability, this indicates that the impact of less important features diminishes over time. We observe that UCAFA converges faster to the baseline, whereas CAFA, SHAP and LIME require more feature to convergence to the baseline.

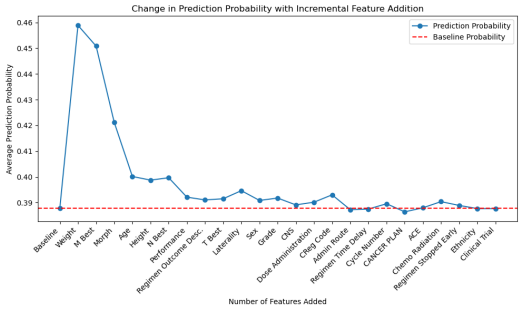
The key contributions of this work are twofold:

- Introducing UCAFA, a novel model-agnostic explanation method that extends CAFA by using a VAE and uncertainty threshold to ensure perturbations remain

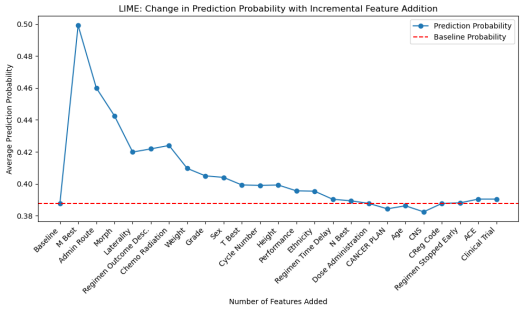
within the expected data distribution for more reliable feature attributions.

- Demonstrating UCAFA's superior performance compared to LIME, SHAP, and CAFA across three healthcare datasets, with faster convergence to baseline probabilities, lower perturbation sensitivity, and reduced error rates, thus enhancing feature attribution reliability and interpretability.

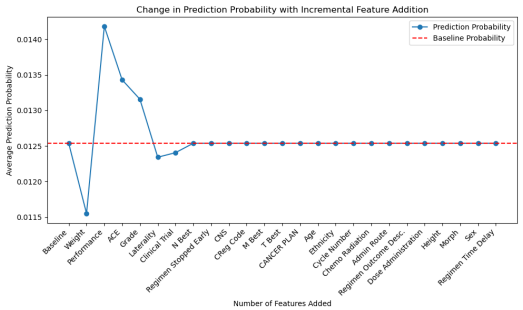
Figure 6.1: Change in prediction probability with incremental feature insertion for the Lung Cancer dataset. The graphs illustrate the average prediction probability as features are incrementally added based on their importance, with the most significant feature included first, followed by the next most important, and so on. The baseline probability is established to represent the scenario where all features are present: (a) SHAP, (b) LIME, (c) CAFA and (d) UCAFA



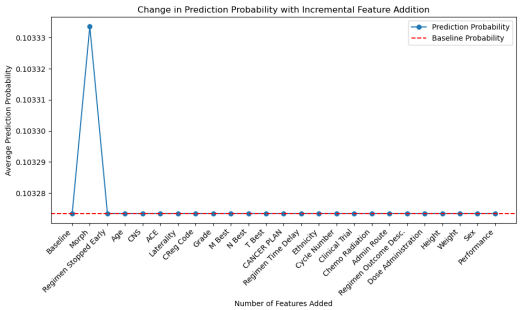
((a)) SHAP



((b)) LIME



((c)) CAFA



((d)) UCAFA

6.2 Variational Autoencoders (VAEs) for Uncertainty Quantification

Variational Autoencoders (VAEs) [KW13] are generative models that learn a low-dimensional latent representation of the data by encoding the input into a latent space and decoding it back to the original space. VAEs optimize the Evidence Lower Bound (ELBO), which consists of a reconstruction term and a regularization term

Figure 6.2: Illustration of the VAE framework that learns parameters θ^* and ψ^* to minimise the distance between q and p.

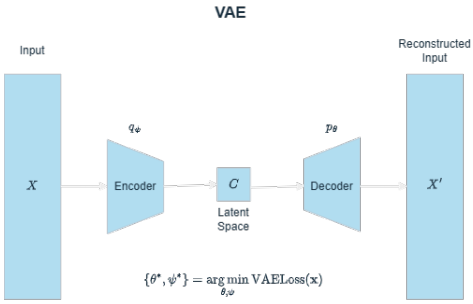
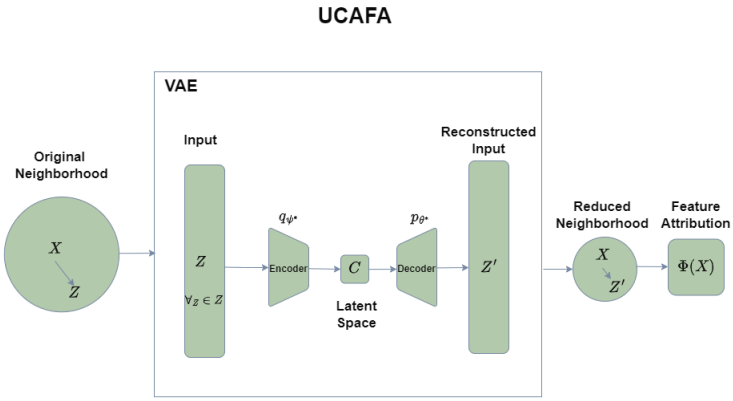


Figure 6.3: UCAFA framework depicting the reducing of the original neighbourhood $\mathcal{Z}$ to the neighbourhood $\mathcal{Z}'$, where the feature attribution values are then calculated for $\mathbf{x}$ in the reduced neighbourhood.



that encourages the latent distribution to be close to a prior distribution, typically a standard Gaussian. The ELBO can be used as a measure of uncertainty, as it quantifies the discrepancy between the true posterior and the variational approximation [WHO+24, WHF23]. Instances with higher ELBO values are considered to be more aligned with the learned data distribution, while instances with lower ELBO values are deemed more uncertain or out-of-distribution.

6.3 Proposed Method: UCAFA

The UCAFA framework shown in figure 6.3 is more formally described such that, given a family of distributions Γ, we let $\gamma(\mathbf{x}) \sim \Gamma$ represent a perturbation function on $\mathbf{x}$, such that we obtain perturbed samples $\mathcal{Z} = \mathbf{x} + \gamma(\mathbf{x})$, where $\mathbf{z} = \langle z^1, \ldots, z^m \rangle \in \mathcal{Z}$ and $\mathcal{Z} \in \mathbb{R}^{n \times m}$. We follow by modifying the surrogate set $\mathcal{Z}$ by restricting the uncertainty in the set. We achieve this by defining a new set $\mathcal{Z}' = \{\mathbf{z} | \, \text{VAELoss}(\mathbf{z}) \leq \delta\}$ where

$$\text{VAELoss}(\mathbf{x}) = -\mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}[log \; p_\psi(\mathbf{x}|\mathbf{c})] + D_{\text{KL}}[q_\theta(\mathbf{c}|\mathbf{x})||p_\psi(\mathbf{c})], \tag{6.1}$$

where $\mathbf{c}$ is the latent space shown in Figure 6.3, and $\delta \in \mathbb{R}$ is the uncertainty threshold and $\mathbf{z}' = \langle z'^{,1}, \ldots, z'^{,m} \rangle$ are the instances in the new neighbourhood $\mathcal{Z}'$ that have an uncertainty value less than or equal to δ.

To provide feature attribution values in this new neighborhood and to make use of valuable game-theoretic properties, we utilize the linear SHAP formulation [LL17]. To achieve this, we define a simple linear regression model F over our neighborhood $\mathcal{Z}'$, producing a vector of predicted values $\mathbf{y}'$ with coefficients β. Here, F is a single black-box model from a set of potential black-box models $\mathcal{F}$, allowing us to define a model-agnostic approach. Therefore, we have:

$$\mathbf{y}' = F(\mathcal{Z}'), \text{ where } F \in \mathcal{F}. \tag{6.2}$$

Following this, we obtain the coefficients:

$$\beta = (\mathcal{Z}'^{T} \mathcal{Z}')^{-1} \mathcal{Z}'^{T} \mathbf{y}'. \tag{6.3}$$

This formulation allows the linear SHAP values to be calculated for each single feature x^j in the perturbed neighborhood of reduced uncertainty $\mathcal{Z}'$ as:

$$\Phi^j(\mathbf{x}) = \beta^j (x^j - \mathbb{E}[\mathcal{Z}'^{,j}]). \tag{6.4}$$

Algorithm 2 UCAFA: Uncertainty-based Controllable Feature Attribution.

Input: Original dataset X, Prediction model f, VAE models (Enc, Dec) with parameters (θ^*, ψ^*), Proximity threshold π, Perturbation function $\gamma \sim \Gamma$, Uncertainty threshold δ, Linear SHAP model F

Output: Feature importance values Φ for each feature in X

Train VAE on X to learn data representation, maximising ELBO to obtain θ^* and ψ^*.
Generate perturbed samples $\mathcal{Z}$ using $\gamma(\mathbf{x})$ for each $\mathbf{x} \in X$.
Initialize filtered neighbourhood $\mathcal{Z}' = \emptyset$.
for each $\mathbf{z} \in \mathcal{Z}$ **do**
    Calculate $\mathcal{L}(\mathbf{z})$ using trained VAE.
    if $\mathcal{L}(\mathbf{z}) \leq \delta$ **then**
        Add $\mathbf{z}$ to $\mathcal{Z}'$.
    end if
end for
Apply linear regression F on $\mathcal{Z}'$ to predict $\mathbf{y}'$.
Calculate coefficients β using $\mathbf{y}'$ and $\mathcal{Z}'$.
for each feature x^j in X **do**
    Calculate $\Phi^j(\mathbf{x}) = \beta^j (x^j - \mathbb{E}[\mathcal{Z}'^{,j}])$.
end for
Return Φ.

This approach ensures that the feature attributions are derived from a neighborhood of reduced uncertainty, providing more reliable explanations for controllable features compared to the original CAFA method. We provide details on the UCAFA implementation in algorithm 2.

6.4 Experiments

6.4.1 Datasets

6.4.1.1 The Simulacrum

The Simulacrum Lung Cancer dataset, developed by Health Data Insight CiC, is derived from anonymous cancer data provided by the National Cancer Registration and Analysis Service[1], a part of Public Health England. We consider a lung cancer subset as presented in [DFB+21], containing 2,242 instances and 24 input features. For demonstration we manually divide features into two categories: controllable and uncontrollable. The set of uncontrollable features that are omitted in our approach are age, ethnicity, sex, and height, the remaining 20 input features are used as controllable features.

6.4.1.2 UCI Breast Cancer

The UCI Breast Cancer dataset [ZS88] contains 286 instances, each characterized by 9 features, to predict cancer recurrence as a binary classification problem. The target variable indicates either "recurrence-events" or "no-recurrence-events". The features include two uncontrollable attributes (age and menopause) and seven controllable diagnostic measurements (tumor size, inv-nodes, node-caps, deg-malig, breast, breast-quad, and irradiate).

6.4.1.3 UK COVID-19 Dataset

For our analysis, we used the dataset provided by [KLS+22], which contains a total of 3,936 instances across 12 UK regions. Each instance consists of both uncontrollable and controllable features. The uncontrollable features include the cases, cumulative cases, deaths, tests, temperature, and humidity. The controllable features encompass various non-pharmaceutical interventions like school closures, indoor and outdoor gathering restrictions, travel limitations, and the operation of public venues. This dataset enables a detailed examination of how controllable measures influenced COVID-19's spread in the UK

[1]The features of the dataset are documented in the Cancer Registration Data Dictionary and the SACT Data Dictionary. Relevant links are accessible via the specified website: https://simulacrum.healthdatainsight.org.uk/available-data/table-descriptions/.

6.4.2 Experimental Setup

For all three datasets, we employed the Random Forest algorithm, which yielded the best accuracy compared to other machine learning models. Hyperparameters are selected through grid search to re optimal performance. The Random Forest model achieved an accuracy of 0.97 on the Lung Cancer dataset, 0.79 on the Breast Cancer dataset, and 0.94 on the UK COVID-19 dataset. These high accuracy scores demonstrate the effectiveness of the model for predicting patient survival, cancer recurrence events, and the impact of control measures on the daily rate of infection (Rt) of COVID-19 (see [KLS+22, KED+21] for dataset details). An overview is provided in table 6.1.

Table 6.1: Summary of datasets and model performance

Dataset	Instances	Features	Task	Accuracy
Lung Cancer	2,242	24 (20 controllable, 4 uncontrollable)	Classification	0.97
Breast Cancer	286	9 (7 controllable, 2 uncontrollable)	Classification	0.79
UK COVID-19	3,936	17 (11 controllable, 6 uncontrollable)	Classification	0.94

6.4.3 Metrics

6.4.3.1 Insertion

The insertion score is a commonly used metric to evaluate the performance of a feature attribution algorithm [LDC+23]. It starts with an all-zero vector and incrementally adds features in order of their attributed importance. The change in the model's prediction probability should reflect the change in feature importance. The baseline prediction probability, which represents the model's prediction when all features are present, serves as a reference point. As important features are added, the prediction probability should quickly converge towards the baseline. A smaller area between the baseline and the changing prediction probability indicates that the attribution method effectively identifies and prioritizes the most influential features.

The faster the convergence to the baseline probability, the more accurate and reliable the feature attributions are considered to be. This metric assesses the attribution method's ability to capture the true importance of features in driving the model's predictions for a given instance.

6.4.3.2 Proximity

The proximity measure in this work, simply describes the average Euclidean distance from an instance to explain to each of the perturbed samples in the neighbourhood. A smaller value of proximity indicates a shorter distance from the instance to explain to all neighbouring samples and thus provides a defined neighbourhood size.

6.4.4 Results

Our evaluation critically examines the performance of the proposed UCAFA methodology, as presented in Table 6.2, which provides a detailed comparison of SHAP, LIME, CAFA and UCAFA models based on the average difference between the baseline prediction probability and the effects of sequentially inserting features, ranked by their importance as determined by each model. This methodology plays a crucial role in assessing how the stepwise addition of features influences a model's prediction probability, show casing a refined understanding of model sensitivity and its capability for interpretability.

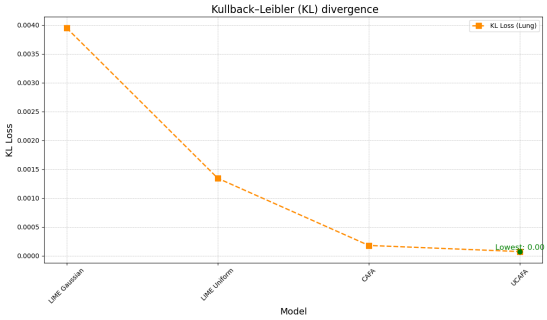
Table 6.2: The average difference between the baseline prediction probability and each sequential insertion of features, ranked by their importance as determined by each model. Lower values indicate a quicker conversion to the baseline probability, indicating better identification of important features and thus a more reliable neighborhood.[a] implies a value < 0.000.

Model	Lung Cancer	Breast Cancer	COVID19
UCAFA	**0.000**[a]	**0.0012**	**0.0021**
CAFA	0.002	0.0031	0.0039
SHAP	0.010	0.0238	0.0071
LIME	0.019	0.0295	0.0075

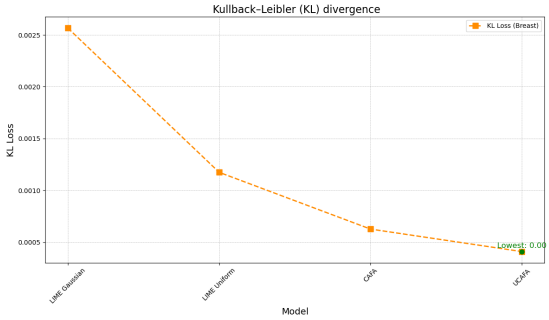
Analysis suggests UCAFA demonstrates better efficiency across all tested datasets: Lung Cancer (0.000[a]), Breast Cancer (0.0012), and COVID-19 (0.0021). This signifies the UCAFA method converges quicker to the baseline prediction probability, thus presenting better assigned feature attribution values. On the other hand CAFA shows a relatively moderate sensitivity to feature insertion, with differences of 0.002, 0.0031, and 0.0039 across the datasets. SHAP and LIME exhibit the largest average differences, especially in Breast Cancer (SHAP: 0.0238, LIME: 0.0295), revealing a slower convergence to the baseline probability with the addition of features. Conclusively, this indicates that feature attribution values are assigned more effectively with the UCAFA method.

We further evaluate UCAFA's ability to generate perturbations that align with the original data distribution by computing the KL divergence between the original dataset and the perturbed instances generated by each method. Figure 6.4 presents the KL divergence results for LIME (with Gaussian and uniform perturbations), CAFA, and UCAFA across the three datasets.

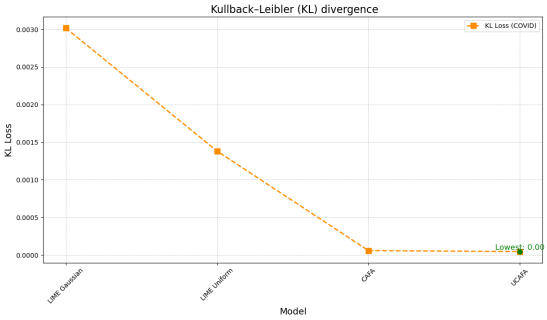
UCAFA consistently achieves the lowest KL divergence across all datasets, indicating that its perturbed instances most closely resemble the original data distribution. CAFA performs well, with lower KL divergence compared to LIME variations. The LIME methods, particularly with Gaussian perturbations, exhibit higher KL divergence, suggesting their perturbed instances deviate more from the original data distribution. These results reinforce UCAFA's superiority in generating perturbations that align



(i) Lung Cancer Dataset



(ii) Breast Cancer Dataset



(iii) UK COVID-19 Dataset

Figure 6.4: KL divergence between the original data distribution and the perturbed instances generated by each method. Lower values indicate better alignment with the original data distribution.

with the data distribution, providing a more reliable basis for deriving accurate and trustworthy feature attributions.

Next, we compare the proposed UCAFA methodology against established models such as the CAFA and LIME with both Uniform and Gaussian perturbations, as they rely on random perturbations. This assessment was conducted across three datasets: the Simulacrum Lung Cancer dataset, the UCI Breast Cancer dataset, and a UK COVID-19 dataset. The essence of our comparison revolves around measuring the average distance over each perturbed instance and the original data instance, which we refer to as the proximity. This metric helps quantify the size of the neighbourhood and alongside insertion score can draw conclusions on the neighourhood size and respective feature attribution performance.

The data presented in Table 6.3 showcase the superior capability of the proposed UCAFA model is designed to preserve the perturbed instances that are in-distribution with the original instance across all examined datasets, with deviations recorded at 0.074 ± 0.008 for the Lung Cancer dataset, 0.071 ± 0.021 for the Breast Cancer dataset, and 0.037 ± 0.004 for the COVID19 dataset. In contrast, the CAFA model, though outperforming the LIME variations, does not match the preservation of a small neighbourhood as given by UCAFA, showing deviations of 0.097 ± 0.005, 0.109 ± 0.010,

Table 6.3: Error comparison between models with respect to the L2 distance between each perturbed instance and the original data instance which needs an explanation, referred to as the error rate. Lower values indicate a smaller neighbourhood. A lower value in conjunction with a low insertion score value indicates that a smaller sized neighbourhood is better for identifying important features.

Model	Lung Cancer	Breast Cancer	COVID
UCAFA	$\mathbf{0.074 \pm 0.008}$	$\mathbf{0.071 \pm 0.021}$	$\mathbf{0.037 \pm 0.004}$
CAFA	0.097 ± 0.005	0.109 ± 0.010	0.044 ± 0.012
LIME (Uniform)	0.28 ± 0.002	0.171 ± 0.002	0.27 ± 0.002
LIME (Gaussian)	0.48 ± 0.006	0.291 ± 0.007	0.47 ± 0.007

and 0.044 ± 0.012 for the respective datasets. Notably, both versions of the LIME model demonstrate considerably larger deviations, highlighting the limitations of their perturbation techniques when not adequately aligned with the native data distribution.

UCAFA presents good performance across the well established insertion metric whilst rely on minimal size neighbourhoods and thus underscores its robustness and dependability in feature attribution, significantly attributed to its novel utilization of VAEs to ensure that perturbations do not stray from the distribution. This approach effectively reduces the risk of generating out-of-distribution samples that could undermine the model's explanatory accuracy and result in erroneous interpretations with respect the instance that needs explanation.

6.5 Conclusion

In this thesis, we provide a model-agnostic explanation method UCAFA, designed to provide explanations in neighbourhoods with reduced uncertainty, thus allowing for more reliable explanations. UCAFA extends the CAFA approach by incorporating a VAE to learn the data representation and applying an uncertainty threshold based on the ELBO. This ensures that the perturbed instances used for deriving feature attributions remain within the expected data distribution. Our experiments on three healthcare datasets (lung cancer, breast cancer, and COVID-19) demonstrated the superior performance of UCAFA compared to existing methods like LIME, SHAP, and CAFA.

