UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Science
School of Electronics and Computer Science

Model Monitoring in the Absence of Labeled Data via Feature Attributions Distributions

DOI: 10.1002/0470841559.ch1

by

Carlos Mougan Navarro

A thesis for the degree of
Doctor of Philosophy

February 2025

University of Southampton

Abstract

Faculty of Engineering and Physical Science
School of Electronics and Computer Science

Doctor of Philosophy

Model Monitoring in the Absence of Labeled Data via Feature Attributions Distributions

by Carlos Mougan Navarro

Model monitoring involves analyzing AI algorithms once they have been deployed and detecting changes in their behaviour. This thesis explores machine learning model monitoring ML before the predictions impact real-world decisions or users. This step is characterized by one particular condition: the absence of labelled data at test time, which makes it challenging, even often impossible, to calculate performance metrics.

The thesis is structured around two main themes: *(i) AI alignment*, measuring if AI models behave in a manner consistent with human values and *(ii) performance monitoring*, measuring if the models achieve specific accuracy goals or desires.

The thesis uses a common methodology that unifies all its sections. It explores feature attribution distributions for both monitoring dimensions. Using these feature attribution explanations, we can exploit their theoretical properties to derive and establish certain guarantees and insights into model monitoring.

For AI Alignment, we explore whether the distributions of feature attributions are distinct for different social groups and propose a new formalization of equal treatment. This novel metric assesses how well AI decisions adhere to ethical standards and political-philosophical values. Our notion of Equal Treatment tests for statistical independence of the explanation distributions over populations with different protected characteristics. We show the theoretical properties of our formalization of equal treatment and devise an equal treatment inspector based on the AUC of a classifier two-sample test.

For performance monitoring, we define *explanation shift* as the statistical comparison between how predictions from training data are explained and how predictions on new data are explained. We propose explanation shift as a key indicator to investigate the interaction between distribution shifts and learned models. We introduce an Explanation Shift Detector that operates on the explanation distributions, providing

more sensitive and explainable changes in interactions between distribution shifts and learned models. We compare explanation shifts with other methods that are based on distribution shifts, showing that monitoring for explanation shifts results in more sensitive indicators for varying model behavior. We provide theoretical and experimental evidence and demonstrate the effectiveness of our approach on synthetic and real data.

Finally, to explain model degradation we use a second model, that predicts the uncertainty estimates of the first

Additionally, we release two open-source Python packages, `skshift` and `explanationspace`, which implement our methods and provide usage tutorials for further reproducibility.

Contents

LIST OF PUBLICATIONS xiii

Declaration of Authorship **xvi**

1 Introduction to Model Monitoring **1**
1.1 AI Alignment and Feature Attributions Challenges 4
1.2 Performance Monitoring with Feature Attributions Distributions 5
1.2.1 Reproducibility and Software Contribution Objectives 6
1.2.2 Summary Requirements for Model Monitoring in the Absence of Labeled Data via Feature Attributions Distributions 7
1.3 Overview of the Thesis 8

2 Foundations **11**
2.1 Mathematical Notation 11
2.2 Feature Attribution Explainations 12
2.2.1 Shapley Values 12
2.2.1.1 Efficiency. 13
2.2.1.2 Uninformativeness. 13
Linear Models and IID Data 14
2.2.2 LIME: Local Interpretable Model-Agnostic Explanations 14
2.3 AI Alignment 15
2.3.1 Distributive Justice Political Philosophy Notions 17
2.3.1.1 Egalitarianism and Equal Opportunity 17
2.3.1.2 Socialism and Equal Outcome 18
2.3.1.3 Liberalism and Equal Treatment 19
2.3.2 Moral Philosophy Notions 20
2.3.2.1 Utilitarianism 20
The Principle of Utility. 20
Consequentialist Evaluation. 21
2.3.2.2 Deontology 21
2.3.2.3 Kantian Critique of Utilitarianism 22
Intrinsic value and the pursuit of the 'good'. 22
The issue of aggregation. 22
Optimizing outcomes. 23
2.4 Technical Types of Fairness 24
2.4.1 Disparate Impact Fairness Metrics: Approaches that Rely on Labeled Data 24
2.4.1.1 Equal Opportunity 24

2.4.1.2 Treatment Equality . . . 25
2.4.2 Disparate Treatment Fairness Metrics: Approaches that Determine Unfairness without Labelled Data . . . 25
2.4.2.1 Fairness Through Unawareness. . . . 25
2.4.2.2 Demographic Parity. . . . 26
2.4.2.3 Individual Fairness. . . . 26
2.4.2.4 Counterfactual Fairness. . . . 27
2.4.2.5 Causal Feature Selection . . . 27
2.5 Model Monitoring . . . 29
2.5.1 Distribution Shift Formalization . . . 30
2.5.2 Types of Distribution Shift . . . 30
2.5.2.1 Covariate Shift . . . 30
2.5.2.2 Predictions or Label Shift . . . 31
2.5.2.3 Concept or Posterior Shift . . . 31
2.5.3 Model Monitoring Limitation . . . 32

3 Measuring Alignment with Equal Treament 33
3.1 Illustrative Examples . . . 35
3.1.1 College Admissions . . . 35
3.1.2 Paper Blind Reviews . . . 36
3.2 Fairness Formalization . . . 38
3.3 A Model for Monitoring Equal Treatment . . . 38
3.3.1 Formalizing Equal Treatment . . . 38
3.3.2 Equal Treatment Inspector . . . 39
3.4 Theoretical Analysis . . . 41
3.4.1 Equal Treatment Given Shapley Values of Protected Attribute . . 41
3.4.2 Equal Treatment vs Demographic Parity vs Fairness of the Input 42
3.4.3 Statistical Independence Test via Classifier AUC Test . . . 43
3.4.4 Testing for Equal Treatment via an Inspector . . . 44
3.4.5 Statistical Independence Test via Classifier AUC Test . . . 44
3.4.6 Explaining Un-Equal Treatment . . . 45
3.4.7 Using AUC as Classifier Two-Sample Test evaluation metric . . . 47
3.5 Experiments . . . 49
3.5.1 Comparison methods and baselines: . . . 49
3.5.2 Experiments with Synthetic Data . . . 50
3.5.2.1 Alternative Feature Attribution Explanation Methods: Shapley Value Calculations True to the Model or True to the Data? . . . 50
3.5.2.2 Alternative Feature Attribution Explanation Methods: LIME as an Alternative to Shapley Values . . . 53
3.6 Experiments on Real Data . . . 55
3.6.1 Equal Treatment vs Demographic Parity: ACS Census . . . 55
3.6.1.1 ACS Income . . . 55
3.6.1.2 ACS Employment . . . 56
3.6.1.3 ACS Travel Time . . . 57
3.6.1.4 ACS Mobility . . . 58
3.6.2 Quantitative Analysis with Demographic Parity . . . 59

3.6.3 Varying Estimator and Inspector . . . 60
3.6.4 Hyperparameters Evaluation . . . 61
3.7 Related Work on Measuring Equal Treatment . . . 63
3.7.1 Fairness and Auditing . . . 63
3.7.2 Explainability for Fair Supervised Learning . . . 63
3.7.3 Classifier Two-Sample Test . . . 64
3.8 Comparison with Related Fairness Metrics . . . 65
3.8.1 Equal Opportunity/Equalized Odds . . . 66
3.8.2 Treament Equality . . . 67
3.8.3 Demographic Parity . . . 67
3.8.4 Individual Fairness . . . 67
3.8.5 Counterfactual Fairness . . . 68
3.8.6 Fairness Through Unawareness . . . 68
3.8.6.1 Decomposition Method Specific of Demographic Parity. 69
3.9 Discussion . . . 71
3.10 Chapter Conclusions . . . 72

4 How Did Distribution Shift Impact the Model? 73
4.1 A Model for Explanation Shift Detection . . . 74
4.2 Relationships between Common Distribution Shifts and Explanation Shifts 75
4.2.1 Explanation Shift vs Data Shift . . . 76
4.2.1.1 Covariate Shift . . . 76
4.2.1.2 Uninformative Features . . . 76
4.2.2 Explanation Shift vs Prediction Shift . . . 77
4.2.3 Explanation Shift vs Concept Shift . . . 77
4.2.4 Analytical Experiments on Synthetic Data with Non-Linear Models 78
4.2.4.1 Detecting multivariate shift . . . 79
4.2.4.2 Detecting concept shift . . . 79
4.2.4.3 Uninformative features on synthetic data . . . 80
4.2.4.4 Explanation shift that does not affect the prediction . . . 80
4.3 Empirical Evaluation . . . 81
4.3.1 Baseline Methods and Datasets . . . 81
4.3.2 Synthetic Data . . . 82
4.3.3 Novel Group Shift . . . 83
4.3.4 StackOverflow Survey Data . . . 87
4.3.5 Geopolitical and Temporal Shift . . . 87
4.3.5.1 Further Experiments on Real Data . . . 89
ACS Employment . . . 89
ACS Travel Time . . . 89
ACS Mobility . . . 90
4.4 Experiments with Modeling Methods and Hyperparameters . . . 93
4.4.1 Varying Models and Explanation Shift Detectors . . . 93
4.4.2 Hyperparameters Sensitivity Evaluation . . . 94
4.5 LIME as an Alternative Explanation Method . . . 97
4.5.1 Runtime . . . 98
4.6 Related Work on Tabular Data . . . 99
4.6.1 Classifier two-sample test . . . 99

4.6.2 Out-Of-Distribution Detection . . . 100
4.6.3 Detecting distribution shift and its impact on model behaviour . 100
4.6.4 Impossibility of model monitoring . . . 101
4.6.5 Model monitoring and distribution shift under specific assumptions: . . . 101
4.6.6 Explainability and distribution shift . . . 101
4.7 Discussion . . . 102
4.8 Conclusions . . . 104

5 Building Indicators of Model Performance Deterioration 105
5.1 Related Work . . . 107
5.1.1 Model Monitoring . . . 107
5.1.2 Uncertainty . . . 108
5.2 Methodology . . . 109
5.2.1 Evaluation of Deterioration Detection Systems . . . 109
5.2.2 Uncertainty Estimation . . . 110
5.2.3 Detecting the Source of Uncertainty/Model Deterioration . . . 110
5.3 Experiments . . . 111
5.3.1 Evaluating Model Deterioration . . . 111
5.3.2 Detecting the Source of Uncertainty . . . 113
5.4 Synthetic Data Experiments . . . 114
5.4.1 Evaluating Model Deterioration . . . 115
5.4.2 Detecting the Source of Uncertainty/Model Deterioration . . . 116
5.5 Computational Performance Comparison . . . 117
5.6 Conclusion . . . 118
5.6.1 Limitations: . . . 119

6 Conclusion and Remarks 121
6.1 Summary of Findings . . . 121
6.2 Contributions to the Field . . . 122
6.2.1 (RQ-1) How can we use feature attributions distributions for post-deployment monitor in the absence of labeled truth data? . . . 122
6.2.2 (RQ-2) Can feature attribution distributions be used to measure alignment of post-deployed models in a unsupervised monitoring set up?? . . . 123
6.2.3 (RQ-3) How did Distribution Shift Impact the Model? . . . 123
6.2.4 (RQ-4) How can we identify and explain changes in model performance using feature attribution distribution? . . . 123
6.2.5 Software Contributions . . . 123
6.2.6 Limitations and Reflections . . . 124
6.2.6.1 Reliability of Explanations . . . 124
6.2.6.2 Alternative Explanation Methods . . . 125
6.2.6.3 Beyond Tabular Data . . . 125
6.2.6.4 Complexity and Practicality of Feature Attribution-Based Monitoring . . . 125
6.3 Future Directions . . . 126
6.3.1 Software Improvement . . . 126
6.3.2 Integration with Real-world Applications . . . 126

6.3.3 Ethical and Societal Implications 126
6.3.4 Longitudinal Studies 127

List of Figures **128**

List of Tables **133**

References **135**

LIST OF PUBLICATIONS

The research presented in Chapter 3 has been disseminated through the following publication:

- **1.Mougan, C**, Ferrara, A., State, L., Ruggieri, S., & Staab, S. Beyond Demographic Parity: Redefining Equal Treatment
  NeurIPS 2023 workshop on AI meets Moral Philosophy
  Mougan et al. (2023d)

The research presented in Chapter 4 has been disseminated through the following publication:

- **2.Mougan, C.**, Broelemann, K., Kasneci, G., Tiropanis, T., and Staab, S. (2022) Explanation shift: Detecting distribution shifts on tabular data via the explanation space. *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications* Mougan et al. (2022)
- **3.** Mougan, C., Broelemann, K., Kasneci, G., Tiropanis, T., and Staab, S. (2025). Explanation shift: How did the distribution shift impact the model? In TMLRMougan et al. (2025)

The research presented in Chapter 5 has been disseminated through the following publication:

- **4.Mougan, C.**, & Nielsen, D. S. (2023).
  Monitoring Model Deterioration with Explainable Uncertainty Estimation via Non-parametric Bootstrap.
  In AAAI Conference on Artificial Intelligence,2023 Mougan and Nielsen (2023)

Additional publications, while not directly incorporated into this thesis, have enriched the understanding of the field and influenced some of the proposed solutions:

- **5.Mougan, C.**, Plant, R., Teng, C., Bazzi, M., Ejea, A. C., Chan, R. S.-Y., Jasin, D. S., Stoffel, M., Whitaker, K. J., and Manser, J. (2023).
  How to data in datathons.
  Advances in Neural Information Processing Systems Mougan et al. (2023c)

- **6.Mougan, C.**, Alvarez, J., Ruggieri, S., and Staab, S. (2023).
  Fairness implications of encoding protected categorical attributes.
  In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES Association for Computing Machinery Mougan et al. (2023a)

- **7.Mougan, C**, Masip, D., Nin, J., and Pujol, O. (2021).
  Quantile encoder: Tackling high cardinality categorical features in regression problems. *Modeling Decisions for Artificial Intelligence - 18th International Conference, MDAI 2021* Mougan et al. (2021b)

- **8.Mougan, C**, Kanellos, G, and Gottron, T.
  Desiderata for Explainable AI in Statistical Production Systems of the European Central Bank. *ECMLPKDD Workshop on Bias (2021)* Mougan et al. (2021a)

- **9.Mougan, C**, Kanellos, G., Micheler, J., Martinez, J., & Gottron, T. (2022).
  Introducing explainable supervised machine learning into interactive feedback loops for statistical production system. *Irving Fisher Committee (IFC) - Bank of Italy workshop on Data science in central banking: Applications and tools.* Mougan et al. (2023b)

- **10.**Yasar, A. G., Chong, A., Dong, E., Gilbert, T. K., Hladikova, S., Maio, R., **Mougan, C**, Shen, X., Singh, S., Stoica, A.-A., Thais, S., and Zilka, M. (2023).
  AI and the EU digital markets act: Addressing the risks of bigness in generative AI.
  In Proceedings of the 1st Workshop on Generative AI and Law. International Conference on Machine Learning (ICML) Yasar et al. (2023)

- **11.**Papageorgiou, I. and **Mougan, C** (2023).
  Processing sensitive personal data for bias monitoring under the AI act: Necessity principle.
  In Beyond Data Protection Conference 2023
  NeurIPS 2023 workshop on Regulatable ML Papageorgiou and Mougan (2023)

- **12.**Alvarez, J. M., Colmenarejo, A. B., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., Ghodsi, S., **Mougan, C.**, Papageorgiou, I., Reyero, P., Russo, M., Scott, K. M., State, L., Zhao, X., and Ruggieri, S.
  Policy advice and best practices on bias and fairness in AI.
  *Ethics and Information Technology*Alvarez et al. (2024)

- **13.**Leslie, D., Ashurst, C., Gonzalez, N. M., Griffiths, F., Jayadeva, S., Jorgensen, M., Katell, M., Krishna, S., Kwiatkowski, D., Martins, C. I., Mahomed, S., **Mougan,**

C., Pandit, S., Richey, M., Sakshaug, J. W., Vallor, S., and Vilain, L.
Frontier AI, power, and the public interest: Who benefits, who decides? *Harvard Data Science Review* Leslie et al. (2024)

- **14.** Ayvaz, D. S., Belenguer, L., He, H., Kanubala, D. D., Li, M., Low, S., **Mougan, C.**, Onwuegbuche, F. C., Pi, Y., Sikora, N., Tran, D., Verma, S., Wang, H., Xie, S., and Pelletier, A. (2025). Measuring fairness in financial transaction machine learning modelsAyvaz et al. (2025)

- **15.**Yasar, A. G., Chong, A., Dong, E., Gilbert, T. K., Hladikova, S., **Mougan, C.**, Shen, X., Singh, S., Stoica, A.-A., and Thais, S. (2024). Integration of generative AI in the digital markets act: Contestability and fairness from a cross-disciplinary perspective. In Working Paper Series. London School of Economics.Yasar et al. (2024)

- **16. Mougan, C.** AI Alignment via Virtue-Based Selected Feedback. Lyceum Project - AI Ethics with Aristotle.

- **17.**Pugnana, A., **Mougan, C**, Saatrup, D., Model Agnostic Explainable Selective Regression via Uncertainty Estimation

- **18.Mougan C.**, Brand, J., Kantian Deontology Meets AI Alignment: Towards Morally Robust Fairness Metrics Mougan and Brand (2023)

- **19.**Castillo-Eslava, F., **Mougan, C**, Romero-Reche, A., and Staab, S. (2023). The role of large language models in the recognition of territorial sovereignty: An analysis of the construction of legitimacy Castillo-Eslava et al. (2023)

Declaration of Authorship

I declare that thesis and the work presented in it are my own and have been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself; Parts of this work have been published as:
   (a) *Mougan, C., Broelemann, K., Kasneci, G., Tiropanis, T., and Staab, S. (2022).Explanation shift: Detecting distribution shifts on tabular data via the explanation space. In NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications* Mougan et al. (2022) — Which has been mainly my contribution (90%) which editing and writing input from the coauthors.
   (b) *Mougan, C., Broelemann, K., Kasneci, G., Tiropanis, T., and Staab, S. (2025).Explanation Shift: How did Distribution Shift Impact the Model? Transaction in Machine Learning Research* — Which has been mainly my contribution (90%) which editing and writing input from the coauthors.

(c) *Mougan, C., Ferrara, A., State, L., Ruggieri, S., & Staab, S. Equal Treatment: Fairness Measures with Explanation Distributions* Which has been mainly my contribution (80%) which editing, writing and mathematical formalization input from the coauthors. Particularly, math formalization into lemmas has been contributed by Salvatore Ruggieri.

(d) *Mougan, C., & Nielsen, D. S. (2022). Monitoring Model Deterioration with Explainable Uncertainty Estimation via Non-parametric Bootstrap. In AAAI Conference on Artificial Intelligence,2023* Which has been an equal contribution my part relied on using uncertainty as a proxy for performance degradation and using explainability techniques to understand the reasons behind it.

Signed:..... Date:.....

Chapter 1

Introduction to Model Monitoring

Model monitoring refers to the practice of observing and understanding the behaviour of Artificial Intelligence (AI) systems in production (Huyen, 2022; Burkov, 2020). It involves collecting and analyzing data from deployed AI models and systems to ensure they are functioning as intended and to identify and diagnose any issues or anomalies that may arise during their operation.

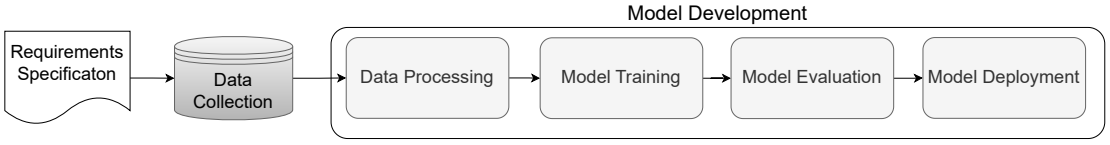

FIGURE 1.1: Model monitoring occurs within the AI system life cycle. During model development, task requirements and data are collected, models are trained and evaluated, and parameters are refined. Finally, the model is assessed against benchmarks and metrics using a hold-out dataset.