UCAFA exhibited faster convergence to the baseline prediction probability and lower sensitivity to feature perturbations, as evidenced by the smaller average differences between the baseline and sequentially inserted features. Moreover, UCAFA consistently achieved the lowest KL divergence values and error rates between the original data distribution and the perturbed instances across all three datasets, confirming its ability to generate perturbations that closely align with the data distribution while maintaining low deviation from the original instances. These findings underscore the importance of accounting for uncertainty when generating perturbations for model explanations. By fo-

cusing on in-distribution perturbations, UCAFA provides more reliable and interpretable feature attributions. The enhanced interpretability and reliability of machine learning models in healthcare, as demonstrated by UCAFA, have significant medical implications. By providing more accurate and trustworthy explanations, UCAFA empowers healthcare professionals to make better-informed decisions regarding diagnosis, treatment, and resource allocation, potentially improving patient outcomes and healthcare efficiency. As such, UCAFA contributes to the growing field of explainable AI in healthcare, paving the way for more transparent and reliable clinical decision support systems.

Part II

XAI in the Finance Domain

Chapter 7

Global Open-Ended Funds: Introduction and Datasets

Contents

7.1 Introduction

Fund performance evaluation naturally differs from conventional analysis on cross-sectional stock returns, even though the same set of risk factors, as input variables,[1] may be under consideration. While evaluation models of stock returns are found to be mostly linear,[2] the impact of similar risk factors on equity fund performance is likely to be more complex and thus nonlinear. The literature has developed a number of solutions to overcome the challenges of fund evaluation. Some have focused on model inputs by designing performance measures tailored to equity funds [KW01, GBRV09, MSWY22], while others have attempted to improve their estimation strategy by incorporating insights from statistical modeling [FC21].

This research joins the latter stream of literature by using machine learning models to analyze the existing factors available for fund evaluation. Evidently, the machine learning techniques involved benefit from the nonlinear framework, which allows risk factors to determine performance in a different manner than with a typical linear regression model. In particular, we focus on explainable artificial intelligence (XAI) models in this researach for two reasons: first, the explainable feature provides the

[1]We use the terms risk factor, input variable, and explainer interchangeably, depending on which context we aim to highlight: finance, machine learning, or XAI.

[2]The factor models in the asset pricing literature usually provide fairly good explainability in the stock returns, e.g., [FF93, Car97].

researcher with richer information on the relevance of each risk factor; second, XAI models also inform the direction of influence on the output variable, in a comparable way as with existing statistical models.

This work first implements a machine learning technique to improve the model's goodness of fit on fund performance and then presents an XAI application to address the following two points. First, we validate that the machine learning model findings are consistent with finance domain knowledge. Second, the results of the XAI model enable us to provide novel implications arising from an open question in the literature on fund performance and international diversification.

It has been demonstrated that machine learning models can generally outperform the conventional linear setting in asset pricing studies.[3] This work consists of two principal elements regarding machine learning. First, we employ eXtreme Gradient Boosting (XGBoost), a state-of-the-art model that usually outperforms other tree-based models such as random forest [HTFF09]. Second, in order to overcome the black-box nature of the machine learning models, we also incorporate an XAI technique,[4] namely SHapley Additive exPlanations (SHAP) as proposed by [LL17], to gain information about the significance and direction of risk factors' influence.

7.1.1 Aims and Contributions

This work considers two different types of risk factors as explanatory features. The first group captures the macro-finance dynamics at the aggregate level, including stock returns, foreign exchange rate returns, and interest rates across countries. The second group contains more granular variables that describe each fund's past performance via an indicator as well as cross-country allocations of investment holdings by their Herfindahl Hirschman Index (HHI) values, which are introduced in subsection 7.2.1. The HHI is a concentration measure commonly applied in finance studies [CHR22]. Our findings can be summarized in three dimensions. First, the XGBoost model significantly enhanced explanatory power with respect to a conventional linear regression benchmark. Second, the directional explanations provided by the SHAP method are consistent with the coefficient signs of the benchmark model, with statistical significance. Third, international diversification is generally advantageous but not always beneficial for fund performance. To the best of our knowledge, this is the first study that employs XAI techniques to examine the relationship between fund performance and international diversification in portfolio holdings.

[3] The application of machine learning models in finance has focused heavily on market returns [GKX20].

[4] There is another widely-used technique in the XAI domain named Local Interpretable Model-agnostic Explanations (LIME), developed by [RSG16]. Both LIME and SHAP belong to the class of additive feature attribution methods. The main difference between them lies in how they provide explanations. While LIME obtains the explanations by solving a penalized linear regression, SHAP considers all possible combinations of explanations in its estimation. Here, we are using SHAP because, by construction, LIME is more likely to suffer from the same problems faced in linear regression analysis [Mol23].

7.2 Background

There have been a number of studies on XAI application in finance, including credit risk management [BGMP21], cryptocurrency investments [BGR22], debt financing [LB22] and prediction of stock splits [LLS23]. In a similar vein to our work, but applied to risk assessments for the stock market, [Ber23] leveraged boosted trees combined with a Shapley values-based XAI model. See also [AB18, Mol23]; and [JLJK23] for an overview of XAI applications in other domains, and [MTvMH+21, DFB+21] for a comparison of XAI models.

7.2.1 Herfindahl Hirschman Index (HHI)

The Herfindahl Hirschman Index (HHI) is a widely used measure of concentration, which in this study, is employed to quantify the degree of international diversification in a fund's portfolio holdings. The HHI is defined as:

$$HHI = \sum_i s_i^2 \tag{7.1}$$

where s_i represents the portfolio holding in country i, and $\sum_i s_i = 1$. The HHI ranges between values close to 0 and 1, with higher values indicating a higher concentration of investment holdings in a few or even a single country. For example, an HHI equal to 1 assumes 100% holdings of US stocks, with $s_{US} = 100\%$ and $s_{non-US} = 0$. Conversely, lower HHI values represent situations where portfolio holdings are well diversified across multiple countries. For instance, an HHI of 0.25 can be obtained from a portfolio with equally weighted holdings in four different countries.

The HHI provides a concise and intuitive measure of international diversification, allowing for a clear understanding of how concentrated or spread out a fund's investments are across different countries. By incorporating the HHI into our analysis, we aim to capture the relationship between the degree of international diversification and fund performance, offering insights into the potential benefits and drawbacks of holding geographically concentrated or diversified portfolios.

7.2.2 Probit Regression

Probit regression [Bli34] is a statistical model used for analyzing binary outcome variables. It assumes that the probability of the binary outcome is related to a linear combination of predictor variables through the cumulative distribution function (CDF) of the standard normal distribution. Probit regression was chosen for this study due to its widespread use in economics and finance literature. It differs from logistic regression in the link function used, with probit regression employing the inverse of the standard normal CDF. While both models often produce similar results, probit regression is preferred when the underlying latent variable is assumed to follow a normal distribution, and its coefficients can be interpreted in terms of changes in the probit index.

The probit model can be expressed as:

$$P(y = 1|\mathbf{x}) = \Phi(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) \tag{7.2}$$

where:

- y is the binary outcome variable
- $\mathbf{x} = (x_1, \ldots, x_p)$ is a vector of predictor variables
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ is a vector of regression coefficients
- $\Phi(\cdot)$ is the CDF of the standard normal distribution

The coefficients $\boldsymbol{\beta}$ are typically estimated using maximum likelihood estimation (MLE). The interpretation of coefficients is in terms of the change in the z-score (probit index) for a one-unit change in the corresponding predictor, holding other variables constant. Model fit is often assessed using pseudo R-squared measures, such as McFadden's R-squared, which compares the log-likelihood of the fitted model to that of a null model. However, these measures have lower values than the R-squared in linear regression and should be interpreted with caution. Probit regression is similar to logistic regression but uses a different link function. The choice between probit and logit models is often based on convenience or tradition, as they generally produce similar results.

In summary, probit regression is a generalized linear model for binary outcomes that uses the normal CDF as the link function. It is widely used in various fields for modeling the relationship between predictors and binary outcomes.

7.3 Datasets

We collected data on Global Open-Ended Funds from the Morningstar Direct database. Global Open-Ended Funds are mutual funds that invest in a diversified portfolio of securities from around the world, providing investors with exposure to international markets. These funds offer investors a convenient way to gain diversified exposure to international markets through a single investment vehicle. The 'open-ended' structure of these funds allows for greater liquidity and flexibility compared to closed-end funds, which have a fixed number of shares that are traded on an exchange. Open-ended funds can issue and redeem shares at any time, enabling investors to buy or sell shares on demand at the fund's current net asset value (NAV). The NAV is calculated by dividing the fund's total assets minus liabilities by the number of outstanding shares and fluctuates based on market conditions and investor demand. This open-ended structure allows the fund to adapt to changing market conditions and investor preferences by issuing or redeeming shares as needed. When investors buy shares, the fund issues new shares and increases in size. Conversely, when investors sell shares, the fund redeems the shares and decreases in size. This flexibility provides investors with the ability to easily enter or exit the fund based on their investment goals and market outlook.

Although the data was available from as far back as 2000, the missing values in the portfolio holdings were substantial. In order to maintain a sufficiently large number of observations, we focused on the most recent years, from 2016 to 2021, in which we reserved the first year (four quarters) to construct an indicator of the fund's past performance. The holdings data was available at a monthly frequency but was transformed to quarterly frequency by keeping only the end-of-quarter observations,[5] as most funds conform to regulations by disclosing their holdings quarterly. We focused our study solely on funds domiciled in the G10 countries.[6] In short, our sample covered the period from 2017:Q1 to 2021:Q3,[7] encompassing 4,330 funds in total.

7.3.1 Macro-finance and Fund-level Variables

As shown in Table 7.1, a larger quantity of funds was available with complete data, containing no null values, from January 2016 to September 2021. This time frame was carefully selected to ensure the dataset's comprehensiveness, capturing long-term financial trends and cycles while upholding a high standard of accuracy and quality. Our final dataset comprised 4,330 funds originating from the G10 countries as shown in Table 7.1.

Table 7.1: **Dataset.**
Equity funds from G10 countries with asset allocations within the G10 countries were analyzed for the period from January 2016 to September 2021. This particular time-period was chosen due to the greater availability of complete data from a larger number of funds, free of null values. Additionally, this extended timeframe ensures that the study captures more extensive financial trends and cycles, while maintaining high standards of data quality and accuracy.

Time line	BEL	CAN	CHE	FRA	GBR	DEU	ITA	JPN	NLD	SWE	USA	$\sum$G10
2001-2021	0	20	0	0	0	0	0	0	0	0	54	74
2006-2021	0	126	3	0	0	22	0	0	0	1	20	505
2011-2021	8	492	50	60	198	106	8	0	19	123	1718	2782
2016-2021	**1**5	**7**96	**1**33	**2**51	**4**63	**1**38	**1**2	**1**7	**5**3	**1**82	**2**270	**4**330

In this study, we address the problem as a binary classification task where the Net Asset Value (NAV) is the target variable. The NAV takes on a value of 0 if the previous quarter's NAV decreased and a value of 1 if the NAV increased compared to the last quarter.

[5]Since regulations vary between countries, quarterly holdings are not always recorded at the end of the quarter. For missing data, we employed a forward-filling strategy by using the most recent observation recorded in the previous two months. If there is no observation in those two months, we leave the data as missing in this particular quarter.

[6]Despite its name, the G10 actually includes 11 countries, namely Belgium (BEL), Canada (CAN), France (FRA), Germany (DEU), Italy (ITA), Japan (JPN), the Netherlands (NLD), Sweden (SWE), Switzerland (CHE), the United Kingdom (GBR), and the United States (USA).

[7]While the whole sample spans 19 quarters, our analysis includes only 18 due to one-period lag applied to the input variables.

For instance, let's assume that in a hypothetical scenario, the NAV recorded in June 2013 was $40,000 and in September 2013, it was $45,000. In this case, the NAV is considered to have increased, and a value of 1 is assigned for September 2013. Conversely, if the NAV recorded in December 2013 was $42,000 and when compared to the NAV recorded in September 2013 $45,000, it is found that the NAV decreased, a value of 0 is assigned for December 2013.

For the chosen time period and its corresponding funds from Table 7.1, we incorporated macroeconomic features such as stock market (ST), interest rates (IR), and exchange rates (ER). The data for ST, IR, and ER were collected from Datastream,[8] an online financial database that provides comprehensive global coverage of financial and macroeconomic data. The data for these features are illustrated in Figures 7.1 and 7.2.

Moreover, In order to better understand the performance of the fund, we have computed a new feature called past performance (PPrfm). This feature is based on the binary values of the fund's NAV over the last four quarters, and represents the number of times the fund has performed better over the four quarters. Specifically, the PPrfm feature takes values between 0 and 4, with 0 indicating that the fund did not perform better in any of the four quarters, and 1 indicating that the fund performed better in one of the four quarters. A value of 2 indicates that the fund performed better in two of the four quarters, while a value of 3 indicates that the fund performed better in three of the four quarters. Finally, a value of 4 indicates that the fund performed better in all four quarters. Additionally, another input feature, the HHI value, is measured using asset allocation weights in the G10 countries and falls within the range of 0 to 1. A low HHI value indicates that the assets are spread out across many different companies, while a high HHI value indicates that the assets are concentrated in a few companies. In the case of asset allocation weights in the G10 countries, a high HHI value would indicate that the assets are concentrated in a few countries, while a low HHI value would indicate that the assets are spread out across many different countries.

After preparing the data, we have the following features for analysis: ST, IR, ER, PPrfm, HHI, and the output prediction is the NAV, which is a binary target: 0 when the previous quarter's net assets value decreased and 1 when compared with the last quarter's net assets value increased.

[8]Datastream is a product of Refinitiv, which offers historical financial and economic information, including stock market data, interest rates, exchange rates, and more. For more information, please visit https://www.refinitiv.com/en/products/datastream-macroeconomic-analysis.

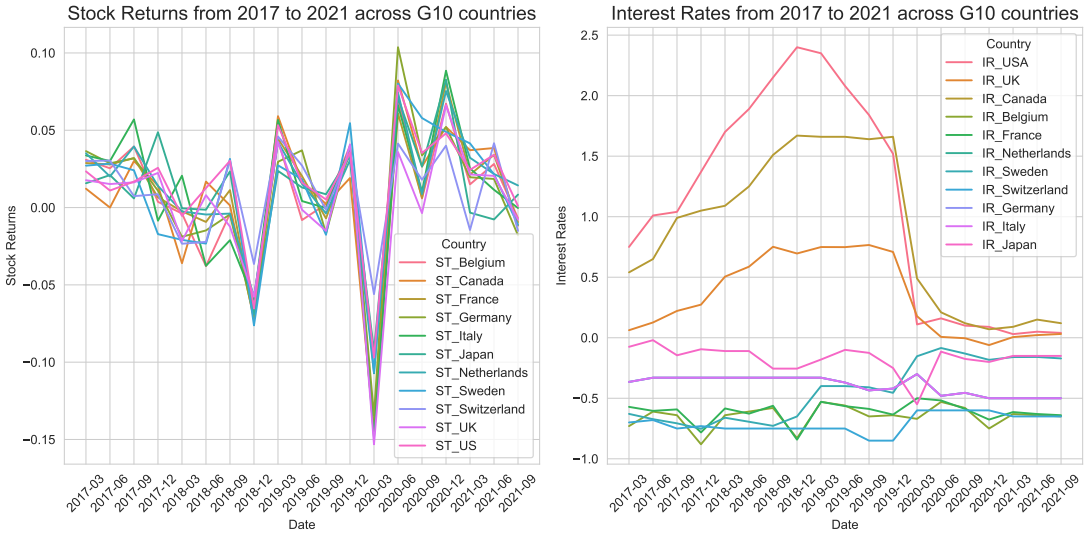


Figure 7.1: **Stock Market and Interest Rates**. The figure explores the temporal trends of interest rates and stock market returns in G10 countries from 2017-January to 2021-September. The left-hand side lineplot displays the stock market returns, while the right-hand side lineplot shows the interest rates across the G10 countries.

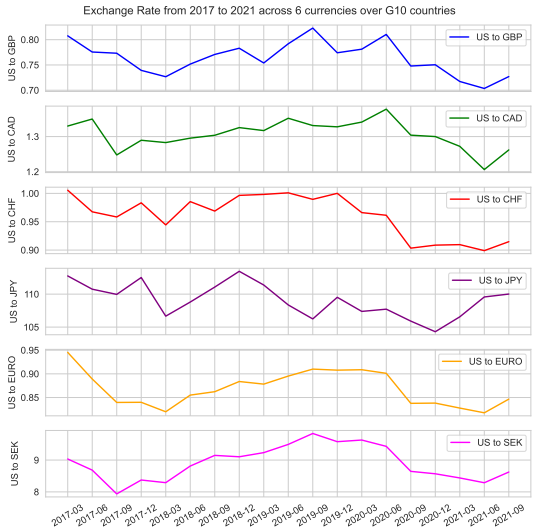


Figure 7.2: **Exchange Rates.** The figure explores the temporal trends of exchange rates across 'GBP', 'CAD', 'CHF', 'JPY', 'EURO' and, 'SEK' over G10 countries from 2017-January to 2021-September.

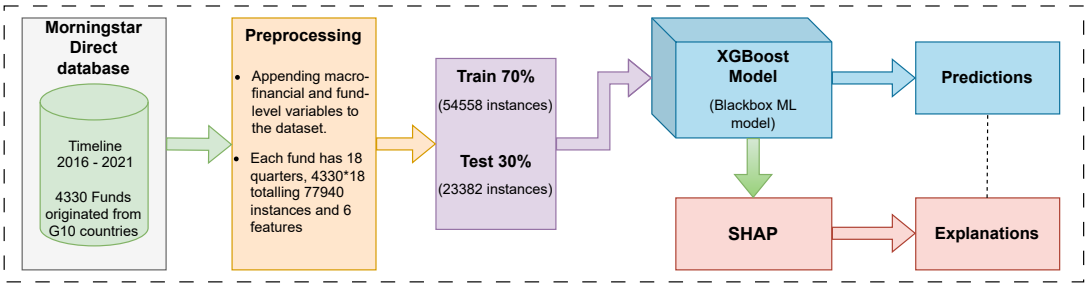


Figure 7.3: **A**rchitecture of the proposed model's workflow. The process starts with the Morningstar Direct dataset, to which macro-financial and fund-level variables are added during preprocessing. The enriched dataset is then split into a 70% training and 30% testing set. These sets are subsequently used as input to the XGBoost model. The model's output is interpreted using the SHAP model, providing comprehensible explanations for the predictions.

Table 7.2: **Descriptive Statistics of Fund Data for G10 Countries.** The table presents the descriptive statistics of the data used in this research. The input features include cross-country macro-finance indicators such as stock market return (ST), interest rate (IR), and exchange rate (ER), extracted using principal component analysis, and fund-level measures such as past performance (PPrfm), and the Herfindahl Hirschman Index (HHI). The target feature Net Asset Value (NAV) is a binary variable, where a value of zero indicates a decrease and one indicates an increase compared with the previous quarter's value. The data at quarterly frequency cover the period from January 2017 to September 2021. In total, there are 18 quarters for each of the 4330 funds. We have also reported the metrics of the data distribution and temporal dependencies, including skewness, kurtosis, and the first-order autocorrelation AR(1).

	ST	IR	ER	PPrfm	HHI	NAV
AR(1)	−0.321	−0.235	0.879	0.695	0.935	−0.036
Skewness	1.582	−0.106	0.120	−0.410	−0.139	−0.503
Kurtosis	2.796	−1.575	−1.212	−0.465	−1.620	−1.746
mean	1.202	0.177	−0.739	2.485	0.526	0.622
Std. dev.	4.353	0.234	3.054	1.069	0.357	0.484
min.	−15.310	−0.220	−8.996	0.000	0.000	0.000
25%	0.520	0.012	−1.719	2.000	0.156	0.000
50%	2.050	0.146	−0.242	3.000	0.587	1.000
75%	3.530	0.410	1.011	3.000	0.880	1.000
max.	10.370	0.600	6.111	4.000	1.000	1.000

Finally, we refer to Figure 7.3 for a detailed view of the proposed model's architectural workflow, from data preprocessing to prediction interpretation using SHAP.

7.4 Summary

This chapter introduced the motivations and aims of this study, focusing on the application of machine learning and explainable AI (XAI) techniques to improve fund performance evaluation. The background section provided an overview of relevant literature and introduced the concept of probit regression. The aims and contributions of this study were discussed, highlighting the novelty of employing XAI techniques to examine the relationship between fund performance and international diversification.

The datasets section detailed the data collection process, focusing on Global Open-Ended Funds from the Morningstar Direct database for the period from 2016 to 2021. The target variable, Net Asset Value (NAV), was introduced as a binary classification task. The macro-finance and fund-level variables used in the study, such as stock market returns, interest rates, exchange rates, past performance, and the Herfindahl Hirschman Index (HHI), were also described. The chapter concluded with descriptive statistics and visualizations of the dataset.