The process of model monitoring unfolds within the broader context of the AI system life cycle, as illustrated in Figure 1.1. After the application task requirements specifications are collected and needed data is selected, AI models undergo training and evaluation using available data during the model developmental phase. In this phase, model parameters are learned, tuned and refined. Then, the model is evaluated using the available data from a hold-out set against established benchmarks and metrics.

Subsequently, based on the results of this evaluation, models are deployed into production environments. This shift marks the transition from the controlled development environment, with available ground truth data, to the dynamic and frequently unpredictable landscape of real-world deployment (Paleyes et al., 2023), as illustrated in Figure 1.1. This stage not only marks a significant point in the AI life cycle but also a moment when understanding how the model is going to behave, which is crucial for the successful and responsible application of AI models.

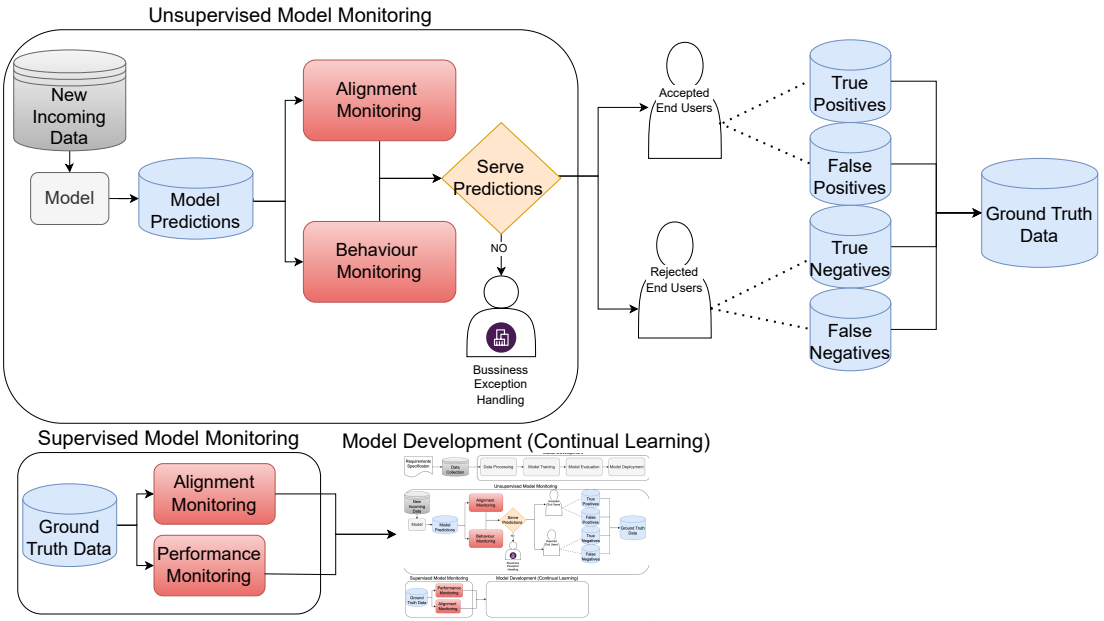

FIGURE 1.2: Life cycle of monitoring a supervised learning binary classifier. Once a model is deployed, it encounters new, unlabelled data, prompting the need to observe and analyse its behaviour to detect potential issues before serving the predictions in the real world. In some cases, after serving the model to users, the acquisition of ground truth data allows for further evaluation. Leveraging these results and the available ground truth data, the original model can undergo retraining.

We can distinguish two types of model monitoring: before serving the model predictions (unsupervised model monitoring) or after, when it is often possible to obtain ground truth data collection (supervised model monitoring). However, it is at the first stage of model monitoring, depicted in Figure 1.2, where the primary focus of this thesis lies; when newly arrived, unlabeled data triggers the model to make predictions, right before serving those predictions to the world (e.g. users). It is precisely during this phase that we observe and analyse the model's predictions to detect any undesirable behaviour before it reaches real-world applications. This specific phase, prior to serving the model predictions, presents a unique challenge: the *absence of ground truth data* that makes it extremely hard and often impossible to estimate conventional model performance metrics such as accuracy, precision, AUC, or F1-score and social inequality metrics such as equal opportunity, average absolute odds or equal odds. Within this step of an AI model's life cycle, we pinpoint two paramount challenges of model monitoring:

- **AI Alignment**: relies on ensuring that the AI systems deployed align with a diverse set of human values.
- **Model Monitoring**: Accounting for changes in the model's behaviour and performance on the new incoming data.

After performing an unsupervised model monitoring assessment, we reach the stage in which we have to decide whether to serve those predictions. This assessment is particular to each task, and the final decision lies within the domain of engineering and business teams. Opting not to serve the predictions to the users leads us back to revisiting the problem definition, as the absence of labeled data prevents retraining. On the other hand, if we accept those predictions and decide to serve them to the users, then in some situations, there is the possibility of collecting ground truth data. Acquiring labeled data after deployment poses a considerable challenge and is often impractical, exhibiting well-known biases, such as confounding, selection, and missingness (Feng et al., 2023; Ruggieri et al., 2023). For example, in a loan lending scenario, after the period is due, we will be able to collect the data for the users who we gave the loan, but we won't be able to collect it for the users who we did not offer the loan, creating then a biased ground truth data.

Once we have obtained ground truth data, the second step of model monitoring can be performed. In this step, other metrics based on the model errors can be performed. If those metrics are satisfactory enough, we can proceed to make predictions in the new incoming data; if they are not and we want to optimize further, in some situation, we can retrain the model on this new data, a process called "continual learning" (Diethe et al., 2019; Pham et al., 2021; Knoblauch et al., 2020).

This thesis centres on the previous step, the unsupervised monitoring stage, where labeled post-deployment data is unavailable. The techniques applicable in this context are relatively limited. Specifically, this thesis seeks to understand how features attribution distribution can be used as this post-deployment monitoring stage.[1]. This focus leads to the first research question of the thesis:

"(Research Question 1) *How can we use feature attributions distributions for post-deployment monitoring in the absence of labelled truth data?*"

We will divide this thesis's overarching research question into two main parts: monitoring for performance and AI Alignment. We now introduce more fine-grained research questions aligned with the current challenges of both dimensions of unsupervised model monitoring. For AI alignment, we start by providing two social science perspectives and then the technical contribution of the thesis that aims to align with them. For model monitoring, we divide the problem into two tasks: understanding changes in the model under distribution shift and building indicators of model performance deterioration. An overarching principle of this thesis is its commitment to open science. To uphold the focus on openness and reproducibility, we will compile a set of requirements designed to ensure replicable research outcomes and the development of practical software tools.