In the next chapter, we will delve into the analysis of global open-ended funds, comparing the performance of probit regression and XGBoost models, examining the XAI results based on input features, and investigating the influence of international diversification on fund performance.

Chapter 8

Analysis of Global Open-Ended Funds

Contents

In this chapter, we delve into the application of XAI techniques to gain insights into the performance of Global Open-Ended Funds. By leveraging state-of-the-art machine learning models, such as XGBoost, and XAI methods, we uncover the key drivers of fund performance and explore the implications of portfolio diversification across G10 countries. Our analysis sheds light on the complex relationships between various explanatory features and fund performance, providing interpretable results that are consistent with traditional statistical models. This chapter serves as an introduction to the potential of XAI in finance and paves the way for future research in this exciting field.

8.1 Probit Regression versus the XGBoost Model

The probit regression model assesses the binary classification problem as found in the XGBoost model by using a linear structure. We report the results of both univariate and multivariate analyses in Table 8.1. Models (1) to (4) are univariate regressions, each using a single input feature, while models (5) and (6) are multivariate regressions incorporating multiple input features. The pseudo R-squared is small or even negative, implying poor model fitness. Alternatively, the input features show strong correlations to the fund performance. The coefficient signs serve as our benchmark, which can be used to verify the reliability of the XAI application.

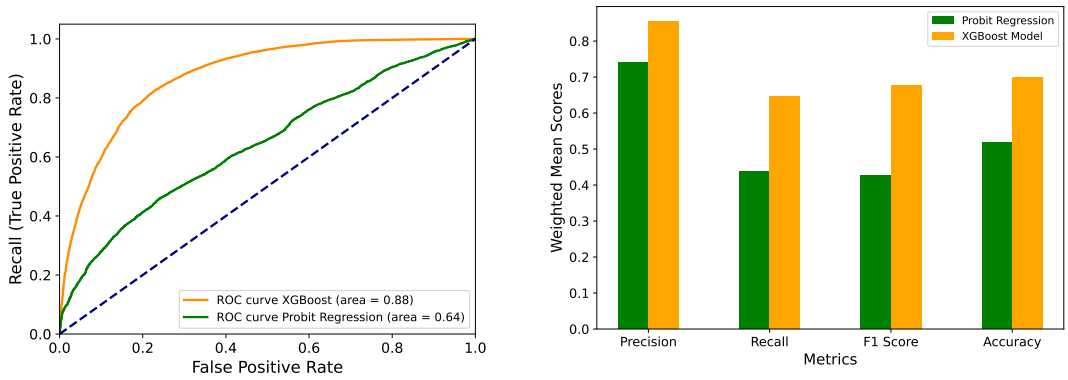
Table 8.1: **Regression Results.** The table presents the results of a probit regression analysis of the relationship between various factors, namely Stock Market, Interest Rates, Exchange Rates, PPrfm, and HHI. The table displays the coefficients and standard errors for six regression models. Additionally, it reports the pseudo R-squared values for each regression model, which measure the goodness of fit of the model. The regression analysis reveals key trends: a rise in Stock Market returns and PPrfm, representing past fund performance, correlates with improved fund performance, while conversely increased Interest and Exchange Rates are linked to decreased fund performance. Moreover, a higher HHI value, signifying market concentration, is associated with increased fund performance. The standard errors are robust to heteroskedasticity and autocorrelation-consistent (HAC) and are reported in parentheses. The significance levels for the coefficients are denoted as *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$. The data cover the period from January 2017 to September 2021.

.

Dependent Variable: Net Asset Value (NAV)						
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Stock Market	0.012***				0.021***	0.021***
	(0.000)				(0.000)	(0.000)
Exchange Rate		−0.066***			−0.051***	−0.050***
		(0.002)			(0.002)	(0.002)
PPrfm			0.121***		0.155***	0.128***
			(0.002)		(0.003)	(0.003)
Interest Rate				−0.217***	−0.549***	−0.633***
				(0.018)	(0.023)	(0.024)
HHI						0.190***
						(0.014)
Pseudo R-squared	−0.032	−0.025	0.004	−0.042	0.049	0.050
Observations	54558	54558	54558	54558	54558	54558

Table 8.2: **Performance Metrics of XGBoost Model.** The table presents the evaluation results of a binary classification model on a synthesized dataset of 28,8869 instances, with two classes, 0 and 1. The metrics shown include support, precision, recall, f1 score, and accuracy. For class 1, the model achieved a precision of 0.82, a recall of 0.88, and an f1 score of 0.85. This means that out of all instances predicted as class 1, 82% were actually class 1, and the model was able to correctly identify 88% of all instances that actually belong to class 1. The f1 score, which is a harmonic mean of precision and recall, was 0.85. For class 0, the values are computed analogously. The data covers the period from 2017 January to 2021 September.

Support	**Precision**	**Recall**	**f1 score**	**Model Accuracy**
0: 14394	0: 0.87	0: 0.80	0: 0.83	0.84
1: 14492	1: 0.82	1: 0.88	1: 0.85	



(i) ROC Curves. (ii) Metrics Scores Comparison

Figure 8.1: **Model performance comparison between Probit Regression vs XGBoost Model. (i) Receiver Operating Characteristic (ROC) curve analysis.** The x- and y-axes represent the false and true positive rates, respectively. The dashed line represents a random classifier. The orange line indicates the ROC curve for the XGBoost model, with an AUC of 0.87 demonstrating superior predictive accuracy compared to the probit regression model. The green line represents the ROC curve for the Probit Regression model, with an AUC of 0.70. **(ii) Metrics Scores Comparison.** The bar plots represent the weighted mean scores of precision, recall, F1, and accuracy metrics. The probit regression and XGBoost models are represented in green and orange, respectively. Across all metrics, the XGBoost model consistently outperforms the probit regression model.

Table 8.2 presents the performance metrics of our XGBoost application. For tuning the hyperparameters, we employ a 70/30 split between the training and testing sets in the XGBoost model. Note that Classes 0 and 1 represent bad and good fund performance, respectively. Counting the percentage of true positives in the XGBoost model, the precision metric for Class 0 (1) indicates that 87% (82%) of predicted of bad (good) performances were actually bad (good). Similarly to precision, the recall metric accommodating the concept of false negatives indicates the fraction that is correctly identified in each class. Last, the F1 score, an average between the previous two metrics, is balanced as a metric between false negatives and false positives. As our values of precision and recall metrics are equally high, it is natural to obtain well-balanced f1 scores across classes. This result establishes the validity of the machine learning model which is essential for the XAI application in the next step.

Building upon these findings, The figure 8.2 illustrates the comparative performance of XGBoost and Probit models across different decision thresholds, which directly influence the trade-off between sensitivity (Recall) and specificity (Precision). The top left plot reveals how the Precision of the models changes with the threshold, while the top right plot focuses on Recall. The bottom left plot represents the harmonic mean of Precision and Recall, or the F1 score, which can be a more balanced metric when

dealing with imbalanced classes. Finally, the bottom right plot demonstrates the overall Accuracy of the models at each threshold. Analyzing these plots together allows us to better understand the performance dynamics of the models and can guide the choice of an appropriate threshold for making final predictions.

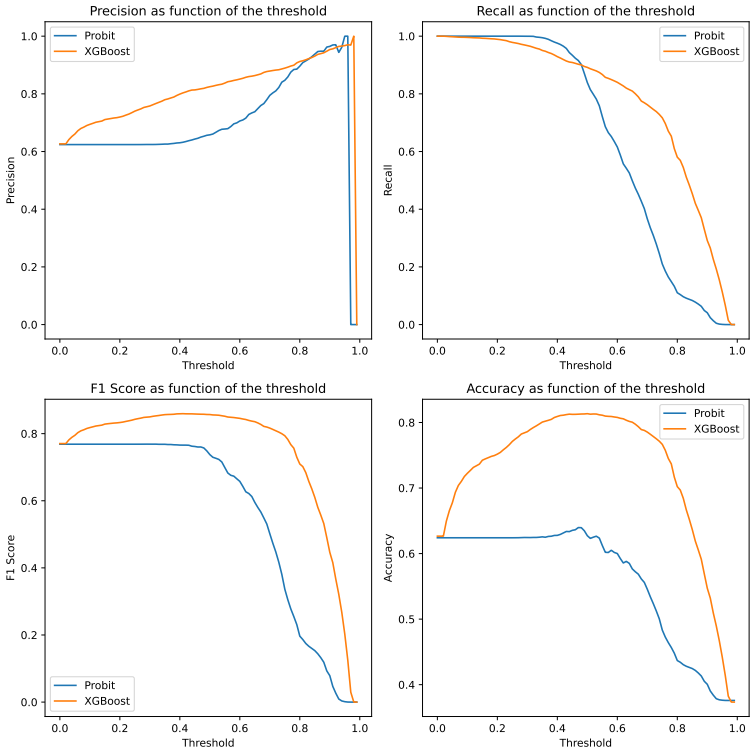


Figure 8.2: **Performance Comparison between XGBoost and Probit Models across Different Thresholds.** The four subplots provide a comprehensive view of the Precision, Recall, F1 Score, and Accuracy for both models. It facilitates the identification of an optimal threshold that balances these metrics.

Furthermore, Figure 8.2 illustrates the comparative performance of XGBoost and Probit models across different decision thresholds, which directly influence the trade-off between sensitivity (Recall) and specificity (Precision). The top left plot reveals how the Precision of the models changes with the threshold, while the top right plot focuses on Recall. Interestingly, the probit regression model's optimal threshold for Recall appears to be slightly less than 0.5, suggesting that a lower threshold results in a better balance between true positives and false negatives for this specific problem and dataset. This can be advantageous in scenarios where the cost of false negatives is higher than that of false positives. However, it is important to consider the model's performance across all metrics, as the XGBoost model generally demonstrates better performance in terms of Precision, F1 Score, and Accuracy. The bottom left plot represents the harmonic mean of Precision and Recall, or the F1 score, which can be a more balanced metric

when dealing with imbalanced classes. Finally, the bottom right plot demonstrates the overall Accuracy of the models at each threshold. Analyzing these plots together allows us to better understand the performance dynamics of the models and can guide the choice of an appropriate threshold for making final predictions.

Finally, in Figure 8.1 and Figure 8.2, we compare the performance of the two models, Probit regression and XGBoost, using an ROC curve[1] and metric comparison chart. XGBoost proves superior in capturing these highly non-linear relationships in explaining NAV changes, making it a more suitable choice for our XAI analysis going forward.

8.2 XAI Results Based on Input Features

Figure 8.3 is the SHAP summary plot which is a graphical representation of the feature importance for a machine learning model. It is based on the concept of Shapley values explained in chapter 2. The plot provides a clear representation of the impact of each feature on the model's prediction, and how it varies across instances of data. Each point on the plot corresponds to a single instance of fund data, and its position along the x-axis represents the input feature's impact on the model's output that is NAV. The y-axis on the left-hand side shows the feature names, Stock market, Exchange Rate, PPrfm, Interest rates, and HHI which are sorted in descending order by their importance; this allows for easy identification of the most impactful features.

Additionally, the y-axis on the right-hand side shows a color gradient that ranges from copper to black. This gradient represents the value of the feature for that particular data instance, with copper indicating a low feature value and black indicating a high feature value. By using this color-coding, we can quickly identify the relationship between feature values and their impact on the model's output.

For instance, when a feature has a high value, such as high stock market returns represented by black color with a positive SHAP value on the X-axis, it suggests that the feature has a strong positive influence on the fund performance. Conversely, when a feature with a high value, such as high values of exchange rate depicted in black color but with a negative SHAP value on the X-axis, it implies that the feature has a strong negative effect on the fund performance.

Similarly, when a feature has a low value, such as low values of exchange rate depicted in copper color with a positive SHAP value on the X-axis, it signifies that the feature has a strong positive impact on the fund performance. On the other hand, when a feature with a low value, such as low stock market returns in copper color but with a negative SHAP value on the X-axis, it indicates that the feature has a strong negative impact on the fund performance. With this in mind, for the fund data which covers the period from January 2017 to September 2021, our SHAP analysis revealed several important relationships between input features and the NAV of global open-ended funds, as shown in Figure 8.3. Firstly, we found a significant positive relationship between

[1]The ROC curve visualizes the performance of a binary classifier, showing the trade-off between sensitivity (true positive rate) and 1-specificity (false positive rate) as the classification threshold changes.

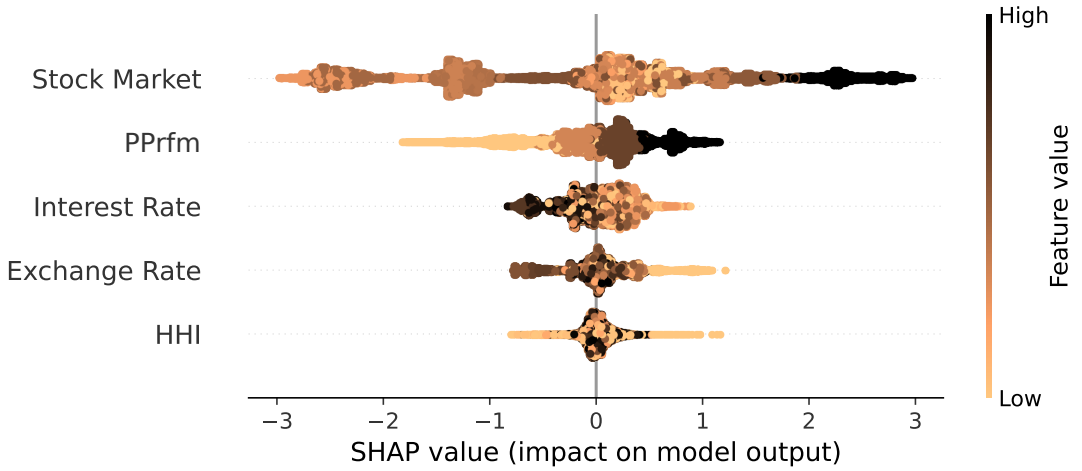


Figure 8.3: **SHAP Summary Plot.** The data covers the timeline from January 2017 to September 2021, illustrating feature importance and relationships with respect to fund performance. The plot showcases the impact of each feature on the model's prediction, represented by its position along the x-axis, while the left y-axis displays feature names sorted by importance. The right y-axis depicts a color gradient indicating feature values, ranging from copper to black. Analysis of the fund data reveals significant relationships such as a positive link between stock market returns and fund performance, a positive correlation with historical fund performance, a negative association between interest rates and fund performance, and a negative impact of exchange rates on fund performance. The relationship with HHI, however, remains ambiguous, as suggested by the copper color on both sides of the plot.

stock market returns and fund performance, with the SHAP values indicating that the stock market input feature was the most important factor in explaining the variation of NAV. These findings are consistent with previous studies that have also reported a positive relationship between stock market returns and fund performance [OP07]. In other words, a higher stock market typically results in a positive fund performance.

Secondly, we observed a positive correlation between historical fund performance and NAV. SHAP values revealed that Fund's Past Performance was the second most important variable in explaining the variation in NAV. Thirdly, our SHAP analysis indicated a negative relationship between interest rates and the fund performance. Interest rates were found to be the third most important variable in explaining the variation in NAV, consistent with prior research suggesting that interest rate changes affect the value of Net Asset in a fund data [LB00]. Fourthly, our analysis showed a negative relationship between exchange rates and fund performance, suggesting that when exchange rates decrease, the fund performance of the funds is more likely to increase. Finally, the relationship between HHI and fund performance was not clear when analysing the whole dataset: High values of HHI can positively and negatively impact the fund performance; the same holds for low values of HHI (see next section

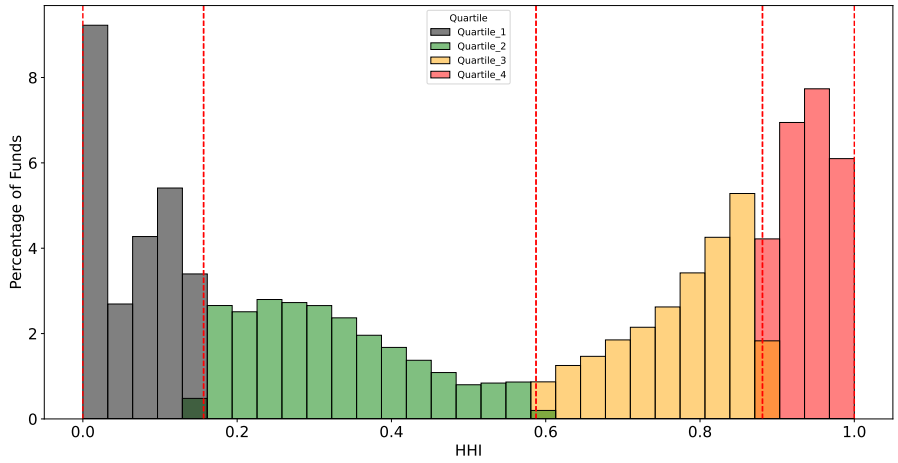


Figure 8.4: **H**istogram of HHI Quartiles. This figure plots Herfindahl-Hirschman Index (HHI) values on fund's portfolio holdings as a histogram. For the first quartile, the HHI value are lower than 0.156. For the second and third quartiles, the ranges of HHI values are (0.156, 0.587] and (0.587, 0.881], respectively. Finally, the fourth quartile contains HHI values exceeding 0.881. The x-axis represents HHI value, while the y-axis represents the percentage of funds. Quartiles are separated by vertical dashed lines.

8.3 for further analysis).

8.3 Influence of International Diversification

Building on the previous section, where we identified an ambiguous relationship between HHI and fund performance, we aim to further elucidate this ambiguity. To do so, we subdivided the fund dataset into four quartiles based on HHI values (see Figure 8.4). The associated summary statistics for these subsamples can be found in Table 8.3. Later we applied the XGBoost classifier to each subsample.

The model showcased robust performance, with an accuracy ranging between 0.83 and 0.84. Key metrics (precision, recall, F1 score, accuracy), are provided in Table 8.4. Specifically, the model attained an accuracy of 0.83 for Quartile 1 (HHI<0.16), 0.84 for Quartile 2 (HHI in the range 0.16–0.58) and Quartile 3 (HHI in the range 0.58–0.88), and 0.83 for Quartile 4 (HHI>0.88).

Figure 8.5 illustrates the SHAP results applied to each quartile. In the first quartile, where the funds are well-diversified, low HHI values lead to a lower fund performance. Conversely, in the fourth quartile (Figure 8.5iv), high HHI values are negatively correlated with performance. With regard to the other explanatory features, the relationships are largely consistent with Section 8.2.

Overall, our results indicate that fund performance varies across different HHI

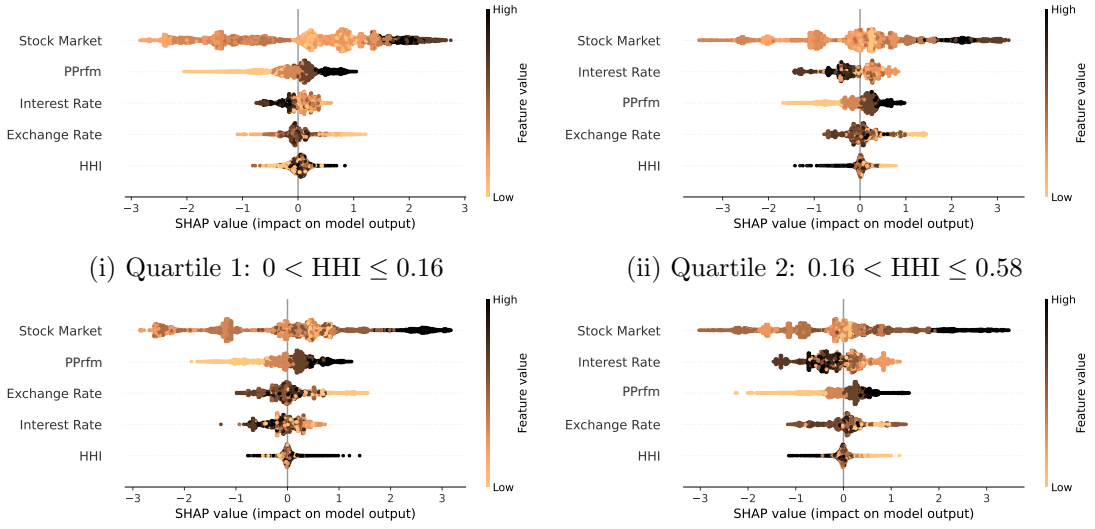
Table 8.3: **Summary Statistics**
The table presents summary statistics of funds data originating from all G10 countries, including the UK, Belgium, France, Canada, the Netherlands, Sweden, Switzerland, Germany, Italy, Japan, and the USA. The input features include stock market return (ST), interest rate (IR), exchange rate (ER), past performance (PPrfm), and Herfindahl Hirschman Index (HHI), with the target feature being Net Asset Value (NAV). The funds data is divided into multiple equal Quartiles based on HHI values, and summary statistics are provided for the whole HHI sample, as well as for the following subsamples: quartile 1 Subsample (HHI<0.16),quartile 2 Subsample (HHI>0.16 and HHI<0.58),quartile 3 Subsample (HHI>0.58 and HHI<0.88) and quartile 4 Subsample (HHI>0.88). The data cover the period from January 2017 to September 2021, comprising 19 quarters, January 2017 to December 2019, comprising 12 quarters, and January 2020 to September 2021 comprising a total of 7 quarters.