[1]The foundations of feature attribution methods will be discussed in Chapter 2

1.1 AI Alignment and Feature Attributions Challenges

The AI alignment field relies on ensuring that AI systems are oriented with a diverse set of human values, encompassing moral, societal, justice-oriented, and other relevant principles (Gabriel, 2020, 2022).

In the field of computer science, researchers have traditionally quantified various statistical disparities that exist among protected attributes. These measures are often labelled using terms from the social sciences (Binns, 2018; Mitchell et al., 2021), such as equal opportunity (Hardt et al., 2016) and equal treatment (Candelas et al., 2023). These notions have been incorporated into the study of algorithmic fairness to address historical and systemic biases that can arise in decision-making systems. Researchers have subsequently proposed methods and algorithmic approaches to improve these metrics. Deciding which metrics are best suited in each situation is an important yet complex task as researchers have demonstrated incompatibility between certain metrics, thus arguing that several AI fairness concepts cannot be achieved simultaneously (Chouldechova, 2017; Kleinberg et al., 2017; Hardt et al., 2016; Hsu et al., 2022). It should be noted that concepts in social sciences are not designed with mathematical precision, which makes it challenging to quantify statistical differences in a straightforward mathematical way directly. Instead, they emerge from the understanding of historical events, social inequalities, and discriminatory practices.

At the intersection of social sciences fields and computer science, there exists an already recognized misalignment (Fazelpour and Lipton, 2020). In light of this, there are several research lines that propose frameworks to help understand existing notions of algorithmic fairness in social science terms (Heidari et al., 2019; Simons et al., 2021). Within the context of this thesis, which focuses on post-deployment monitoring without labeled data and explores the use of feature attribution distributions, we pose the second research question:

"(Research Question 2) *Can feature attribution distributions be used to measure alignment of post-deployed models in a unsupervised monitoring set-up?*"

This thesis will introduce novel metrics based on feature attribution distributions, aligned with two core principles of AI alignment, demonstrating how these distributions can be utilized to effectively measure them:

- Moral values grounded in Kantian deontological philosophy.
- Equality notions based on distributive justice principles that are derived from political philosophy.

1.2 Performance Monitoring with Feature Attributions Distributions

The second and subsequent part of this PhD thesis centres on performance aspects of model monitoring, covering chapters 4 and 5. In supervised machine learning, model monitoring of predictive performance is an important step of the ML model pipeline as it directly relates to the quality of the application (Huyen, 2022; Burkov, 2020). Monitoring machine learning models once they have been deployed is not an easy task. Model monitoring ensures that a machine learning application in a production environment displays consistent behaviour over time. Understanding how distribution shift impacts the model behaviour is paramount in deployed ML models. It enables us to gauge how well models adapt to changing data distributions, ensuring their reliability in dynamic real-world scenarios. Moreover, such investigations aid in addressing critical potential societal issues and ethical concerns by identifying and mitigating post-deployment biases promoting responsible AI deployment. By understanding how models respond to distribution shifts, we can enhance model accountability, allocate resources efficiently, and meet regulatory requirements.

With this challenge in mind and the usage of feature attribution distributions this thesis covers the following research question:

"Research Question (3) *By looking at changes in the distributions of explanations, can we measure how did distribution shift impact the model?*"[2]

Current methods to detect shifts in the input or output data distributions have limitations in identifying model behaviour changes, where we recollect our research requirement *(REQ-05)*, measuring the impact of distribution shift in input data in the model. Extensive related work has aimed at detecting that data is from out-of-distribution. To this end, they have created several benchmarks that measure whether data comes from in-distribution or not (Koh et al., 2021; Sagawa et al., 2021; Malinin et al., 2021a, 2022, 2021b). However, many methods developed for detecting out-of-distribution data are specific to neural networks processing image and text data and can not be applied to traditional machine learning techniques. These methods often assume that the relationships between predictor and response variables remain unchanged, i.e., no concept shift occurs. In this thesis, we focus on unlabeled monitoring, where ground truth from deployment data is not available, forming research requirement *(REQ-06)*. In this thesis, a research requirement is a statistical comparison between how predictions from training data are explained and how predictions on new data are explained, *(REQ-07)*. Also, identifying and having explanation capabilities of the changes in the model behaviour is a research requirement *(REQ-08)*

[2] To be answered in Chapter 4

Once we have proposed and studied methods based on feature attributions distributions to understand the impact of distribution shifts on the model, the next step is to assess whether these shifts have affected the model's performance. This task is a well known impossible challenge without further assumption, therefore developing an approach to measure performance decay is notoriously difficult. In the final research question of this thesis, we will study this open ended challenge, exploring how performance degradation can be identified and explained under a specific set of conditions using feature attributions distributions:

"Research Question (4) *How can we identify and explain changes in model performance using feature attribution distribution?*"[3]

1.2.1 Reproducibility and Software Contribution Objectives

Having well-defined software requirements in a computer science thesis is crucial to ensure clarity, reproducibility, and applicability of the research. Open-source software enables other researchers to replicate the study findings, validating the results and ensuring their reliability across different settings, which is a fundamental aspect of scientific research, as captured in requirement *REQ-09*. In addition to releasing open-source packages, the software used for experiments should be available in the form of a software repository. This facilitates systematic evaluation, scrutiny, and comparison of the developed experiments against other benchmarks or existing solutions, as detailed in requirement *REQ-10*.

[3] To be answered in Chapter 5

1.2.2 Summary Requirements for Model Monitoring in the Absence of Labeled Data via Feature Attributions Distributions

TABLE 1.1: Requirements Elicitation Table for Equal Treatment

Requirement	Description	Source
Alignment Requirements		
REQ-01 Unlabeled Data	Metrics should prioritize assessing disparities in model decisions rather than fixating on model errors.	Society: Deontological Ethics
REQ-02 Group Discrimination	Decisions based on protected characteristics that individuals did not choose at birth.	Society: Politics & Ethics
REQ-03 Measuring Proxy Discrimination	Detecting that certain features may indirectly contribute to biased decisions of non-chosen factors at birth (protected attributes)	Society(Liberal Political Philosophy)
REQ-04 Explanation Capability	Offer both theoretical underpinnings and empirical validation of the sources driving discrimination.	End User
Performance Monitoring		
REQ-05 Model-Data Interactions	The system shall measure and quantify the impact of distribution shifts in input data in the model	Monitoring Objectives
REQ-06 Unlabeled Data	The system should not require labelled test data	Monitoring Objectives
REQ-07 Explanation Distributions	The system shall identify model explanation changes resulting from distribution shifts.	Monitoring Objectives
REQ-08 Explanation Capabilities	The system shall provide insights into how models respond to different types of distribution shifts.	System Users
Software Requirements		
REQ-09 Open Source	An open-source software package with clear and comprehensive documentation, including installation guides, usage instructions, tutorials covering both basic and advanced usage.	Developers & Scientific Community
REQ-10 Reproducibility	The experiments should be available with sample datasets and examples showcasing the experiments used in this thesis. Additionally, there should be a feedback mechanism for users to report issues and ask questions.	Developers & Scientific Community

1.3 Overview of the Thesis

The opening chapter of this thesis sets the stage for what is unsupervised model monitoring within machine learning systems. This chapter is divided into two main sections, each addressing different facets of AI monitoring: AI alignment challenges and performance monitoring.