	January 2017 to September 2021						**January 2017 to December 2019**						**January 2020 to September 2021**					
	ST	**IR**	**ER**	**PPrfm**	**HHI**	**NAV**	**ST**	**IR**	**ER**	**PPrfm**	**HHI**	**NAV**	**ST**	**IR**	**ER**	**PPrfm**	**HHI**	**NAV**
	Whole Sample																	
mean	1.20	0.17	-0.73	2.48	0.52	0.62	0.99	0.26	-0.24	2.39	0.52	0.55	1.62	0.0	-1.72	2.67	0.53	0.76
std	4.35	0.23	3.05	1.07	0.35	0.48	2.91	0.23	3.04	1.11	0.35	0.50	6.29	0.06	2.82	0.95	0.36	0.43
min	-15.31	-0.22	-8.99	0.00	0.00	0.00	-7.62	-0.22	-8.99	0.00	0.00	0.00	-15.31	-0.19	-8.70	0.00	0.00	0.00
25%	0.52	0.01	-1.71	2.00	0.16	0.00	0.14	0.14	-1.55	2.00	0.16	0.00	2.14	0.00	-2.16	2.00	0.16	1.00
50%	2.05	0.14	-0.24	3.00	0.58	1.00	1.61	0.31	0.03	3.00	0.58	1.00	3.53	0.02	-1.22	3.00	0.59	1.00
75%	3.53	0.41	1.01	3.00	0.88	1.00	2.73	0.46	1.91	3.00	0.87	1.00	4.80	0.03	0.06	3.00	0.89	1.00
max	10.37	0.60	6.11	4.00	1.00	1.00	5.91	0.60	6.11	4.00	1.00	1.00	10.37	0.12	3.67	4.00	1.00	1.00
	quartile 1 Subsample (HHI<0.16)																	
mean	1.16	0.16	-0.74	2.44	0.06	0.61	0.96	0.24	-0.26	2.35	0.07	0.54	1.56	-0.01	-1.71	2.63	0.06	0.74
std	4.36	0.23	3.07	1.08	0.05	0.49	2.93	0.24	3.08	1.13	0.05	0.50	6.32	0.06	2.84	0.96	0.05	0.44
min	-15.31	-0.22	-8.99	0.00	0.00	0.00	-7.62	-0.22	-9.00	0.00	0.00	0.00	-15.31	-0.19	-8.70	0.00	0.00	0.00
25%	0.23	0.00	-1.94	2.00	0.00	0.00	-0.00	0.06	-1.72	2.00	0.00	0.00	2.05	-0.00	-1.97	2.00	0.00	0.00
50%	2.04	0.06	-0.24	3.00	0.06	1.00	1.61	0.26	0.04	2.00	0.07	1.00	3.41	0.02	-1.22	3.00	0.07	1.00
75%	3.53	0.38	1.49	3.00	0.11	1.00	2.79	0.42	1.92	3.00	0.11	1.00	4.80	0.03	0.06	3.00	0.11	1.00
max	10.37	0.60	6.11	4.00	0.16	1.00	5.91	0.60	6.11	4.00	0.16	1.00	10.37	0.12	3.67	4.00	0.16	1.00
	quartile 2 Subsample (HHI>0.16 and HHI<0.58)																	
mean	1.38	-0.02	-1.71	2.68	0.33	0.77	0.86	0.19	-0.2	2.37	0.32	0.55	1.38	-0.02	-1.71	2.68	0.33	0.77
std	6.73	0.08	2.9	0.95	0.11	0.42	2.98	0.25	3.03	1.14	0.11	0.50	6.73	0.08	2.90	0.95	0.11	0.42
min	-15.31	-0.19	-8.7	0.00	0.16	0.00	-7.62	-0.22	-9.00	0.00	0.16	0.00	-15.31	-0.19	-8.70	0.00	0.16	0.00
25%	1.96	-0.08	-2.16	2.00	0.24	1.00	-0.14	-0.09	-1.10	2.00	0.23	0.00	1.96	-0.08	-2.16	2.00	0.24	1.00
50%	3.53	0.02	-1.17	3.00	0.32	1.00	1.52	0.25	0.12	2.00	0.30	1.00	3.53	0.02	-1.17	3.00	0.32	1.00
75%	5.20	0.03	0.10	3.00	0.42	1.00	2.84	0.42	1.66	3.00	0.39	1.00	5.20	0.03	0.10	3.00	0.42	1.00
max	10.37	0.12	3.67	4.0	0.58	1.00	5.91	0.60	6.11	4.00	0.58	1.00	10.37	0.12	3.67	4.00	0.58	1.00
	quartile 3 Subsample (HHI>0.58 and HHI<0.88)																	
mean	1.19	0.21	-0.75	2.48	0.78	0.62	1.06	0.30	-0.29	2.39	0.78	0.55	1.45	0.01	-1.70	2.66	0.77	0.77
std	4.34	0.23	3.07	1.06	0.08	0.48	2.87	0.22	3.09	1.10	0.08	0.50	6.38	0.05	2.82	0.95	0.08	0.42
min	-15.31	-0.21	-9.00	0.00	0.58	0.00	-7.62	-0.21	-9.00	0.00	0.58	0.00	-15.31	-0.17	-8.70	0.00	0.58	0.00
25%	0.52	0.02	-1.80	2.00	0.72	0.00	0.23	0.19	-1.72	2.00	0.72	0.00	2.14	0.01	-1.94	2.00	0.72	1.00
50%	2.04	0.19	-0.24	3.00	0.79	1.00	1.61	0.34	0.04	3.00	0.79	1.00	3.41	0.02	-0.89	3.00	0.79	1.00
75%	3.53	0.42	1.01	3.00	0.84	1.00	2.57	0.47	1.92	3.00	0.84	1.00	4.80	0.03	0.06	3.00	0.84	1.00
max	10.37	0.60	6.11	4.00	0.88	1.00	5.91	0.60	6.11	4.00	0.88	1.00	10.37	0.12	3.67	4.00	0.88	1.00
	quartile 4 Subsample (HHI>0.88)																	
mean	1.43	0.22	-0.76	2.55	0.94	0.64	1.11	0.33	-0.23	2.46	0.94	0.57	2.05	0.01	-1.78	2.72	0.94	0.77
std	4.10	0.23	3.00	1.04	0.03	0.48	2.89	0.21	2.99	1.09	0.03	0.50	5.69	0.05	2.74	0.92	0.03	0.42
min	-15.31	-0.22	-9.00	0.00	0.88	0.00	-7.62	-0.22	-9.00	0.00	0.88	0.00	-15.31	-0.19	-8.70	0.00	0.88	0.00
25%	0.52	0.02	-1.54	2.00	0.91	0.00	0.52	0.25	-1.16	2.00	0.91	0.00	2.44	0.01	-1.80	2.00	0.91	1.00
50%	2.34	0.19	-0.24	3.00	0.94	1.00	1.61	0.38	0.12	3.00	0.94	1.00	3.53	0.02	-1.22	3.00	0.94	1.00
75%	3.53	0.42	0.98	3.00	0.97	1.00	2.79	0.47	1.92	3.00	0.97	1.00	4.80	0.03	0.06	3.00	0.97	1.00
max	10.37	0.60	6.11	4.00	1.00	1.00	5.91	0.60	6.11	4.00	1.00	1.00	10.37	0.12	3.67	4.00	1.00	1.00

groups, with a moderate level of diversification having the potential to enhance fund performance.

Table 8.4: **Performance metrics of XGBoost model**
The table shows the evaluation results of a XGBoost binary classification model on four distinct subsamples of a dataset, classified according to their Herfindahl-Hirschman Index (HHI) values. Each subsample's HHI range and performance metrics, including precision, recall, F1 score, support, and model accuracy, are listed in the table. The metrics for both classes (0 and 1) are reported, along with their corresponding support values (number of samples in each class). The dataset covers the period between January 2017 and September 2021.

HHI	Total Instances	Support	Precision	Recall	f1 score	Model Accuracy
Q1 Subsample (<0.16)	7806	0- 3867 1- 3939	0- 0.85 1- 0.83	0- 0.82 1- 0.86	0- 0.83 1- 0.84	0.83
Q2 Subsample (0.16 to 0.58)	8028	0- 4037 1- 3991	0- 0.85 1- 0.83	0- 0.82 1- 0.85	0- 0.84 1- 0.84	0.84
Q3 Subsample (0.58 to 0.88)	7992	0- 3919 1- 4073	0- 0.87 1- 0.82	0- 0.80 1- 0.88	0- 0.83 1- 0.85	0.84
Q4 Subsample (>0.88)	8177	0- 4137 1- 4040	0- 0.87 1- 0.81	0- 0.79 1- 0.88	0- 0.83 1- 0.84	0.83



(i) Quartile 1: $0 < \text{HHI} \leq 0.16$

(ii) Quartile 2: $0.16 < \text{HHI} \leq 0.58$

(iii) Quartile 3: $0.58 < \text{HHI} \leq 0.88$

(iv) Quartile 4: $0.88 < \text{HHI} \leq 1.00$

Figure 8.5: **SHAP Summary Plot in HHI Quartiles.** This figure illustrates subsamples of funds by the HHI quartiles from January 2017 to September 2021. The high (low) HHI values within each quartile are marked by dark (light) color. In the first quartile Figure 8.5i, the light tail on the left and dark tail on the right shows a positive correlation between the HHI values and fund performance, which is found to be reversed in the forth quartile in Figure 8.5iv, implying underperformance of funds with extreme HHI values that are too close to 0 or 1. Note that relationships for the other features are consistent with Section 8.2.

8.4 Robustness Tests

Our empirical analysis provides consistent results for fund performance evaluation. In this section, we will examine whether the model is robust in the cross-section and time series. Specifically, we check if the findings are driven by specific countries or time periods.

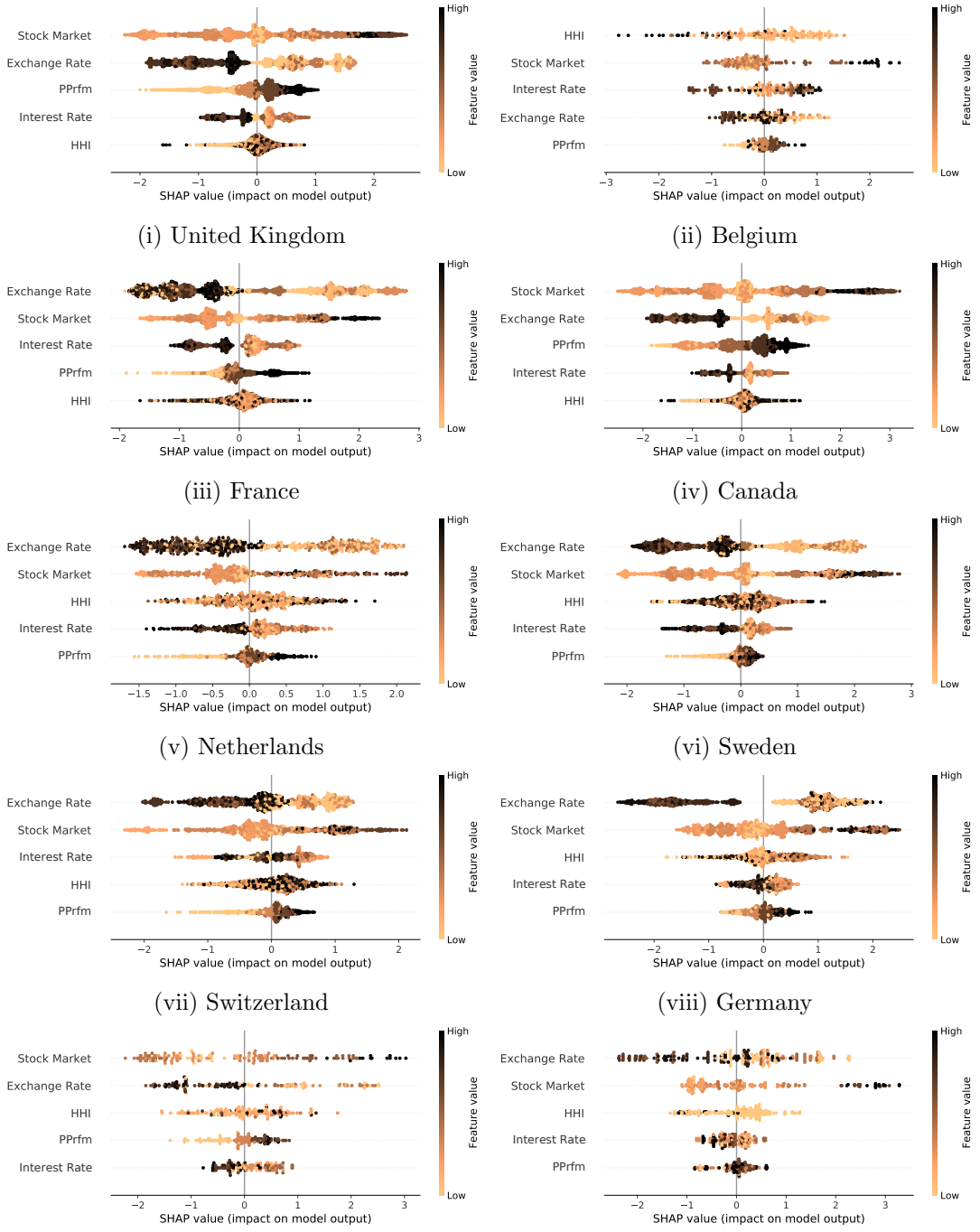
8.4.1 Country-Level Robustness

To investigate whether the results are driven by one single country, we designed two exercises. First, we retrained the XGBoost model with subsamples of 10 out of 11 countries. The performance metrics, namely precision, recall, and F1-score, remain fairly stable across different subsets of countries. Note that the model's accuracy is remarkably stable, ranging over the narrow range between 0.83 and 0.84. Comprehensive details of these metrics can be found in Table 8.5 Additionally, we calculate the Pearson correlation between fund performance and features for each country. Table 8.6 shows that the correlations are consistent across countries, albeit with a few exceptions. For instance, the correlation with past performance in Sweden, Germany, and Japan is found to have the opposite sign to the whole-sample analysis in both the XGBoost and conventional linear regression models. This is likely due to the small samples for these countries.

In addition, we present a detailed examination of the input financial variables and their effects on fund performance for each G10 country using SHAP summary plots as shown in Figure 8.6. Each subplot in the SHAP summary plot corresponds to a specific G10 country, highlighting the impact of input financial variables on fund performance within that country.

Figure 8.6i displays the SHAP summary plot for funds originating from the United Kingdom, indicating that high stock market returns, low exchange rates, positive past performance, and low interest rates have a clear positive impact on fund performance. However, the effect of HHI values remains unclear. Similarly, Figure 8.6ii for Belgium-originated funds shows that low HHI values, high stock returns, low exchange rates, and positive fund performance positively affect fund performance, but the influence of interest rates is ambiguous.

Figures 8.6iii, 8.6iv, 8.6v, 8.6vi, 8.6vii, 8.6viii, and 8.6ix all exhibit similar patterns for funds from France, Canada, Netherlands, Sweden, Switzerland, Germany, and Italy, respectively. High stock market returns, low exchange rates, positive past performance, and low interest rates contribute to a positive effect on fund performance. Nevertheless, the impact of HHI values remains unclear in these cases. In contrast, Figure 8.6x for Japanese-originated funds demonstrates that high stock market returns, low exchange rates, low HHI values, and low interest rates have a clear positive effect on fund performance, but the influence of past performance is not well-defined.



(i) United Kingdom (ii) Belgium

(iii) France (iv) Canada

(v) Netherlands (vi) Sweden

(vii) Switzerland (viii) Germany

(ix) Italy (x) Japan

Figure 8.6: **SHAP Summary Plot(G10 Countries).** SHAP explanation for the funds originated from G10 countries: United Kingdom, Belgium, France, Canada, Netherlands, Sweden, Switzerland, Germany, Italy, Japan

Table 8.5: **Performance of XGBoost model: robustness test at the country level**
This table reports the metrics of the XGBoost model's performance, obtained from the leave-one-out cross-validation analysis. In each column, the sign '¬' indicates the country that is excluded from our robustness test, while the column of 'G10' reports the whole sample results as found in Table 8.3.

		Metrics			
Countries	**Observations**	**Precision**	**Recall**	**f1 score**	**Model Accuracy**
G10	77940	0- 0.87 1- 0.82	0- 0.80 1- 0.88	0- 0.83 1- 0.85	0.84
¬ BEL	77670	0- 0.85 1- 0.81	0- 0.79 1- 0.86	0- 0.83 1- 0.84	0.83
¬ CAN	63612	0- 0.87 1- 0.81	0- 0.79 1- 0.88	0- 0.83 1- 0.84	0.83
¬ CHE	75546	0- 0.86 1- 0.81	0- 0.80 1- 0.87	0- 0.83 1- 0.84	0.83
¬ FRA	73422	0- 0.85 1- 0.82	0- 0.81 1- 0.86	0- 0.83 1- 0.84	0.83
¬ GBR	69606	0- 0.86 1- 0.81	0- 0.80 1- 0.88	0- 0.83 1- 0.84	0.84
¬ DEU	75456	0- 0.86 1- 0.81	0- 0.79 1- 0.88	0- 0.83 1- 0.84	0.83
¬ ITA	77724	0- 0.86 1- 0.81	0- 0.80 1- 0.87	0- 0.83 1- 0.84	0.83
¬ JPN	77634	0- 0.85 1- 0.82	0- 0.81 1- 0.86	0- 0.83 1- 0.84	0.83
¬ NLD	76986	0- 0.86 1- 0.81	0- 0.80 1- 0.87	0- 0.83 1- 0.84	0.83
¬ SWE	74664	0- 0.86 1- 0.82	0- 0.81 1- 0.86	0- 0.83 1- 0.84	0.84
¬ USA	37080	0- 0.85 1- 0.82	0- 0.82 1- 0.86	0- 0.83 1- 0.84	0.83

8.4.2 Influence of COVID-19

Following the cross-sectional robustness tests, we next investigate the dimension of the time series. Specifically, we focus on the influence of COVID-19 by dividing the data into two subperiods: before and after December 2019. The performance metrics of XGBoost model before and after the COVID-19 outbreak can be found in Table 8.7 Figure 8.7i

Table 8.6: **Correlation between Fund Performance and the Features.** The figure represents Pearson Correlation coefficients and associated p-values for several key financial parameters from January 2017 to September 2021. These parameters include the stock market (defined as the logarithmic return), interest rates, exchange rates, PPrfm (a measure summarizing fund performance over the most recent four quarters), and the Herfindahl-Hirschman Index (HHI, an indicator of market concentration and diversification). Significance levels are as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The superscript 'a' signifies values represented as 0.000 or −0.000, namely minimal amounts that round to zero at the third decimal place.

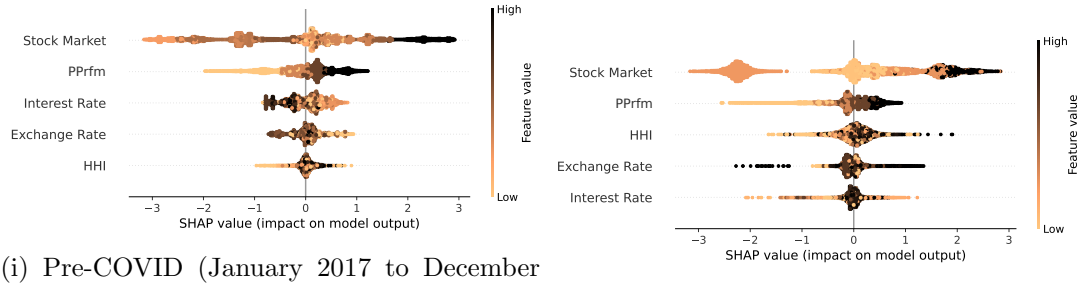
			Features				
Country	**Sample Size**	**Model Accuracy**	**Stock Market**	**Interest Rates**	**Exchange Rate**	**PPrfm**	**HHI**
G10 ex. USA	37080	0.84	0.140*** (0.000)	−0.090*** (0.000)	−0.240*** (0.000)	0.070*** (0.000)	0.010*** (0.000)
USA	40860	0.84	0.150*** (0.000)	−0.140*** (0.000)	−0.200*** (0.000)	0.090*** (0.000)	0.020*** (0.000)
BEL	270	0.68	0.140** (0.022)	−0.020 (0.713)	−0.150** (0.015)	0.010 (0.882)	−0.120** (0.043)
CAN	14328	0.85	0.150*** (0.000)	−0.110*** (0.000)	−0.260*** (0.000)	0.140*** (0.000)	0.010* (0.099)
CHE	2394	0.80	0.140*** (0.000)	−0.040** (0.041)	−0.180*** (0.000)	0.040* (0.059)	0.050** (0.025)
FRA	4518	0.84	0.130*** (0.000)	−0.080*** (0.000)	−0.260*** (0.000)	0.020** (0.029)	−0.010* (0.081)
GBR	8334	0.83	0.150*** (0.000)	−0.070*** (0.000)	−0.230*** (0.000)	0.050*** (0.000)	−0.010** (0.040)
DEU	2484	0.84	0.110*** (0.000)	−0.070*** (0.000)	−0.260*** (0.000)	−0.040** (0.033)	0.000^a (0.957)
ITA	216	0.73	0.220*** (0.001)	−0.140** (0.042)	−0.200*** (0.003)	0.040 (0.516)	0.080 (0.221)
JPN	306	0.74	0.230*** (0.000)	−0.100* (0.075)	−0.190*** (0.001)	−0.070 (0.244)	−0.050 (0.359)
NLD	954	0.80	0.110*** (0.000)	−0.140*** (0.000)	−0.220*** (0.000)	0.020 (0.617)	0.030 (0.351)
SWE	3276	0.86	0.130*** (0.000)	−0.060*** (0.000)	−0.270*** (0.000)	−0.040*** (0.001)	-0.000^a (0.819)

Table 8.7: **Performance metrics for the pre-COVID and COVID periods**
This table presents Summary of Experimental Results on Global open-ended funds originating from all G10 countries: the UK, Belgium, France, Canada, the Netherlands, Sweden, Switzerland, Germany, Italy, and Japan.We experimented over three different timelines. For each timeline, we carried analysis with we experimented with whole HHI and 4 HHI Sub samples. The timelines were:2017 January to 2019 December, with 2017 January to 2018 December as the training data and 2019 January to 2019 December as the test data; and 2020 January to 2021 September, with 2020 January to 2020 December as the training and 2021 January to 2021 September as the test data.