Chapter 2, titled "Foundations," provides the theoretical framework essential for understanding and addressing model monitoring and AI alignment. The chapter begins by establishing the necessary mathematical notation. It then provides the foundations for feature attribution explanations (Shapley values and LIME). Then, AI alignment explores how AI can adhere to distributive justice and moral philosophy principles such as egalitarianism, socialism, liberalism, utilitarianism, and deontology. The chapter also introduces the different types of distribution shifts and model monitoring challenges.

Chapter 3, titled "Measuring AI Alignment with Equal Treatment," is dedicated to exploring and establishing methodologies for ensuring AI systems align with philosophical notions of equal treatment. The chapter uses illustrative examples, such as college admissions and paper-blind reviews, to demonstrate the practical implications and challenges of AI in decision-making processes that require impartiality. It present a methodology for monitoring equal treatment in AI systems. The chapter engages with theoretical analyses, comparing various fairness metrics like equal treatment, demographic parity, and individual fairness, and includes experiments with both synthetic and real data to test these concepts. Through these discussions and experiments, the chapter aims to refine the tools and methods needed to evaluate and ensure that AI systems administer equal treatment across different scenarios, aligning closely with both distributive justice concepts and moral values.

Chapter 4, titled "How Did Distribution Shift Impact the Model?", examines the effects of various types of distribution shifts on machine learning models, particularly focusing on how these shifts influence explanations provided by models. The chapter introduces a model designed to detect explanation shifts and explores relationships between common distribution shifts—like covariate shift, prediction shift, and concept shift—and their impacts on model explanations. It studies analytical and empirical evaluations using both synthetic data and real-world datasets, such as StackOverflow survey data and ACS derived datasets. It concludes with a discussion on the challenges of monitoring models under distribution shifts and the implications for model explainability.

Chapter 5, titled "Building Indicators of Model Performance Deterioration," develops an indicator to detect and evaluate the deterioration in machine learning model performance over time. It reviews related works on model monitoring and the role of

uncertainty in assessing model performance. The chapter describes various methods to evaluate model deterioration detection systems and estimate the uncertainty that can lead to model decay. Experiments are conducted using synthetic and classical ML data benchmarks to assess these methods. It concludes with a discussion on the limitations of current approaches and the potential for further research in the area of dynamic model monitoring and maintenance.

Chapter 6, the conclusion of the thesis, synthesizes the key findings, delineates the contributions made to the field, and outlines future research directions. It summarizes the results of investigations into AI model monitoring, fairness, and performance.

Chapter 2

Foundations

In this chapter, we present the foundations on which this thesis is build. We first introduce the formalization and then provide *(i)* explainable AI foundations *(ii)* model monitoring and finally *(iii)* AI alignment foundations.

2.1 Mathematical Notation

TABLE 2.1: Notation used throughout the paper.

Notation	Meaning
f_θ	Supervised learning model.
X	Predictive features.
Y	Target feature.
(x_i, y_i)	Observation in the training set.
$\text{dom}(Y)$	Domain of the target feature.
$f_\theta(x)$	Model prediction for x
$\mathcal{D}^{tr}$	Training set.
$\mathcal{D}^{new}$	Unseen dataset at training.
$\mathcal{D}_X$	Projection of $\mathcal{D}$ on X.
$\mathcal{D}_X^{tr}$	Covariate training dataset input.
$\mathcal{D}_X^{new}$	Covariate unseen dataset input.
$f_\theta(\mathcal{D}_X)$	Empirical prediction distribution.
N	Size of X.
B	Number of bootstrap samples.
b	The index of a particular bootstrap sample.
$f_\theta(x^\star)$	Instance to be predicted.
$\mathcal{S}(f, x^\star)$	Shapley values for $x^\star$ and model f.
$\mathcal{S}_i(f, x^\star)$	Shapley value of feature X_i.
$\mathcal{S}_j(f_\theta, \mathcal{D}_X)$	Shapley value of feature j calculated from dataset $\mathcal{D}_X$.
$\mathbf{P}(\mathcal{D}_X)$	Distribution of dataset X.
$\mathbf{P}(\mathcal{S}(f_\theta, \mathcal{D}_X))$	Distribution of Shapley values calculated from dataset $\mathcal{D}_X$.

In supervised learning, a function $f_\theta : X \to Y$, also called a model, is induced from a set of observations, called the training set, $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \sim X \times Y$, where $X = \{X_1, \ldots, X_p\}$ represents predictive features and Y is the target feature. The domain of the target feature is $\text{dom}(Y) = \{0, 1\}$ (binary classification) or $\text{dom}(Y) = \mathbb{R}$ (regression). For binary classification, we assume a probabilistic classifier, and we denote by $f_\theta(x)$ the estimates of the probability $P(Y = 1|X = x)$ over the (unknown) distribution of $X \times Y$. For regression, $f_\theta(x)$ estimates $E[Y|X = x]$. We call the projection of $\mathcal{D}$ on X, written $\mathcal{D}_X = \{x_1, \ldots, x_n\} \sim X$, the empirical *input distribution*. The dataset $f_\theta(\mathcal{D}_X) = \{f_\theta(x) \mid x \in \mathcal{D}_X\}$ is called the empirical *prediction distribution*.

2.2 Feature Attribution Explainations

Explainability has become an important concept in legal and ethical data processing guidelines and machine learning applications Selbst and Barocas (2018). A wide variety of methods have been developed to account for the decision of algorithmic systems Guidotti et al. (2018b); Mittelstadt et al. (2019); Arrieta et al. (2020).