Sno	**HHI**	**Precision**	**Recall**	**f1 score**	**Support**	**Model Accuracy**
Panel A: 2017-January to 2019-December						
1	Whole Sample (0-1)	0- 0.85 1- 0.77	0- 0.75 1- 0.86	0- 0.79 1- 0.81	0- 8712 1- 8613	0.80
2	Q1 Subsample (<0.16)	0- 0.82 1- 0.82	0- 0.81 1- 0.83	0- 0.82 1- 0.83	0- 2276 1- 2368	0.82
3	Q2 Subsample (0.16 to 0.58)	0- 0.82 1- 0.77	0- 0.76 1- 0.83	0- 0.79 1- 0.80	0- 2380 1- 2350	0.79
4	Q3 Subsample (0.58 to 0.88)	0- 0.87 1- 0.77	0- 0.73 1- 0.89	0- 0.80 1- 0.83	0- 2396 1- 2404	0.80
5	Q4 Subsample (>0.88)	0- 0.84 1- 0.77	0- 0.75 1- 0.86	0- 0.79 1- 0.81	0- 2407 1- 2382	0.79
Panel B: 2020-January to 2021-September						
1	Whole Sample (0-1)	0- 0.90 1- 0.84	0- 0.84 1- 0.90	0- 0.86 1- 0.87	0- 6561 1- 6474	0.86
2	Q1 Subsample (<0.16)	0- 0.90 1- 0.81	0- 0.78 1- 0.91	0- 0.84 1- 0.86	0- 1584 1- 1579	0.85
3	Q2 Subsample (0.16 to 0.58)	0- 0.90 1- 0.86	0- 0.85 1- 0.91	0- 0.88 1- 0.88	0- 1625 1- 1660	0.88
4	Q3 Subsample (0.58 to 0.88)	0- 0.91 1- 0.84	0- 0.83 1- 0.92	0- 0.87 1- 0.88	0- 1592 1- 1601	0.87
5	Q4 Subsample (>0.88)	0- 0.90 1- 0.84	0- 0.82 1- 0.91	0- 0.86 1- 0.87	0- 1674 1- 1720	0.86

presents the SHAP summary plot prior to the onset of the COVID-19 pandemic, showing a similar pattern to the whole-sample results in Figure 8.3. Conversely, the COVID-19 subperiod in Figure 8.7ii shows a much higher importance, but with an opposite influence, of Interest Rate on fund performance. While this finding of low Interest Rate values associating with poor fund performance is not surprising, it does underline the importance of having a complete sample for the XAI analysis that is balanced between crisis and non-crisis periods.

Furthermore, from the Figure 8.7i the relationship between the HHI and fund performance appeared unclear during both the pre-COVID era (January 2017 - December 2019) and the COVID-19 period (January 2020 - September 2021). The linkages between other input factors, such as interest rates and exchange rates, also exhibited ambiguity



(i) Pre-COVID (January 2017 to December 2019)

(ii) COVID (January 2020 to September 2021)

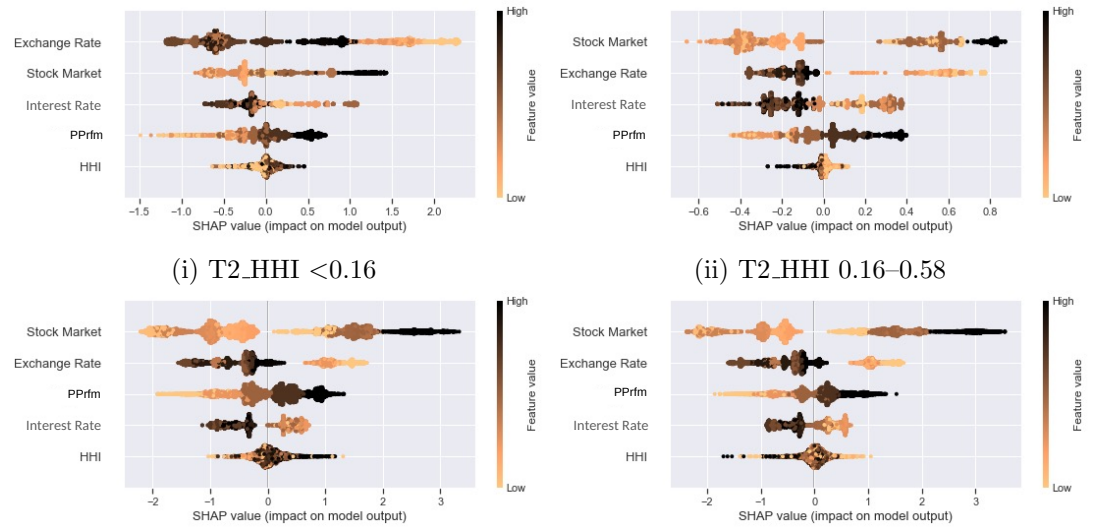
Figure 8.7: **SHAP Summary Plot in Subperiods.** The figure depicts two distinct periods: pre-COVID (on the left) and COVID (on the right). In each figure, the x-axis represents SHAP values. The color gradient on the right y-axis indicates the values of the various features, while the left y-axis lists these features according to their significance.

during the pandemic, possibly due to economic turmoil. To further explore these unclear relationships, we subdivided the fund dataset into four quartiles based on HHI values (refer to Figure 8.4). The XGBoost classifier was then applied to each subsample. Performance metrics for the pre-COVID and COVID periods are presented in Table 8.7.

pre-COVID period

Our analysis reveals that the relationship between portfolio diversification and fund performance is dependent on the level of concentration. In the lower quartile, we found that higher HHI values are associated with a greater likelihood of better fund performance, indicating that excessive portfolio diversification can have a negative impact on fund performance. This is demonstrated in SHAP summary plot 8.8i. In contrast, in the higher quartile, we found that higher HHI values are associated with a lower likelihood of better fund performance, suggesting that excessive portfolio concentration can also lead to a negative impact on fund performance. This is shown in SHAP summary plot 8.8iv. These findings highlight the importance of finding an appropriate level of portfolio diversification that balances the benefits of risk reduction with the negative effects of over-diversification and concentration.

In medium Quartiles, Figure 8.8ii shows that lower HHI values are related to a higher possibility of an increase in fund performance, while higher HHI values are associated with a lower likelihood of better fund performance. This suggests that an appropriate or balanced level of diversification can yield a positive impact on fund performance. On the other hand, Figure 8.8iii indicates that the relationship between HHI and fund performance is not very clear. Overall, we find that a moderate HHI value, indicating an appropriate level of portfolio diversification, can have a positive impact on fund performance, while excessively high levels of diversification and excessive concentration can both have negative impacts. These results suggest that the relationships are consistent irrespective of different samples of data extracted from different timelines.



(i) T2_HHI <0.16 (ii) T2_HHI 0.16–0.58

(iii) T2_HHI 0.58–0.88 (iv) T2_HHI>0.88

Figure 8.8: **SHAP Summary Plot (pre-COVID).** Figure illustrate different Quartiles based on HHI values, covering the time period from January 2017 to December 2019. In (a), which corresponds to HHI values less than 0.16, higher HHI values are associated with a higher likelihood of positive fund performance, indicating the negative impact of excessive portfolio diversification. In (b), the quartile ranging from 0.16 to 0.58, lower HHI values are linked to a higher likelihood of positive fund performance, highlighting the importance of moderate portfolio diversification. In (c), covering the range from 0.58 to 0.88, the relationship between HHI and fund performance is not clearly defined. Lastly, in (d), for HHI values exceeding 0.88, higher HHI values indicate a lower likelihood of positive fund performance, emphasizing the negative effect of excessive portfolio concentration.

COVID period

Throughout the COVID-19 pandemic, unlike previous periods, our analysis did not detect a distinct relationship between HHI and fund performance within the medium quartiles. Furthermore, in both lower and upper quartiles, high levels of diversification and over concentration could potentially have detrimental impacts, as illustrated in Figures 8.9i, 8.9ii, 8.9iii, and 8.9iv.

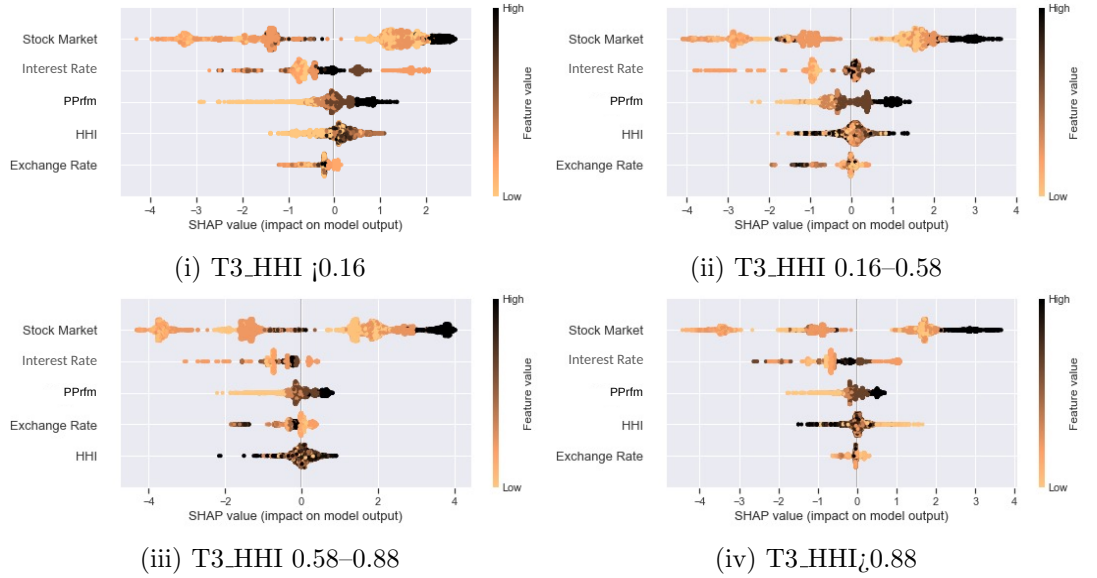
Overall, our analysis suggests that stock market returns continue to have the strongest positive impact on fund performance. past performance also had a positive impact towards fund performance. Exchange rates did not significantly impact fund performance during this period, and the relationship between interest rates, HHI, and fund performance was not clear. Additionally, our analysis of the relationship between HHI and fund performance suggests that the impact of HHI may be more nuanced during periods of economic downturn.

In summary, the moderate level of portfolio diversification, as indicated by a moderate

Table 8.8: **Correlation between Fund Performance and the Features**
The figure represents Pearson Correlation coefficients and associated p-values for several key financial parameters from January 2017 to December 2019. These parameters include the stock market (defined as the logarithmic return), interest rates, exchange rates, PPrfm (a measure summing up fund performance over the last four quarters), and the Herfindahl-Hirschman Index (HHI, an indicator of market concentration and diversification). Significance levels as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

			Features				
Country	**Sample Size**	**Model Accuracy**	**Stock Market**	**Interest Rates**	**Exchange Rate**	**PPrfm**	**HHI**
G10	51960	0.82	0.200***	-0.100***	-0.130***	0.100***	0.020***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
BEL	180	0.66	0.180**	-0.040	-0.050	0.080	-0.160**
			(0.022)	(0.582)	(0.540)	(0.292)	(0.024)
CAN	9552	0.82	0.230***	-0.130***	-0.120***	0.150***	0.010
			(0.000)	(0.000)	(0.000)	(0.000)	(0.832)
CHE	1596	0.77	0.160***	-0.090***	-0.060***	0.050**	0.040*
			(0.000)	(0.000)	(0.000)	(0.037)	(0.095)
FRA	3012	0.79	0.130***	-0.080**	-0.260***	0.020**	-0.010
			(0.000)	(0.000)	(0.000)	(0.029)	(0.82)
GBR	5556	0.81	0.160***	-0.090***	-0.220***	0.050***	-0.010
			(0.000)	(0.000)	(0.000)	(0.000)	(0.511)
DEU	1656	0.82	0.150***	-0.080***	-0.130***	-0.060**	-0.020
			(0.000)	(0.000)	(0.000)	(0.011)	(0.534)
ITA	144	0.72	0.260***	-0.160*	-0.150*	0.040	0.000
			(0.000)	(0.061)	(0.066)	(0.636)	(0.990)
JPN	204	0.84	0.190***	-0.170**	-0.030	-0.050	-0.040
			(0.000)	(0.015)	(0.613)	(0.422)	(0.589)
NLD	636	0.75	0.160***	-0.150***	-0.170***	0.040	0.080**
			(0.000)	(0.000)	(0.000)	(0.371)	(0.044)
SWE	2184	0.83	0.200***	-0.040*	-0.130***	-0.050**	0.000
			(0.000)	(0.092)	(0.000)	(0.011)	(0.535)
USA	27240	0.82	0.220***	-0.080***	-0.100***	0.120***	0.020***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

HHI value, along with high stock market returns, low exchange rates, low interest rates, and positive past performance, can have a positive impact on fund performance. However, excessively high levels of diversification and excessive concentration can both have negative impacts. During the January 2017 to September 2021 and January 2017 to December 2019 periods, we found similar trends in the relationships between input features and fund performance. Specifically, moderate diversification, indicated by a moderate HHI value, along with high stock market returns, low exchange rates, low interest rates, and positive past performance, can have a positive impact on fund performance. However, during the January 2020 to September 2021 period, our analysis suggests that the relationship between HHI and fund performance may be more complex



(i) T3_HHI ¡0.16

(ii) T3_HHI 0.16–0.58

(iii) T3_HHI 0.58–0.88

(iv) T3_HHI¿0.88

Figure 8.9: **SHAP Summary Plot (COVID-19 period).** Figure illustrate different Quartiles based on HHI values, covering the time period from January 2020 to September 2021. In (a), which corresponds to HHI values less than 0.16, higher HHI values are associated with a higher likelihood of positive fund performance, indicating the negative impact of excessive portfolio diversification. In (b), the quartile ranging from 0.16 to 0.58, the relationship is inconclusive. In (c), covering the range from 0.58 to 0.88, the relationship between HHI and fund performance is not clearly defined. Lastly, in (d), for HHI values exceeding 0.88, higher HHI values indicate a lower likelihood of positive fund performance, emphasizing the negative effect of excessive portfolio concentration.

during periods of economic downturn. While the importance of stock market returns and past performance remained consistent, our analysis suggests that the impact of HHI on fund performance may be more nuanced during economic downturns. Overall, our analysis underscores the importance of continually monitoring the relationships between input features and fund performance, as they may vary over time and across different economic conditions.

8.5 Conclusion

Explainable artificial intelligence (XAI) techniques currently show great potential to enrich information content in financial and economic studies. On one hand, our findings supplement [GKX20] by providing further information on the importance and directional influence specific to each explanatory feature. We showed that a state-of-the-art machine learning model (specifically, XGBoost) together with XAI techniques can produce reliable and interpretable results in financial studies. On the other, we leveraged the advantage of machine learning models in analyzing highly nonlinear problems with large datasets as

in [LLS23]. We further examined the diversification implications for portfolio holdings across the G10 countries, finding good performance of equity funds to be associated with a moderate degree of diversification. Moreover, our results as implied by XAI were able to replicate statistical characteristics such as signs and significance of the benchmark linear regression model at both aggregate and country levels. We therefore advocate the benefits of applying XAI to complex questions that remain open in the finance literature. Additionally, adaptations of the XAI approach that cater to the endogenous or exogenous nature: Controllable versus uncontrollable features in the context of [KLS+22] of input variables form a prospective direction for future research with the enhancement of interpretability.

Part III

XAI in the Socio-Economic Domain

Chapter 9

Socio-Economic Datasets: Preliminary Analysis and CAFA-driven Interpretations

Contents

9.1 Criteria and Rationale for Dataset Selection

In the domain of socio-economic research, the selection of datasets plays a crucial role in developing fair, unbiased, and socially responsible AI models. The criteria for dataset selection in this study are carefully designed to ensure the development of explainable AI techniques that can address issues of fairness, accountability, and transparency in socio-economic decision-making processes. Key considerations include the relevance of the datasets to real-world socio-economic problems, the presence of sensitive attributes, and the potential for bias and discrimination in the data [BS16].

The chosen datasets must contain a mix of demographic, social, and economic features that are representative of the complexities found in real-world scenarios. They should include both continuous and categorical variables, enabling the demonstration of the capabilities of the proposed explainable AI techniques in handling diverse data types.

The datasets should also have a sufficient number of instances to allow for meaningful analysis and the training of robust models [GMV+18].

Accessibility and public availability of the datasets are prioritized to ensure reproducibility and promote open research. The datasets should have a track record of being used in relevant studies, allowing for benchmarking and comparison with existing approaches. Furthermore, the datasets must adhere to ethical guidelines and privacy regulations, ensuring that sensitive information is adequately protected [HHN+18].

Expanding upon these criteria, the following factors are considered in detail:

1. **Socio-economic relevance**: The datasets should be directly related to socio-economic decision-making processes, such as credit lending, hiring, or criminal risk assessment. They should contain features that are commonly used in these domains and have a significant impact on individuals' lives [RR14].

2. **Presence of sensitive attributes**: The datasets should include sensitive attributes such as race, gender, age, or other protected characteristics. This allows for the examination of potential biases and discrimination in the decision-making process and the development of techniques to mitigate these issues [Žli17].

3. **Potential for bias and discrimination**: The datasets should have a history of being studied in the context of fairness and discrimination. They should exhibit patterns or biases that are representative of real-world challenges in socio-economic decision-making [Cho17].

4. **Feature diversity**: The datasets should contain a mix of demographic, social, and economic features, including both continuous and categorical variables. This diversity enables the evaluation of the proposed techniques' ability to handle different data types and capture complex relationships [OCDK19].

5. **Data volume and quality**: The datasets should have a sufficient number of instances to support robust analysis and model training. They should also be of high quality, with minimal missing values and well-documented data collection processes [GMV+18].

6. **Public availability and accessibility**: The datasets should be publicly available and easily accessible to researchers, ensuring reproducibility and facilitating open collaboration. They should have permissive licenses that allow for academic use and publication [HHN+18].

7. **Prior use in research**: The datasets should have a history of being used in relevant studies, particularly in the context of fairness, explainability, and socio-economic decision-making. This allows for benchmarking and comparison with existing approaches [ZRHL21].

8. **Ethical and privacy considerations**: The datasets must adhere to ethical guidelines and privacy regulations. They should be properly anonymized and have the necessary permissions for use in research [BS16].

Based on these criteria, three widely-used datasets in the socio-economic domain were selected for this study:

1. UCI Adult Income Dataset [Koh96]
2. Statlog German Credit Dataset [Hof00]
3. ProPublica's COMPAS Dataset [ALMK16]

These datasets cover various aspects of socio-economic decision-making, including income prediction, credit risk assessment, and criminal risk assessment. They contain a mix of sensitive attributes and have been extensively studied in the context of fairness and discrimination. By applying the proposed explainable AI techniques to these datasets, we aim to gain insights into the factors influencing the decisions and develop approaches to enhance fairness and transparency in socio-economic models.

9.2 Overview of Chosen Datasets

9.2.1 UCI Adult Income Dataset

The UCI Adult Income Dataset, also known as the "Census Income" dataset, is a widely used benchmark dataset in the field of socio-economic research [Koh96]. The dataset contains information extracted from the 1994 US Census database and consists of 48,842 instances with 14 attributes. The primary task associated with this dataset is to predict whether an individual's income exceeds $50,000 per year based on various demographic and employment-related factors.

The dataset includes a mix of categorical and numerical attributes, such as age, workclass, education, marital status, occupation, race, sex, capital gain/loss, hours per week, and native country. This diverse set of features allows for a comprehensive analysis of the factors influencing income levels and provides an opportunity to examine potential biases and discrimination in income prediction models.

The UCI Adult Income Dataset has been extensively studied in the context of algorithmic fairness and has served as a benchmark for evaluating various machine learning algorithms. Its socio-economic relevance, large sample size, and well-documented attributes make it a valuable resource for understanding and mitigating biases in income prediction tasks.

9.2.2 German Credit Dataset

The German Credit Dataset, obtained from Professor Dr. Hans Hofmann at the University of Hamburg, is another widely used dataset in the domain of credit risk assessment [Hof00]. The dataset consists of 1,000 instances, each representing a person who takes credit from a bank. The goal is to predict the credit risk associated with each individual, classified as either good or bad credit risks.

Feature	Type
Age	Continuous
Workclass	Categorical
fnlwgt	Continuous
Education	Categorical
Education-num	Continuous
Marital-status	Categorical
Occupation	Categorical
Relationship	Categorical
Race	Categorical
Sex	Categorical
Capital-gain	Continuous
Capital-loss	Continuous
Hours-per-week	Continuous
Native-country	Categorical
Income (Target)	Binary

Table 9.1: UCI Adult Income Dataset Features

The dataset contains 20 attributes, including both categorical and integer variables. These attributes capture various aspects of the credit applicant, such as their account status, credit history, purpose of the loan, credit amount, savings account/bonds, employment status, personal status, other debtors/guarantors, property ownership, age, housing arrangements, number of existing credits, job category, and telephone ownership.

	Predicted	
Actual	Good	Bad
Good	0	1
Bad	5	0

Table 9.2: Cost Matrix for Credit Risk Prediction

The rows represent the actual classification, and the columns represent the predicted classification. It is considered worse to classify a customer as good when they are bad (cost of 5) than it is to classify a customer as bad when they are good (cost of 1). This cost matrix reflects the real-world implications of credit risk assessment, where the consequences of misclassifying a bad credit risk as good are more severe than the reverse.

Table 9.3 presents a detailed description of the attributes in the German Credit Dataset.

The German Credit Dataset provides a valuable resource for evaluating the fairness and explainability of credit risk assessment models. Its well-defined attributes, real-

Feature	Type
Checking Account Status	Categorical
Credit Duration	Integer
Credit History	Categorical
Loan Purpose	Categorical
Credit Amount	Integer
Savings Account/Bonds	Categorical
Employment Duration	Categorical
Installment Rate	Integer
Personal Status and Sex	Categorical
Other Debtors/Guarantors	Categorical
Residence Duration	Integer
Property Ownership	Categorical
Age	Integer
Other Installment Plans	Categorical
Housing Situation	Categorical
Number of Existing Credits	Integer
Job Type	Categorical
Number of People Liable	Integer
Telephone Availability	Binary
Foreign Worker	Binary
Credit Risk (Target)	Binary

Table 9.3: German Credit Dataset Features

world relevance, and the inclusion of a cost matrix make it particularly suitable for investigating the potential biases and discrimination in algorithmic decision-making within the financial domain.

9.2.3 ProPublica's COMPAS Dataset

ProPublica's COMPAS dataset is a landmark dataset in the study of algorithmic fairness and bias in criminal justice risk assessment [ALMK16]. The dataset contains information on 6,172 individuals who were arrested and assessed for their likelihood of recidivism (i.e., committing a future crime) using the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk assessment tool.

The dataset includes various attributes such as the individual's age, gender, ethnicity, criminal history, and the COMPAS risk scores. The target variable is a binary indicator of whether the individual actually recidivated within two years of their initial assessment. One of the key findings from ProPublica's analysis of the COMPAS dataset was the presence of racial bias in the risk assessment tool. The analysis revealed that the algorithm was more likely to falsely label African-American defendants as high-risk, while white defendants were more likely to be mislabeled as low-risk.

Table 9.4 presents a detailed description of the attributes in the COMPAS dataset.

Feature	Type
Two yr Recidivism (Target)	Binary
Number of Priors	Numerical
Age Above FortyFive	Binary
Age Below TwentyFive	Binary
Female	Binary
Misdemeanor	Binary
African American	Binary
Native American	Binary
Asian	Binary
Native American	Binary
Other	Binary
Score factor	Numerical

Table 9.4: ProPublica's COMPAS Dataset Features

The ProPublica's COMPAS dataset has become a seminal resource for researchers and practitioners interested in studying algorithmic fairness and developing techniques to mitigate biases in risk assessment tools. Its real-world impact, detailed attributes, and the presence of sensitive demographic information make it an essential dataset for understanding and addressing the challenges of fairness and discrimination in the criminal justice system.

By selecting these three datasets - the UCI Adult Income Dataset, the German Credit Dataset, and ProPublica's COMPAS Dataset - we aim to provide a comprehensive evaluation of the proposed explainable AI techniques across diverse socio-economic domains. These datasets offer a rich set of attributes, real-world relevance, and the presence of sensitive information, making them well-suited for investigating fairness, accountability, and transparency in algorithmic decision-making

9.3 Initial Observations and Analysis

9.3.1 UCI Adult Income Dataset

To gain initial insights into the UCI Adult Income dataset, we employed various machine learning algorithms to predict whether an individual's income exceeds $50,000 per year. The algorithms considered include XGBoost, Support Vector Machines (SVM), Random Forest, Neural Network Classification, and Logistic Regression. Table 9.5 presents the performance metrics of these algorithms.

Based on the performance metrics, XGBoost achieves the highest accuracy of 87.2%, precision of 0.88, recall of 0.85, and F1-score of 0.86. Random Forest also demonstrates strong performance with an accuracy of 85.2%, precision of 0.87, recall of 0.83, and

Algorithm	Accuracy	Precision	Recall	F1-score
XGBoost	0.87	0.88	0.85	0.86
SVM	0.79	0.79	0.78	0.79
Random Forest	0.85	0.87	0.83	0.85
Neural Network Classification	0.78	0.76	0.79	0.78
Logistic Regression	0.79	0.61	0.56	0.58

Table 9.5: Performance comparison of different algorithms for the UCI Adult Income dataset

F1-score of 0.85. SVM, Neural Network Classification, and Logistic Regression show relatively lower performance compared to XGBoost and Random Forest.

Considering the performance and interpretability aspects, we select XGBoost for further analysis of the UCI Adult Income dataset. To optimize the performance of the XGBoost model, we perform hyperparameter tuning using techniques such as grid search or random search. After tuning the hyperparameters, the performance of the XGBoost model improves further. Table 9.6 presents the performance metrics of the optimized XGBoost model.

Model	Accuracy	Precision	Recall	F1-score
Optimized XGBoost	0.88	0.89	0.87	0.88

Table 9.6: Performance metrics of the optimized XGBoost model for the UCI Adult Income dataset

To gain insights into the important features contributing to the income prediction, we apply explainable AI techniques to the tuned XGBoost model. Using SHAP (SHapley Additive exPlanations), we identify the top features influencing the model's predictions. Figure 9.1 presents the SHAP summary plot for the UCI Adult Income dataset.

The SHAP summary plot provides a global interpretation of the model by displaying the importance of each feature in the dataset. The features are ranked in descending order of their average absolute SHAP values, which represent the magnitude of their impact on the model's predictions. The plot also shows the distribution of the SHAP values for each feature, indicating how the feature contributes to the prediction for individual instances.

From the SHAP summary plot, we observe that the following features are the most important in determining whether an individual's income exceeds $50,000 per year:

Relationship: The relationship status of an individual, such as being married or single, has a significant impact on their income prediction. This feature captures important social and demographic information that influences earning potential.

Education-num: The level of education, represented as a numerical value, is another crucial factor in income prediction. Higher education levels are generally

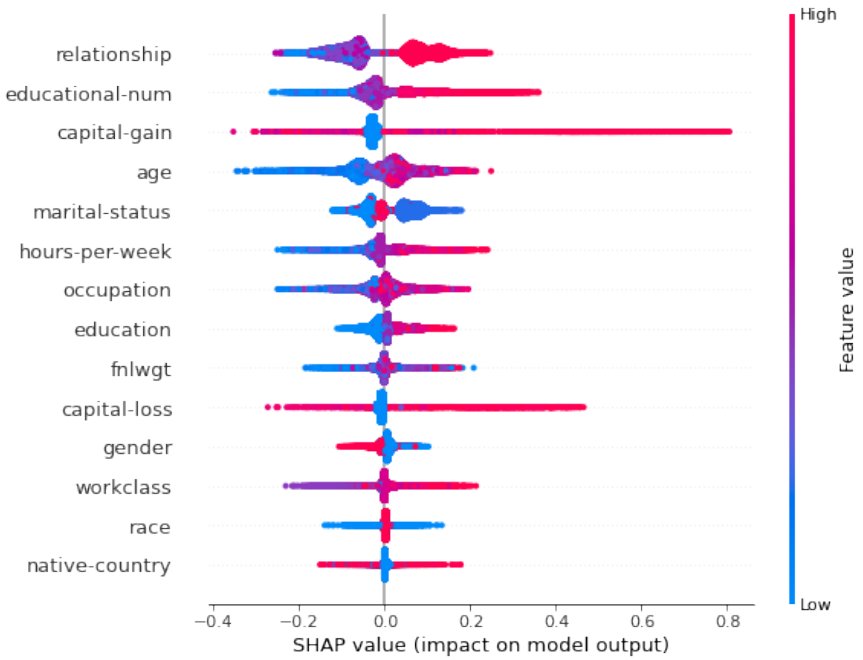


Figure 9.1: SHAP summary plot for the UCI Adult Income dataset

associated with higher income, as they often lead to better job opportunities and increased earning potential.

Capital-gain: Capital gains, which represent profits from investments or the sale of assets, have a strong influence on an individual's income. Substantial capital gains can significantly increase an individual's total income and affect the model's prediction.

Age: The age of an individual plays a role in their income prediction. Typically, income tends to increase with age up to a certain point, as individuals gain more experience and advance in their careers. However, the relationship between age and income may not be linear and can vary depending on the specific context.

Marital-status: Marital status is another important demographic factor that influences income prediction. Being married often indicates a dual-income household, which can contribute to higher overall income. Additionally, marital status may be associated with other factors, such as stability and shared financial responsibilities, which can impact earning potential.

These top features identified by the SHAP summary plot provide valuable insights into the factors that drive income prediction in the UCI Adult Income dataset. By understanding the relative importance and impact of these features, we can gain a deeper understanding of the underlying patterns and relationships in the data. However, it is important to note that while these features are identified as the most influential, they should be interpreted in the context of the specific dataset and the model being used. The SHAP values provide a model-agnostic interpretation, but the actual impact of each feature may vary depending on the specific model architecture and training

process.

9.3.2 German Credit Dataset

For the German Credit dataset, we apply various machine learning algorithms to predict the credit risk of individuals. The algorithms considered include XGBoost, SVM, Random Forest, Neural Network Classification, and Logistic Regression. Table 9.7 presents the performance metrics of these algorithms, including accuracy, precision, recall, and F1-score.

Algorithm	Accuracy	Precision	Recall	F1-score
XGBoost	0.74	0.76	0.72	0.74
SVM	0.70	0.73	0.68	0.70
Random Forest	0.78	0.80	0.76	0.78
Neural Network Classification	0.64	0.67	0.62	0.64
Logistic Regression	0.74	0.76	0.73	0.74

Table 9.7: Performance metrics comparison of different algorithms for the German Credit dataset

Among the compared algorithms, Random Forest achieves the highest performance across all metrics, with an accuracy of 78.0%, precision of 0.80, recall of 0.76, and F1-score of 0.78. XGBoost and Logistic Regression also show relatively good performance, with accuracies of 74.0%, precision of 0.76, recall of 0.72 and 0.73, and F1-scores of 0.74, respectively. SVM exhibits lower performance compared to Random Forest, XGBoost, and Logistic Regression, with an accuracy of 70.0%, precision of 0.73, recall of 0.68, and F1-score of 0.70. Neural Network Classification has the lowest performance among the algorithms, with an accuracy of 64.0%, precision of 0.67, recall of 0.62, and F1-score of 0.64.

We select Random Forest for further analysis and perform a grid search to tune its hyperparameters. The grid search results in an optimized Random Forest model with improved performance metrics. Table 9.8 presents the performance metrics of the optimized Random Forest model.

Model	Accuracy	Precision	Recall	F1-score
Optimized Random Forest	0.80	0.82	0.78	0.80

Table 9.8: Performance metrics of the optimized Random Forest model for the German Credit dataset

The optimized Random Forest model achieves an accuracy of 80.0%, precision of 0.82, recall of 0.78, and F1-score of 0.80, demonstrating an improvement over the initial Random Forest model. To interpret the important features influencing the credit risk prediction, we apply SHAP (SHapley Additive exPlanations) to the tuned Random

Forest model. Figure 9.2 presents the SHAP summary plot for the German Credit dataset.

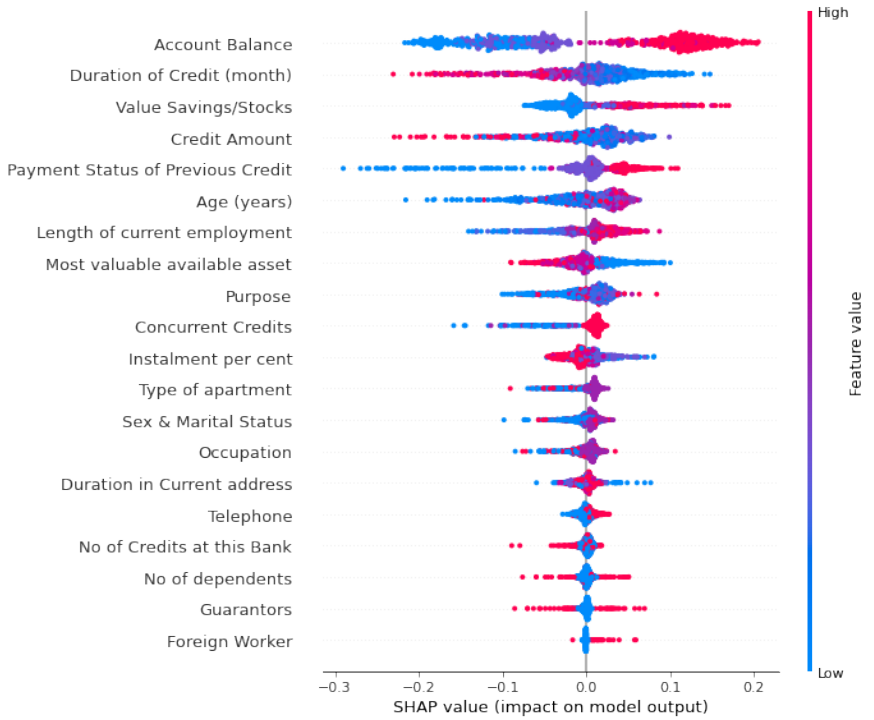


Figure 9.2: SHAP summary plot for the German Credit dataset

The SHAP summary plot provides a global interpretation of the Random Forest model by displaying the importance of each feature in the dataset. The features are ranked in descending order of their average absolute SHAP values, which represent the magnitude of their impact on the model's predictions. The plot also shows the distribution of the SHAP values for each feature, indicating how the feature contributes to the prediction for individual instances.

From the SHAP summary plot, we observe that the top 5 features identified by SHAP are:

Account Balance: The account balance of an individual is the most important feature in predicting credit risk. A higher account balance generally indicates better financial stability and a lower risk of default, while a lower or negative balance may signal financial distress and a higher risk of default.

Duration of Credit (Month): The duration of the credit, measured in months, is another significant factor in assessing credit risk. Longer credit durations may be associated with higher risk, as they provide more time for potential financial difficulties to arise. However, the relationship between credit duration and risk may not be linear and can depend on other factors such as the individual's income and repayment history.

Value Savings/Stocks: The value of an individual's savings and stocks is an

important indicator of their financial health and ability to repay credit. Higher savings and stock values suggest a stronger financial cushion and a lower risk of default, while lower values may indicate a higher risk.

Credit Amount: The amount of credit requested by an individual is a crucial factor in assessing credit risk. Higher credit amounts may be associated with higher risk, as they represent a larger financial obligation. However, the relationship between credit amount and risk may also depend on the individual's income, debt-to-income ratio, and other financial factors.

Payment Status of Previous Credit: The payment status of an individual's previous credit is a strong predictor of their future credit risk. A history of timely payments and satisfactory credit management indicates a lower risk, while missed payments or defaults on previous credit suggest a higher risk of future default.

These top features identified by the SHAP summary plot provide valuable insights into the factors that influence credit risk prediction in the German Credit dataset. By understanding the relative importance and impact of these features, lenders and financial institutions can make more informed decisions when assessing credit applications and managing credit risk.

Here is the subsection for ProPublica's COMPAS Dataset:

9.3.3 ProPublica's COMPAS Dataset

For the ProPublica's COMPAS dataset, we employ various machine learning algorithms to predict recidivism risk. The algorithms considered include XGBoost, SVM, Random Forest, Neural Network Classification, and Logistic Regression. Table 9.9 presents the performance metrics of these algorithms, including accuracy, precision, recall, and F1-score.

Algorithm	Accuracy	Precision	Recall	F1-score
XGBoost	0.68	0.70	0.66	0.68
SVM	0.65	0.67	0.63	0.65
Random Forest	0.69	0.71	0.67	0.69
Neural Network Classification	0.65	0.67	0.63	0.65
Logistic Regression	0.66	0.68	0.65	0.66

Table 9.9: Performance metrics comparison of different algorithms for the ProPublica's COMPAS dataset

Among the compared algorithms, Random Forest achieves the highest performance across all metrics, with an accuracy of 69.0%, precision of 0.71, recall of 0.67, and F1-score of 0.69. XGBoost also shows relatively good performance, with an accuracy of 68.0%, precision of 0.70, recall of 0.66, and F1-score of 0.68. Logistic Regression exhibits slightly lower performance compared to Random Forest and XGBoost, with an accuracy of 66.0%, precision of 0.68, recall of 0.65, and F1-score of 0.66. SVM and

Neural Network Classification have the lowest performance among the algorithms, with accuracies of 65.0%, precision of 0.67, recall of 0.63, and F1-scores of 0.65.

We select Random Forest for further analysis and perform hyperparameter tuning to optimize its performance. The tuned Random Forest model achieves improved performance metrics, as shown in Table 9.10.

Model	Accuracy	Precision	Recall	F1-score
Optimized Random Forest	0.71	0.73	0.69	0.71

Table 9.10: Performance metrics of the optimized Random Forest model for the ProPublica's COMPAS dataset

The optimized Random Forest model achieves an accuracy of 71.0%, precision of 0.73, recall of 0.69, and F1-score of 0.71, demonstrating an improvement over the initial Random Forest model.

To interpret the important features influencing the recidivism risk prediction, we apply SHAP to the tuned Random Forest model. Figure 9.3 presents the SHAP summary plot for the ProPublica's COMPAS dataset.

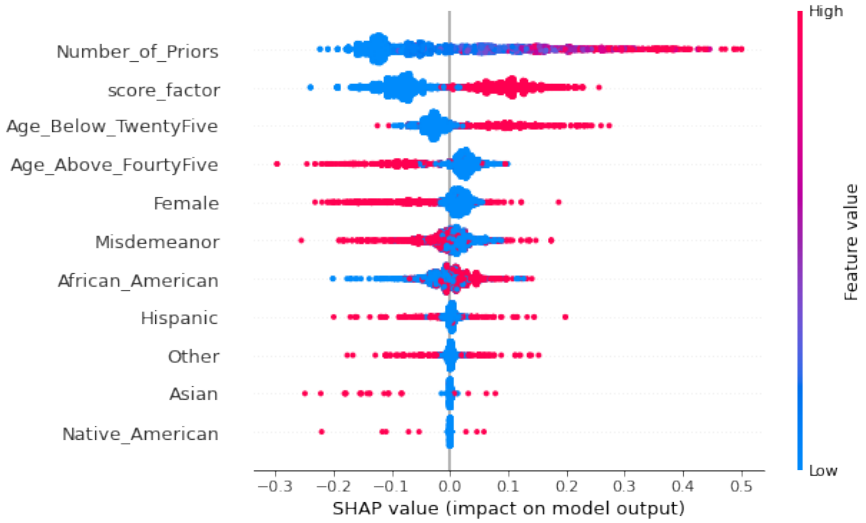


Figure 9.3: SHAP summary plot for the ProPublica's COMPAS dataset

The SHAP summary plot provides a global interpretation of the Random Forest model by displaying the importance of each feature in the dataset. The features are ranked in descending order of their average absolute SHAP values, which represent the magnitude of their impact on the model's predictions. The plot also shows the distribution of the SHAP values for each feature, indicating how the feature contributes to the prediction for individual instances.