2.2.1 Shapley Values

One of the most popular approaches to explaining machine learning models has been the use of Shapley values to attribute relevance to features used by the model Lundberg et al. (2020a); Lundberg and Lee (2017). The Shapley value is a concept from coalition game theory that aims to allocate the surplus generated by the grand coalition in a game to each of its players Shapley (1953). The Shapley value $\mathcal{S}_j$ for the j'th player can be defined via a value function $\text{val} : 2^N \to \mathbb{R}$ of players in T:

$$\mathcal{S}_j(\text{val}) = \sum_{T \subseteq N \setminus \{j\}} \frac{|T|!(p - |T| - 1)!}{p!} (\text{val}(T \cup \{j\}) - \text{val}(T)) \tag{2.1}$$

In machine learning, $N = \{1, \ldots, p\}$ is the set of features occurring in the training data, and T is used to denote a subset of N. Given that x is the feature vector of the instance to be explained, and the term $\text{val}_x(T)$ represents the prediction for the feature values in T that are marginalized over features that are not included in T:

$$\text{val}_{f,x}(T) = E_{X|X_T = x_T}[f(X)] - E_X[f(X)] \tag{2.2}$$

The Shapley value framework satisfies several theoretical properties, which we will describe in the following section. Our approach works with feature attribution techniques that fulfill efficiency and non-informative properties, and we use Shapley values as an example.

2.2.1.1 Efficiency.

The feature contributions must add up to the difference of prediction $x^\star$ and the expected value:

$$\sum_{j \in N} \mathcal{S}_j(f, x^\star) = f(x^\star) - E[f(X)]) \tag{2.3}$$

It is important to differentiate between the theoretical Shapley values and the different implementations that approximate them. We use TreeSHAP as an efficient implementation of an approach for tree-based models of Shapley values Lundberg et al. (2020a); Molnar (2019); Zern et al. (2023), particularly we use the observational (or path-dependent) estimation Chen et al. (2022a); Frye et al. (2020); Chen et al. (2020) and for linear models we use the correlation dependent implementation that takes into account feature dependencies Aas et al. (2021).

The following property only holds for the interventional variant (SHAP values), but not for the observational variant.

2.2.1.2 Uninformativeness.

A feature X_j that does not change the predicted value (i.e., for all x, x'_j: $f(x_{N\setminus\{j\}}, x_j) = f(x_{N\setminus\{j\}}, x'_j)$) have a Shapley value of zero, i.e., $\mathcal{S}_j(f, x^\star) = 0$.

There are two variants for the term $\text{val}(T)$ Aas et al. (2021); Chen et al. (2020); Zern et al. (2023): the *observational* and the *interventional*. When using the observational conditional expectation, we consider the expected value of f over the joint distribution of all features conditioned to fix features in T to the values they have in $x^\star$:

$$\text{val}(T) = E[f(x^\star_T, X_{N\setminus T}) | X_T = x^\star_T] \tag{2.4}$$

where $f(x^\star_T, X_{N\setminus T})$ denotes that features in T are fixed to their values in $x^\star$, and features not in T are random variables over the joint distribution of features. Opposed, the interventional conditional expectation considers the expected value of f over the marginal distribution of features not in T:

$$\text{val}(T) = E[f(x^\star_T, X_{N\setminus T})] \tag{2.5}$$

In the interventional variant, the marginal distribution is unaffected by the knowledge that $X_T = x_T^\star$ (Aas et al., 2021). In general, the estimation of (2.4) is difficult, and some implementations (e.g., SHAP) consider (2.5) as the default one Chen et al. (2020). In the case of decision tree models, TreeSHAP offers both possibilities(Lundberg et al., 2020a).

The interventional SHAP values are said to stay "true to the model", which means they will only give credit to the model's features and do not consider any correlation between inputs. In contrast, the full conditional SHAP values stay "true to the data", considering how the model would behave when respecting the correlations in the input data (Chen et al., 2020). Therefore, the full conditional SHAP values only credit the features relevant to the model behavior, while the interventional SHAP values allow us to explore the effect of changing inputs.

The only option supported for sparse cases is the interventional option. Thus, interventional SHAP values can be useful for sparse models, while full conditional SHAP values benefit models with highly correlated inputs. Overall, both interventional and full conditional SHAP values are useful for explaining the behaviour of machine learning models and providing insights into the most important features of model performance.

Linear Models and IID Data In the case of a linear model $f_\beta(x) = \beta_0 + \sum_j \beta_j \cdot x_j$, the SHAP values turns out to be $\mathcal{S}(f, x^\star) = \beta_i(x_i^\star - \mu_i)$ where $\mu_i = E[X_i]$. For the observational case, this holds only if the features are independent (Aas et al., 2021).

2.2.2 LIME: Local Interpretable Model-Agnostic Explanations

Another widely used feature attribution technique is LIME (Local Interpretable Model-Agnostic Explanations). The intuition behind LIME is to create a local linear model that approximates the behavior of the original model in a small neighbourhood of the instance to explain (Ribeiro et al., 2016b,a), whose mathematical intuition is very similar to the Taylor/Maclaurin series. Instead of attempting to interpret the entire model globally, LIME focuses on understanding predictions at a local level. It achieves this by perturbing or sampling data points in the vicinity of the prediction of interest and fitting a simple, interpretable model to approximate the complex model's behavior in that local region. The interpretable model, often a linear model, decision tree or a decision rule, can be easily understood and analyzed (Guidotti et al., 2018a).

This approach effectively "explains" a model's prediction by providing insight into which features were most influential in the local context. LIME quantifies the feature importance, enabling users to grasp why the model made a particular prediction for a specific data point. This local approximation is crucial because it reflects the model's behavior regarding the instance in question, which may differ from its global behavior.

In this thesis we use the implementation based on the Euclidean version of LIME, called "tabular LIME" and it involves the following steps:

- **Selection of Data Point**: Choose the data point for which you want an explanation.
- **Data Perturbation:** Generate a dataset of similar data points by perturbing the chosen instance. This step is crucial for approximating local behavior.
- **Prediction Collection:** Obtain model predictions for the perturbed data points.
- **Surrogate Model Creation:** Fit an interpretable model, such as a linear regression model, to the perturbed data and their corresponding model predictions. This surrogate model approximates the complex model's behavior locally.
- **Explanation Generation:** The surrogate model's coefficients are used to estimate the importance of each feature for the chosen prediction, resulting in an explanation of the model's decision for that instance.

LIME is a versatile and model-agnostic technique, which means it can be applied to virtually any machine learning model without requiring knowledge of the model's internal architecture. This makes it a valuable tool for interpretable AI in diverse domains.

2.3 AI Alignment

AI value alignment aims to ensure that AI is properly aligned with human values (Gabriel, 2020). As computer systems continue to gain greater autonomy and impact, AI value alignment takes on greater importance. The rapid decision-making capabilities of AI systems, combined with their autonomy, raise the stakes when it comes to ensuring that these systems align with human values. In real-time scenarios, AI may encounter complex ethical dilemmas, making it increasingly challenging for human oversight and evaluation of each action. This lack of oversight can lead to issues of accountability and ethical risks, underscoring the need to establish mechanisms that ensure AI acts according to a set of human values.