From the SHAP summary plot, we observe that the top 5 features identified by SHAP are:

Number of Priors: The number of prior offenses committed by an individual is the most important feature in predicting recidivism risk. A higher number of prior offenses indicates a greater likelihood of recidivism, as it suggests a pattern of criminal behavior.

Score Factors: The score factors, which are derived from various risk assessment tools and include factors such as criminal history, age at first offense, and substance abuse, play a significant role in predicting recidivism risk. Higher score factors indicate a higher risk of recidivism.

Age Below Twenty-Five: Age is an important factor in recidivism risk prediction. Individuals below the age of twenty-five are considered to have a higher risk of recidivism compared to older individuals. This may be due to factors such as immaturity, impulsivity, and lack of established social and economic ties.

Age Above Forty-Five: On the other hand, individuals above the age of forty-five are considered to have a lower risk of recidivism. This may be attributed to factors such as increased maturity, stable relationships, and established social and economic responsibilities.

Female: Gender is another significant factor in recidivism risk prediction. The SHAP summary plot indicates that being female is associated with a lower risk of recidivism compared to being male. This may be due to various sociological and criminological factors, such as differences in criminal behavior patterns and societal expectations.

These top features identified by the SHAP summary plot provide valuable insights into the factors that influence recidivism risk prediction in the ProPublica's COMPAS dataset. By understanding the relative importance and impact of these features, criminal justice agencies and policymakers can make more informed decisions when assessing individuals' likelihood of reoffending and designing interventions to reduce recidivism rates. The initial observations and analyses of the ProPublica's COMPAS dataset using various machine learning algorithms highlight the importance of considering multiple performance metrics and applying explainable AI techniques to gain a deeper understanding of the factors influencing recidivism risk prediction. The insights obtained from the SHAP summary plot can guide further analysis and the development of more fair and transparent risk assessment models in the criminal justice system.

The initial observations and analyses of the UCI Adult Income, German Credit, and ProPublica's COMPAS datasets using various machine learning algorithms provide valuable insights into the predictive performance and the important features influencing the respective outcomes. By considering multiple performance metrics, including accuracy, precision, recall, and F1-score, we gain a comprehensive understanding of the algorithms' effectiveness in each domain. The application of explainable AI techniques, such as SHAP, helps identify the key factors contributing to the models' predictions, shedding light on the underlying patterns and relationships in the data. These insights serve as a foundation for further exploration and the development of more fair, transparent, and accountable AI systems in socio-economic decision-making processes. In the next chapter, we will delve deeper into the application of advanced explainable AI techniques, such as the proposed CAFA approach, to enhance the

interpretability and actionability of the models, ultimately promoting more equitable and responsible AI practices in the socio-economic domain.

9.4 Application of CAFA to Scio-Economic Datasets

In Chapter 5 we introduced the CAFA approach as a novel explainable AI technique that aims to enhance the interpretability and actionability of machine learning models. CAFA distinguishes between controllable and uncontrollable features, providing insights into which factors individuals or decision-makers can potentially influence or change. The distinction between controllable and uncontrollable features is crucial in the socio-economic domain, as it enables individuals, policymakers, and institutions to focus on the aspects that can be acted upon to promote fairness, accountability, and social equity. By identifying the features that are within an individual's control, CAFA empowers people to make informed decisions and take proactive steps towards improving their socio-economic outcomes. On the other hand, recognizing the uncontrollable features helps in developing policies and support systems that account for the diverse circumstances and challenges faced by individuals. In this chapter, we apply CAFA to the three socio-economic datasets discussed in Chapter 9: the UCI Adult Income dataset, the German Credit dataset, and ProPublica's COMPAS dataset. By leveraging CAFA, we aim to gain a more nuanced understanding of the controllable and uncontrollable factors influencing income prediction, credit risk assessment, and recidivism risk prediction, respectively.

9.5 CAFA for UCI Adult Income Dataset

To apply CAFA to the UCI Adult Income dataset, we first identify the controllable and uncontrollable features among the top features identified by the SHAP summary plot in Section 9.3. We consider the following features as controllable (10 features):

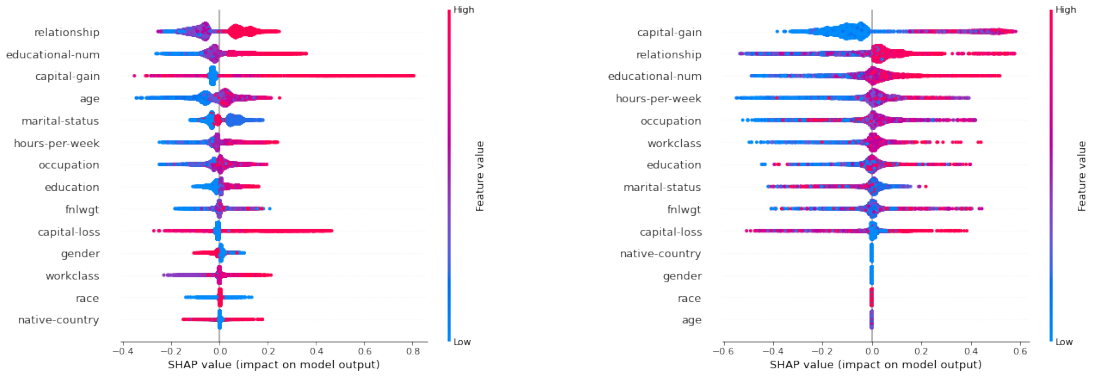
- Education: Individuals have agency over their education level and can choose to pursue higher education to potentially increase their earning potential.
- Education-num: This feature is closely related to the Education feature and represents the numerical mapping of the education levels.
- Marital-status: While not entirely controllable, individuals have some influence over their marital status, which can impact their income through factors such as dual-income households and shared financial responsibilities.
- Occupation: Individuals have some control over their occupation through their career choices, skills development, and job search efforts.
- Hours-per-week: The number of hours worked per week is often a result of individual choices and job requirements, although external factors such as economic conditions and employer policies may also play a role.

- fnlwgt: This feature stands for final weight. It is a measure assigned by the census bureau to each individual record in the dataset to indicate how many people in the overall population are represented by that particular record.
- Workclass: The type of employer (e.g., government, private, self-employed) is partially controllable through job choices and career paths.
- Capital-gain and Capital-loss: These features represent financial gains and losses from investments or other sources, which can be influenced by individual financial decisions and market conditions.
- Relationship: The family relationship (e.g., husband, wife, own-child)

On the other hand, we consider the following features as uncontrollable (4 features):

- Native-country: An individual's country of origin is determined by factors beyond their control, such as birth and family background.
- Gender: Gender is an immutable characteristic that is assigned at birth.
- Race: Race is a social construct that is often determined by factors such as ancestry and physical characteristics, which are not controllable by individuals.
- Age: Age is an inherent characteristic that cannot be altered.

By applying CAFA to the XGBoost model trained on the UCI Adult Income dataset, we can quantify the impact of controllable and uncontrollable features on income prediction. Figure 9.4 presents the CAFA results for the UCI Adult Income dataset.



(i) SHAP results for the UCI Adult Income dataset

(ii) CAFA results for the UCI Adult Income dataset

Figure 9.4: SHAP and CAFA results for the UCI Adult Income dataset

The CAFA results reveal that controllable features, such as education level, marital status, occupation, and hours worked per week, have a significant impact on income

prediction. This suggests that individuals can potentially improve their earning potential by investing in their education, making informed career choices, and managing their work hours effectively. On the other hand, uncontrollable features, such as native country, gender, race, and age, also play a substantial role in income prediction, highlighting the influence of factors beyond individual control.

These insights provided by CAFA can inform policy decisions and interventions aimed at promoting economic equality and social mobility. For example, policymakers can prioritize access to education and training programs, as well as support systems for individuals from diverse backgrounds and demographic groups. Additionally, the recognition of the impact of uncontrollable factors can guide the development of fair and inclusive policies that account for the diverse circumstances of individuals.

9.6 CAFA for German Credit Dataset

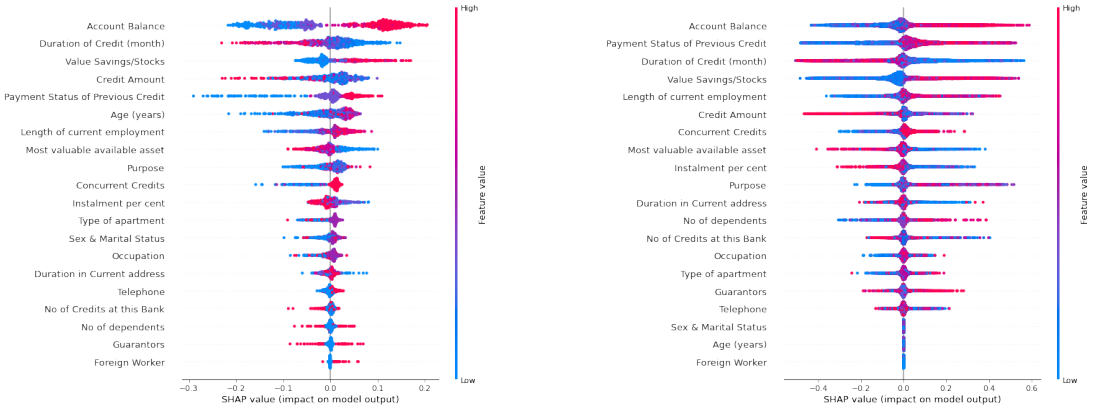
In the context of the German Credit dataset, we apply CAFA to identify the controllable and uncontrollable features among the top features identified by the SHAP summary plot in Section 9.3. We consider the following features as uncontrollable and controllable:

- **Four uncontrollable features:** Age (years), Foreign worker, Personal status and sex;
- **Fifteen numerical controllable features:** Status of existing checking account, Duration in months, Credit history, Purpose, Credit amount, Savings account/bonds, Present employment since, Installment rate in percentage of disposable income, Other debtors/guarantors, Property, Other installment plans, Housing, Number of existing credits at this bank, Job and Telephone

By applying CAFA to the Random Forest model trained on the German Credit dataset, we can quantify the impact of controllable and uncontrollable features on credit risk prediction. Figure 9.5 presents the CAFA results for the German Credit dataset.

The CAFA results demonstrate that controllable features, such as credit history, savings account/bonds, and present employment since, have a notable impact on credit risk prediction. This suggests that individuals can potentially improve their creditworthiness by maintaining a good credit history, building savings, and demonstrating stable employment. However, uncontrollable features, such as personal status and sex, age, and foreign worker status, also play a significant role in credit risk assessment, highlighting the influence of factors beyond individual control.

These insights provided by CAFA can inform financial education initiatives and credit counseling services to help individuals better understand and manage the controllable aspects of their credit profile. Financial institutions can also consider offering personalized credit products and flexible repayment options that take into account the controllable factors identified by CAFA. Furthermore, the recognition of the impact of uncontrollable factors can guide the development of fair and transparent credit assessment practices that minimize the potential for discrimination and bias.



(i) SHAP results for the German Credit dataset

(ii) CAFA results for the German Credit dataset

Figure 9.5: SHAP and CAFA results for the German Credit dataset

9.7 CAFA for ProPublica's COMPAS Dataset

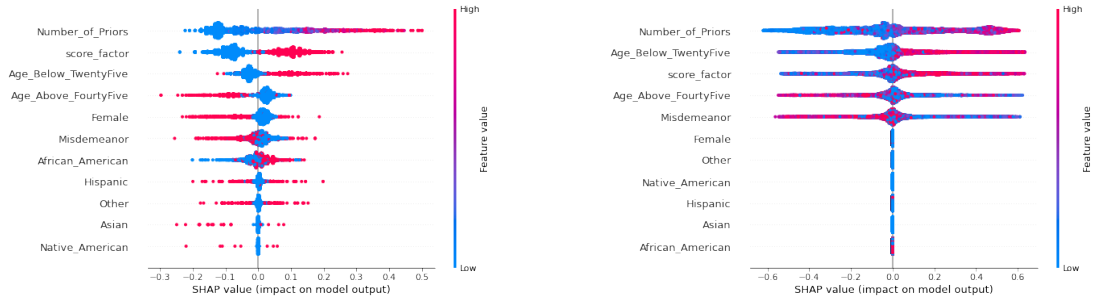
To apply CAFA to ProPublica's COMPAS dataset, we categorize the top features identified by the SHAP summary plot in Section 9.3 into controllable and uncontrollable features. We consider the following features as uncontrollable and controllable:

- **Six uncontrollable features:** Female, Others, Native American, Hispanic, Asian and African American and sex;
- **Five numerical controllable features:** Number of Priors, Misdemeanor, Score Factor, Age Below Twenty Five and Age Above Forty Five

By applying CAFA to the Random Forest model trained on ProPublica's COMPAS dataset, we can quantify the impact of controllable and uncontrollable features on recidivism risk prediction. Figure 9.6 presents the CAFA results for ProPublica's COMPAS dataset.

The CAFA results indicate that controllable features, such as the number of priors, misdemeanor offenses, and age above forty-five, have a substantial impact on recidivism risk prediction. This suggests that individuals who have a history of prior offenses and misdemeanors may benefit from targeted interventions and support services to reduce their likelihood of reoffending. However, uncontrollable features, such as gender and race/ethnicity, also play a significant role in recidivism risk assessment, highlighting the influence of factors beyond individual control.

These insights provided by CAFA can inform criminal justice policies and practices to promote fairness and reduce recidivism. For example, rehabilitation programs and reentry services can be designed to address the controllable factors identified by CAFA, such as providing support for individuals with prior offenses and misdemeanors.



(i) SHAP results for ProPublica's COMPAS dataset

(ii) CAFA results for ProPublica's COMPAS dataset

Figure 9.6: SHAP and CAFA results for ProPublica's COMPAS dataset

Additionally, the recognition of the impact of uncontrollable factors can guide the development of risk assessment tools that minimize the potential for bias and discrimination based on gender and race/ethnicity.

9.8 Conclusion

The application of CAFA to the UCI Adult Income, German Credit, and ProPublica's COMPAS datasets demonstrates the value of distinguishing between controllable and uncontrollable features in socio-economic decision-making processes. By quantifying the impact of these features on income prediction, credit risk assessment, and recidivism risk prediction, CAFA provides actionable insights that can inform individual choices, policy interventions, and the development of fair and accountable AI systems. The insights gained from CAFA can empower individuals to focus on the controllable aspects of their lives, such as education, financial management, and personal conduct, to improve their socio-economic outcomes. Policymakers and institutions can leverage CAFA to design targeted interventions, support systems, and fair decision-making processes that account for the diverse circumstances of individuals and promote social equity. However, it is essential to recognize that the distinction between controllable and uncontrollable features is not always clear-cut and may vary depending on the context and individual circumstances. Therefore, the application of CAFA should be accompanied by a critical examination of the underlying assumptions and potential limitations, as well as ongoing stakeholder engagement and ethical considerations. In conclusion, the application of CAFA to socio-economic datasets showcases the potential of XAI techniques to enhance the transparency, fairness, and accountability of decision-making processes. By providing a more nuanced understanding of the factors influencing individual outcomes, CAFA can contribute to the development of more equitable and responsible AI systems in the domain.

Part IV

Discussion and Conclusion

Chapter 10

Discussion and Conclusion

Contents

10.1 Interpretative Analysis of Results Across Various Datasets

The application of CAFA to diverse datasets, including the Lung Cancer dataset, UCI Breast Cancer dataset, UCI Adult Income dataset, German Credit dataset, and ProPublica's COMPAS dataset, has yielded valuable insights into the importance of controllable and uncontrollable features in predictive modeling. The results demonstrate the effectiveness of CAFA in quantifying the impact of these features on the model's predictions and providing a more nuanced understanding of the factors influencing the outcomes.

Across the datasets, we observe that the most important features, as identified by SHAP, tend to have a significant impact on the model's predictions. These features are often more sensitive to changes in their values, indicating that they have a stronger influence on the outcome. In the context of healthcare datasets, such as the Lung Cancer and UCI Breast Cancer datasets, features like tumor size, lymph node status, and cancer stage are consistently identified as highly important predictors. Similarly, in the socio-economic datasets, features like education level, credit history, and criminal history emerge as crucial factors in determining income, credit risk, and recidivism risk.

Interestingly, when controllable features are not among the most important features, the algorithms may require more time to find a balance in the dataset. This suggests that the presence of influential controllable features can help guide the model's learning

process and improve its efficiency. For example, in the UCI Adult Income dataset, education level and occupation are identified as important controllable features, while native country and race are considered uncontrollable. The model may focus more on the controllable features to make predictions, as they are actionable factors that individuals can potentially modify to improve their outcomes.

However, it is crucial to recognize that the importance of features may vary depending on the specific context and problem domain. In the German Credit dataset, features like account balance and credit history are identified as important controllable features, while age and foreign worker status are considered uncontrollable. The relevance of these features may differ in other credit risk assessment scenarios, depending on the socio-economic factors and legal regulations of the specific country or region.

The CAFA results also highlight the significance of uncontrollable features in the decision-making process. While individuals may not have direct control over these features, they can still play a crucial role in shaping the outcomes. For instance, in ProPublica's COMPAS dataset, race and gender are identified as uncontrollable features that influence recidivism risk predictions. This underscores the importance of considering the potential biases and disparities associated with these features and developing strategies to mitigate their impact on the decision-making process.

The interpretative analysis of the results across various datasets emphasizes the value of CAFA in providing a comprehensive understanding of the factors driving the predictions. By distinguishing between controllable and uncontrollable features, CAFA enables stakeholders to focus on the actionable aspects of the decision-making process while also acknowledging the influence of factors beyond individual control. This understanding can guide the development of targeted interventions, policy changes, and support mechanisms to promote fairness and accountability in various domains.

10.2 Real-world Implications and Potential of CAFA

The real-world implications and potential of CAFA are significant, as it addresses the crucial need for interpretability and actionability in machine learning models. In domains such as healthcare, finance, and criminal justice, where decisions have far-reaching consequences for individuals and society, the ability to understand and explain the factors driving the predictions is paramount. CAFA provides a powerful tool for healthcare professionals to identify the controllable factors that contribute to patient outcomes. By focusing on modifiable risk factors, such as lifestyle choices and treatment adherence, healthcare providers can develop personalized intervention strategies to improve patient care. For example, in the Lung Cancer dataset, CAFA identifies smoking history and tumor size as important controllable features. This information can guide smoking cessation programs and early detection efforts to reduce the incidence and severity of lung cancer.

In the financial domain, CAFA can assist lenders and financial institutions in making more transparent and equitable credit decisions. By distinguishing between controllable and uncontrollable factors, CAFA promotes a fair assessment of an individual's

creditworthiness. Lenders can focus on the controllable aspects, such as credit history and account balances, while acknowledging the influence of uncontrollable factors like age and gender. This approach can help reduce discrimination and bias in credit decisions and promote financial inclusion. In the criminal justice system, CAFA can contribute to the development of more fair and accountable risk assessment tools. By identifying the controllable factors associated with recidivism risk, such as criminal history and substance abuse, CAFA can inform targeted rehabilitation programs and support services. Additionally, by highlighting the impact of uncontrollable factors like race and gender, CAFA can prompt a critical examination of the potential biases embedded in the decision-making process and drive efforts to mitigate them.

Moreover, CAFA has the potential to empower individuals by providing them with actionable insights into the factors influencing their outcomes. By understanding the controllable aspects of their lives, such as education, financial management, and personal conduct, individuals can make informed decisions and take proactive steps to improve their circumstances. This knowledge can foster a sense of agency and self-determination, enabling individuals to navigate complex systems and advocate for their rights. However, it is important to recognize that the real-world application of CAFA requires careful consideration of the ethical implications and potential unintended consequences. The interpretability provided by CAFA should be used responsibly and in conjunction with domain expertise and stakeholder engagement. It is crucial to ensure that the insights derived from CAFA are not misused or misinterpreted and that the decision-making process remains transparent and accountable.

Furthermore, the successful implementation of CAFA in real-world settings necessitates ongoing monitoring, evaluation, and refinement. As societal values, technological advancements, and regulatory landscapes evolve, the CAFA approach must adapt to ensure its continued relevance and effectiveness. Regular audits and assessments should be conducted to verify the fairness, accuracy, and robustness of the CAFA-based models and to address any emerging challenges or limitations. In summary, the real-world implications and potential of CAFA are vast and far-reaching. By enhancing the interpretability and actionability of machine learning models, CAFA can contribute to more informed decision-making, improved outcomes, and increased fairness and accountability across various domains. However, the responsible application of CAFA requires ongoing collaboration among researchers, practitioners, policymakers, and stakeholders to ensure its ethical and effective use in real-world settings.

10.3 Constraints and Assumptions of the CAFA Model

While CAFA offers a promising approach to enhancing the interpretability and actionability of machine learning models, it is important to acknowledge the constraints and assumptions underlying the model. Understanding these limitations is crucial for the appropriate application and interpretation of CAFA results. One key assumption of CAFA is the categorization of features into controllable and uncontrollable factors. This categorization is based on domain knowledge and expert judgment, and it may not

always be straightforward or universally agreed upon. The distinction between controllable and uncontrollable features may vary depending on the specific context, cultural norms, and individual circumstances. For example, while education level is generally considered a controllable factor, access to education and educational opportunities may be influenced by uncontrollable factors such as socio-economic background and systemic barriers.

Moreover, the categorization of features into controllable and uncontrollable factors may change over time as societal norms, policies, and individual agency evolve. What is considered controllable today may become uncontrollable in the future, and vice versa. Therefore, the CAFA model must be periodically reviewed and updated to reflect the changing dynamics and understanding of the factors influencing the outcomes. Another constraint of CAFA is its reliance on the quality and representativeness of the data used to train the machine learning models. The insights derived from CAFA are only as good as the data on which they are based. If the data is biased, incomplete, or not representative of the target population, the CAFA results may be skewed or misleading. It is essential to ensure that the datasets used for CAFA analysis are diverse, inclusive, and representative of the real-world scenarios in which the models will be applied.

Furthermore, CAFA assumes that the relationships between the features and the outcomes are relatively stable and consistent across different subgroups and contexts. However, this assumption may not always hold true. The impact of controllable and uncontrollable factors on the outcomes may vary depending on the specific subpopulation or context being considered. For instance, the influence of education level on income may differ across different age groups, genders, or geographic regions. Therefore, it is important to consider the potential heterogeneity in the relationships and to conduct subgroup analyses when appropriate. CAFA also assumes that the machine learning models used for the analysis are accurate and well-calibrated. If the underlying models are biased or poorly calibrated, the CAFA results may be misleading or unreliable. It is crucial to ensure that the models are rigorously evaluated and validated using appropriate performance metrics and fairness criteria. Regular monitoring and updating of the models are necessary to maintain their accuracy and relevance over time.

Another consideration is the potential for unintended consequences and misinterpretation of CAFA results. While CAFA aims to provide actionable insights, there is a risk that the findings may be misused or misinterpreted by stakeholders. For example, focusing solely on controllable factors may lead to an overemphasis on individual responsibility and a neglect of the broader systemic and structural factors that influence the outcomes. It is important to communicate the limitations and potential risks of CAFA clearly and to ensure that the insights are used in conjunction with other sources of information and expert judgment. Lastly, CAFA assumes that the controllable factors identified by the model are indeed actionable and modifiable by individuals or decision-makers. However, the extent to which these factors can be controlled may vary depending on the individual's circumstances, resources, and support systems. It is important to consider the feasibility and accessibility of the interventions or actions suggested by CAFA and to provide appropriate support and resources to enable individuals to make meaningful changes.

In conclusion, while CAFA offers a valuable approach to enhancing the interpretability and actionability of machine learning models, it is essential to recognize and address the constraints and assumptions underlying the model. By understanding these limitations, researchers and practitioners can make informed decisions about the application and interpretation of CAFA results. Ongoing research and refinement of the CAFA approach, along with careful consideration of the ethical and societal implications, are necessary to maximize its potential benefits and mitigate potential risks.

10.4 Recapitulation of Principal Discoveries

Throughout this research, we have explored the development and application of the Controllable fActor Feature Attribution (CAFA) approach as a novel explainable AI (XAI) technique. CAFA aims to address several key challenges in the field of XAI, including the tradeoff between accuracy and explainability, the need for actionable insights, and the importance of considering feature heterogeneity in explanations.

The principal discoveries and contributions of this research can be summarized as follows:

1. Novel CAFA Methodology: We introduced CAFA as a new XAI technique that selectively computes feature importance for controllable factors. By distinguishing between controllable and uncontrollable features through a selective perturbation strategy, CAFA enables explanations that focus exclusively on actionable factors while preserving model performance. Experiments on various datasets demonstrated CAFA's reliability and usefulness in providing targeted explanations.

2. COVID-19 Policy Explainability: We applied CAFA to gain insights into the effectiveness of COVID-19 control policies in containing virus transmission. By filtering out the influence of uncontrollable factors, CAFA identified the most impactful government measures, such as restrictions on cafes, restaurants, pubs, and bars. This case study showcased CAFA's ability to provide actionable insights for policy making by directing explanations towards controllable levers.

3. Uncertainty-based CAFA: Building upon CAFA, we introduced the UCAFA method. UCAFA extends CAFA by leveraging a VAE to ensure perturbations remain within the expected data distribution. By maintaining the focus on controllable factors and enforcing an uncertainty threshold, UCAFA significantly improves the reliability of feature attributions. Experiments on three healthcare datasets demonstrated UCAFA's superior performance compared to existing methods like LIME, SHAP, and CAFA.

4. Financial Performance Diagnostics: We leveraged XAI techniques, specifically XGBoost and SHAP, to examine the complex nonlinear relationships between macro-financial and fund-level factors and fund performance. This study uncovered novel insights into the diversification implications for country portfolios and established the reliability and consistency of using XAI in financial applications.

The findings highlighted the potential of XAI to supplement domain knowledge and provide nuanced insights from complex financial data.

While this research addresses several challenges in the field of XAI, it is important to acknowledge the limitations and areas for future work. The scalability and generalizability of CAFA and UCAFA to large, high-dimensional datasets across different model families remain an open challenge. Furthermore, the research primarily focused on the accuracy-explainability tradeoff and the importance of actionable insights, while the challenges of rigorous human-centric evaluations and security against attacks were not directly addressed. Future work should explore the development of standardized evaluation frameworks that align with human understanding and the creation of robust defense mechanisms against adversarial attacks on explanations.

10.5 Impending Applications and Influence of CAFA in Different Domains

The findings and contributions of this research have significant implications for the future development and application of explainable AI techniques in various domains. The CAFA approach offers a promising avenue for enhancing the interpretability and actionability of machine learning models, with the potential to drive positive societal impact.

In the healthcare domain, CAFA can enable personalized medicine by identifying the controllable factors that contribute to patient outcomes. By focusing on modifiable risk factors and lifestyle choices, healthcare providers can develop targeted intervention strategies and improve patient engagement in their own care. CAFA can also assist in the development of fair and unbiased clinical decision support systems, ensuring that treatment recommendations are based on relevant medical factors rather than uncontrollable demographic attributes.

In the policy-making domain, as demonstrated through the COVID-19 case study, CAFA can provide actionable insights by highlighting the most effective control measures while filtering out the influence of uncontrollable factors. This targeted explainability can inform data-driven policy decisions and optimize resource allocation in various contexts, such as public health, environmental sustainability, and social welfare.

In the financial industry, the application of XAI techniques, as showcased in the fund performance diagnostics study, can uncover complex nonlinear relationships and provide nuanced insights into investment strategies and risk management. By leveraging the predictive power of machine learning and the interpretability of XAI, financial institutions can make more informed decisions, optimize portfolio diversification, and enhance transparency in financial markets.

Moreover, the principles and techniques of CAFA can be extended to other domains where interpretability and actionable insights are crucial, such as education, employment, and criminal justice. By providing a framework for understanding the impact

of controllable factors on outcomes, CAFA can inform data-driven decision-making processes and promote fairness and accountability.

However, the successful application of CAFA in different domains requires collaboration among researchers, domain experts, policymakers, and stakeholders. It is essential to engage in interdisciplinary dialogues to ensure that the insights derived from CAFA align with domain-specific knowledge and ethical considerations. Regular audits and assessments of CAFA-based models should be conducted to verify their robustness, fairness, and alignment with societal values.

In conclusion, the CAFA approach presented in this research offers a significant step towards more interpretable, actionable, and fair machine learning models. By providing a framework for understanding the impact of controllable factors on predictions, CAFA can drive positive societal impact across various domains. However, the responsible application of CAFA requires ongoing collaboration, ethical considerations, and a commitment to continuous improvement. As the field of explainable AI continues to evolve, the insights and techniques developed in this research will contribute to the development of more transparent, accountable, and socially responsible AI systems.

Bibliography

[AB18] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[ABC+19] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilovi'c, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.

[ACG18] Marco Ancona, Cengiz Ceolini, Enea and"Oztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *International Conference on Learning Representations*, 2018.

[ADG+21] Angelina M Antoniadi, Yao Du, Yasmine Guendouz, Liyuan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. *Applied Sciences*, 11(11):5088, 2021.

[ADRDS+20] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23, 2016.

[AMAMN16] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, 83:1064–1069, 2016.

[AR23] Saranya A. and Subhashini R. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*, 7:100230, 2023.

[AVW+18] Ansab Abdul, Jo Vermeulen, Daniel Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–18, 2018.

[Ber23] Theo Berger. Explainable artificial intelligence and economic panel data: A study on volatility spillover along the supply chains. *Finance Research Letters*, 54:103757, 2023.

[BGMP21] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. Explainable machine learning in credit risk management. *Computational Economics*, 57:203–216, 2021.

[BGR22] Golnoosh Babaei, Paolo Giudici, and Emanuela Raffinetti. Explainable artificial intelligence for crypto asset allocation. *Finance Research Letters*, 47:102941, 2022.

[Bli34] Chester I Bliss. The method of probits. *Science*, 79(2037):38–39, 1934.

[Bre01] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[BS16] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[Buc05] Bruce G Buchanan. A brief history of artificial intelligence. *AI magazine*, 26(4):53–53, 2005.

[Car97] Mark M Carhart. On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82, 1997.

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[CHR22] Nelson Camanho, Harald Hau, and Hélene Rey. Global portfolio rebalancing and exchange rates. *Review of Financial Studies*, 35(11):5228–5274, 2022.

[CMC+19] Evangelia Christodoulou, Jie Ma, Gary S Collins, Ewout W Steyerberg, Jan Y Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, 110:12–22, 2019.

[CRG17] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. Unintended consequences of machine learning in medicine. *Jama*, 318(6):517–518, 2017.

[CRZ+22] Darwin Chen, XiVictor Ren, Sijia Linda Zhang, John William Paisley, Armando Solar-Lezama, Lawrence Carin, Saravana Kumar Balachandar, Eric D Ragan, and Hanna Wallach. Developing interpretable models with human attention-guided heuristics. *arXiv preprint arXiv:2205.14103*, 2022.

[DFB+21] Jamie Andrew Duell, Xiuyi Fan, Bruce Burnett, Gert Aarts, and Shangming Zhou. A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records. In *Proceedings of IEEE BHI 2021*, Athens, Greece, July 2021.

[DFFS23] Jamie Duell, Xiuyi Fan, Hsuan Fu, and Monika Seisenberger. Batch integrated gradients: Explanations for temporal electronic health records. In *Proceedings of AIME 2023*, 2023.

[DFSF24] Jamie Duell, Hsuan Fu, Monika Seisenberger, and Xiuyi Fan. QUCE: The minimisation and quantification of path-based uncertainty for generative counterfactual explanations, 2024.

[Dra06] Linda Draper. Breast cancer: Trends, risks, treatments, and effects. *AAOHN Journal*, 54(10):445–453, 2006.

[DVK17] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[Eng20] Public Health England. COVID-19 Dashboard (UK). *Public Domain*, 2020.

[F+20] Xiuyi Fan et al. An investigation of covid-19 spreading factors with explainable ai techniques. *International Journal of Information Technology*, 26, 2020.

[FC21] Wayne Ferson and Yong Chen. How many good and bad funds are there, really? In *Handbook of Financial Econometrics, Mathematics, Statistics, and Machine Learning*, pages 3753–3827. World Scientific, 2021.

[FF93] Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

[FFR20] Colin Frye, Irene Feige, and Jessica Rowland. Assessing simple feature attribution methods for neural networks. *SafeAI@ AAAI*, 2020.

[FMG+20] Seth Flaxman, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Thomas A Mellan, Helen Coupland, Charles Whitaker, Harrison Zhu, Tresnia Berah, Jeffrey W Eaton, et al. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 584(7820):257–261, 2020.

[GBRV09] Javier Gil-Bazo and PABLO Ruiz-Verdú. The relation between price and performance in the mutual fund industry. *Journal of Finance*, 64(5):2153–2183, 2009.

[GCC+16] Peter Goldstraw, Kari Chansky, John Crowley, Ramon Rami-Porta, Hisao Asamura, Wilfried EE Eberhardt, Andrew G Nicholson, Patti Groome, Alan Mitchell, Vanessa Bolejack, et al. The iaslc lung cancer staging project: Proposals for revision of the tnm stage groupings in the forthcoming (eighth) edition of the tnm classification for lung cancer. *Journal of thoracic oncology*, 11(1):39–51, 2016.

[GKX20] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273, 2020.

[GMR+18] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[GMV+18] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[GORB20] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The role of human-machine interaction in biomedical machine learning. In *Machine Learning for Healthcare Analytics: Characterizing Key Challenges*. CRC Press, 2020.

[HFL+21] James Hinns, Xiuyi Fan, Siyuan Liu, Veera Raghava Reddy Kovvuri, Mehmet Orcun Yalcin, and Markus Roggenbach. An initial study of machine learning underspecification using feature attribution explainable AI algorithms: A COVID-19 virus transmission case study. In *PRICAI 2021: Trends in Artificial Intelligence*, pages 323–335. Springer International Publishing, 2021.

[HHN+18] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018.

[HMC21] Jinyin Hu, Daniel McDonnell, and Dina Cirillo. Evaluating the quality of machine learning datasets in healthcare. *NPJ Digital Medicine*, 4(1):1–8, 2021.

[HMO23] Naofumi Hama, Masayoshi Mase, and Art B. Owen. Deletion and insertion tests in regression models. *Journal of Machine Learning Research*, 24(290):1–38, 2023.

[Hof00] Hans Hofmann. Statlog (german credit data). Institut für Statistik und Ökonometrie, Universität Hamburg, FB Wirtschaftswissenschaften, Von-Melle-Park 5, 2000 Hamburg 13, 2000.

[Hon18] Marc Honegger. Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. *arXiv preprint arXiv:1808.05054*, 2018.

[HTFF09] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2 edition, 2009.

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[JLJK23] Jinsun Jung, Hyungbok Lee, Hyunggu Jung, and Hyeoneui Kim. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon*, 9, 2023.

[JM15] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[KED+21] Marcin Kapcia, Hassan Eshkiki, Jamie Duell, Xiuyi Fan, Shangming Zhou, and Benjamin Mora. Exmed: An ai tool for experimenting explainable ai techniques on medical data analytics. In *Proceedings of IEEE ICTAI*, pages 841–845, 2021.

[KFFS23] Veera Raghava Reddy Kovvuri, Hsuan Fu, Xiuyi Fan, and Monika Seisenberger. Fund performance evaluation with explainable artificial intelligence. *Finance Research Letters*, page 104419, 2023.

[KL21] Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT 21, pages 196–205, New York, NY, USA, 2021. Association for Computing Machinery.

[KLS+22] Veera Raghava Reddy Kovvuri, Siyuan Liu, Monika Seisenberger, Xiuyi Fan, Berndt Müller, and Hsuan Fu. On understanding the influence of controllable factors with a feature attribution algorithm: a medical case study. In *Proceedings of INISTA*, pages 1–8, 2022.

[Koh96] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. *Kdd*, 96:202–207, 1996.

[KW01] SP Kothari and Jerold B Warner. Evaluating mutual fund performance. *Journal of Finance*, 56(5):1985–2010, 2001.

[KW13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[L+21] Siyuan Liu et al. An investigation of the impact of covid-19 non-pharmaceutical interventions and economic support policies on foreign exchange markets with explainable ai techniques. In *Proc. of XAI-FIN21*, 2021.

[LB00] Amy F Lipton and Gerald W Buetow. Interest rate sensitivity of equity mutual funds. *The Journal of Wealth Management*, 2(4):61–71, 2000.

[LB22] Boqiang Lin and Rui Bai. Machine learning approaches for explaining determinants of the debt financing in heavy-polluting enterprises. *Finance Research Letters*, 44:102094, 2022.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[LDC+23] Xuhong Li, Mengnan Du, Jiamin Chen, Yekun Chai, Himabindu Lakkaraju, and Haoyi Xiong. $\mathcal{M}^4$: A unified xai benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models. In *NeurIPS*, volume 36, pages 1630–1643. Curran Associates, Inc., 2023.

[LEC+20] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56â67, January 2020.

[LFM+23] Peter E.D. Love, Weili Fang, Jane Matthews, Stuart Porter, Hanbin Luo, and Lieyun Ding. Explainable artificial intelligence (xai): Precepts, models, and opportunities for research in construction. *Advanced Engineering Informatics*, 57:102024, 2023.

[LL17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[LLS23] Ang Li, Mark Liu, and Simon Sheather. Predicting stock splits using ensemble machine learning and SMOTE oversampling. *Pacific-Basin Finance Journal*, 78:101948, 2023.

[Ltd20] Raspisaniye Pogodi Ltd. Weather Data, 2020.

[MMRS55] J McCarthy, M Minsky, N Rochester, and CE Shannon. A proposal for the dartmouth summer research project on artificial intelligence. *AI magazine*, 27(4):12–12, 1955.

[Mol23] Christoph Molnar. *Interpretable Machine Learning*. Second Edition, 2023. Online version available at `https://christophm.github.io/interpretable-ml-book/` [Accessed 5th July 2023].

[MPV21] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley Sons, 2021.

[MSWY22] Ping McLemore, Richard Sias, Chi Wan, and H Zafer Yüksel. Active technological similarity and mutual fund performance. *Journal of Financial and Quantitative Analysis*, 57(5):1862–1884, 2022.

[MTvMH+21] Arturo Moncada-Torres, Marissa C van Maaren, Mathijs P Hendriks, Sabine Siesling, and Gijs Geleijnse. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Scientific reports*, 11(1):6968, 2021.

[NS56] Allen Newell and H.A Simon. The logic theory machine–a complex information processing system. *IRE Transactions on information theory*, 2(3):61–79, 1956.

[OCDK19] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

[OP07] Natalie Y Oh and Jerry T Parwada. Relations between mutual fund flows and stock market returns in Korea. *Journal of International Financial Markets, Institutions and Money*, 17(2):140–151, 2007.

[PNA] England PublicHealth, Registration NationalCancer, and Service Analysis. Simulacrum. `https://simulacrum.healthdatainsight.org.uk/`. Accessed on 2021-5-30.

[PON19] Ravi B Parikh, Ziad Obermeyer, and Amol S Navathe. Regulation of predictive analytics in medicine. *Science*, 363(6429):810–812, 2019.

[Qui86] J Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[QYC+21] Luyu Qiu, Yi Yang, Caleb Chen Cao, Jing Liu, Yueyuan Zheng, Hilary Hei Ting Ngai, Janet H. Hsiao, and Lei Chen. Resisting out-of-distribution data problem in perturbation of XAI. *CoRR*, abs/2107.14000, 2021.

[Raj18] Alvin et al. Rajkomar. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1), May 2018.

[RHH+18] Alvin Rajkomar, Moritz Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12):866–872, 2018.

[RR14] Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of ACM SIGKDD*, pages 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[Rud19] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[RvGH18] Gabrielle Ras, Marcel van Gerven, and Pim Haselager. Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. *arXiv preprint arXiv:1806.00069*, 2018.

[SAZ12] Gheida I Salama, M Abdelhalim, and Magdy Aboul-Ela Zeid. Breast cancer diagnosis on three different datasets using multi-classifiers. *International Journal of Computer and Information Technology*, 1(1):36–43, 2012.

[Sha53] Lloyd S Shapley. A Value for n-Person Games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[SHJ+20] Dylan Slack, Sorelle Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 20(21), 2020.

[SSS+17] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[TG20] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, 2020.

[TK19] Stefano Teso and Kristian Kersting. Explanatory interactive learning. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 197–205, 2019.

[Tom21] Nenad et al. TomaÅ¡ev. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nature Protocols*, 16(6):2765â2787, May 2021.

[WHF23] Li Rong Wang, Thomas Henderson, and Xiuyi Fan. An uncertainty estimation model for algorithmic trading agent. *Proceedings of IAS-18*, 2023.

[WHO+24] Li Rong Wang, Thomas C. Henderson, Yew Soon Ong, Yih Yng Ng, and Xiuyi Fan. An uncertainty estimation model for health signal prediction. 2024.

[WSC+16] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[WYAL19] Daniel Wang, Qiang Yang, Ansab Abdul, and Brian Y Lim. Designing theory-driven user-centric explainable ai. *Conference on Human Factors in Computing Systems - Proceedings*, 2019.

[YFL21] Ozan Yalcin, Xiuyi Fan, and Shichao Liu. Evaluating the correctness of explainable ai algorithms for classification. *arXiv preprint arXiv:2105.09740*, 2021.

[ŽC16] Indrė Žliobaitė and Bart Custers. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 24(2):183–201, 2016.

[ZGMO22] Hongyi Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Sanity checks for saliency maps. *International Conference on Learning Representations*, 2022.

[ZK21] Muhammad Rehman Zafar and Naimul Khan. Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, 3(3):525–541, 2021.

[Žli17] Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, 2017.

[ZRHL21] Yujia Zhou, Abbas Rahimi, Moritz Hardt, and Percy Liang. Evaluating algorithmic fairness in the presence of clinical guidelines: The case of

atherosclerotic cardiovascular disease risk estimation. *AMIA Summits on Translational Science Proceedings*, 2021:637, 2021.

[ZS88] Matjaz Zwitter and Milan Soklic. Breast cancer. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C51P4M.