We can divide the alignment challenge into two parts (Gabriel, 2020). A first *normative* question asks which values AI should align. A second challenge is *technical* and focuses on mathematically formalising those values or principles in AI systems to capture the previously discussed notions.

The first part of the challenge raises questions about the very nature of "value" and its subjectivity, as it serves as a placeholder for a wide array of concepts. Values are fundamental beliefs or principles that guide behavior, decision-making, and judgments

about what is right or wrong, good or bad, desirable or undesirable. This leads us into a meta-normativity tecno-legal debate about the objectivity of values. In this thesis, we acknowledge this meta-normativity debate, but for the purpose of practicality, we opt for a classic and shallow definition of values:

"*In its broadest sense, 'value theory' is a catch-all label used to encompass all branches of moral philosophy, social and political philosophy, aesthetics, and sometimes feminist philosophy and the philosophy of religion*" Schroeder (2021)

In this thesis, we will limit the discussion of AI alignment to two main branches of philosophy and two specific values of those branches, which will be discussed in the next section: *(i)* the branch of political philosophy with the value of distributive justice and ***(ii*** the branch of ethics, centred on the value of moral rightness.

The second part of the challenge relies on the mathematical formalization of those values. The challenge becomes especially intricate when considering the diverse and dynamic nature of human values, ranging from ethical principles and cultural norms to personal preferences and moral priorities. Developing a mathematical framework that captures this intricate web of values demands a balance between flexibility and precision. It often requires acknowledging that human values are not fixed or universally agreed upon. A fundamental question is, *if values lack an objectively quantifiable foundation, how can AI be developed to align with them?* Gabriel (2020) proposes a distinction in order to ease this translation of human values into computational metrics:

> *Here it is useful to draw a distinction between minimalist and maximalist conceptions of value alignment. The former involves tethering artificial intelligence to some plausible schema of human value and avoiding unsafe outcomes. The latter involves aligning artificial intelligence with the correct or best scheme of human values on a society-wide or global basis. While the minimalist view starts with the sound observation that optimizing exclusively for almost any metric could create bad outcomes for human beings, we may ultimately need to move beyond minimalist conceptions if we are going to produce fully aligned AI. This is because AI systems could be safe and reliable but still a long way from what is best—or from what we truly desire.*(Gabriel, 2020)

In this thesis, we deviate from minimalist alignments and explore the foundational aspects of two distinct branches of philosophy: political philosophy and moral philosophy. These branches can serve to establish the fundamental humanities' fundamental pillars, guiding systems in accordance with the values, principles and standards found within these fields. We will now present the fundamental concepts underpinning each of these values and relate them to their significance in the landscape of AI alignment.

2.3.1 Distributive Justice Political Philosophy Notions

In philosophy, longstanding discussions about what constitutes a fair or unfair political system have led to established frameworks of distributive justice. Justice is a broad and foundational concept that pertains to the fair and equitable treatment of individuals and groups in society. It encompasses the idea of treating people with fairness, impartiality, and according to moral or legal principles. Even though there are several types of justice (retributive, procedural, restorative, or corrective justice), this thesis focuses on the intersection of distributive justice and supervised learning algorithms. Distributive justice is a specific subset of justice that deals with the fair allocation and distribution of resources, benefits, and burdens in a society. It is concerned with questions of how economic and social goods (such as income, wealth, education, healthcare, and opportunities) should be distributed among individuals and groups, emphasizing that circumstances not chosen at birth should not be the basis for disparate distribution (Rawls, 1971; Lamont and Favor, 2017).

Different theories of distributive justice propose varying principles for achieving a fair distribution of goods in society. We will now introduce a few of them discussed in machine learning research.

2.3.1.1 Egalitarianism and Equal Opportunity

Egalitarianism is a moral and political philosophy that advocates for the ideal of equality among individuals in society. At its core, egalitarianism stands for the belief that all people should be treated with equal dignity, respect, and worth, regardless of their inherent characteristics, social status, or circumstances of birth. This philosophy places a strong emphasis on the principle of opportunity, asserting that individuals should have equal access to resources, opportunities, and benefits. Egalitarianism strives to create a society characterized by fairness and justice, aiming to minimize, if not eliminate, disparities in access to income, wealth, and essential services. It asserts that factors beyond an individual's control should not be translated into social inequalities.

In this context, **equal opportunity** is a fundamental mechanism for achieving egalitarian goals. Egalitarians argue that by providing individuals with an equal chance to access opportunities like education, employment, and social services, society can move closer to the ideal of fairness, as individuals are not systematically disadvantaged due to factors beyond their control.

Equal opportunity is a practical means for addressing structural barriers that hinder individuals from realizing their potential and is instrumental in levelling the playing field. Egalitarians often advocate for policies and practices that promote both equal

opportunity and their broader egalitarian objectives, such as affirmative action to rectify historical injustices, progressive taxation to reduce economic disparities, and social safety nets to ensure that all individuals have access to essential services. In essence, equal opportunity is a fundamental component of the egalitarian vision, working hand in hand with egalitarianism to reduce inequalities and foster a more just and equitable society.

The concept of equal opportunity has been translated into computable metrics with the same name (Hardt et al., 2016), as we will discuss later when we review fairness metrics. From a machine learning perspective, the technical drawback is that metrics for equal opportunity require label annotations for true positive outcomes, which are not always available after the deployment of a model (Feng et al., 2023).

2.3.1.2 Socialism and Equal Outcome

Socialism is a political and economic ideology that advocates for collective or government ownership of the means of production, distribution, and exchange [1]. Socialism has been shaped by numerous influential thinkers throughout its history, such as Marx and Engels (2019); Engels (2023).

The basic intuition is that it aims to reduce economic inequalities and promote social equality by redistributing wealth and resources from the affluent to the less privileged. While socialism does not necessarily mandate equal outcomes for all individuals, it does aim to reduce extreme disparities in wealth and income, fostering a more equitable distribution of resources. This makes socialism inherently concerned with reducing the gap in outcomes between different socio-economic groups.

In this context, **equal outcome**, is the concept of ensuring that every individual in a society attains the same or very similar results in terms of wealth, income, and social well-being. While some forms of socialism aspire to achieve more equal outcomes, it's not an inherent feature of all socialist systems. The pursuit of equal outcomes can be seen as a more extreme or utopian vision, as it often requires substantial intervention and redistribution of resources. Many socialist systems seek to balance the principles of equality and individual merit, providing a safety net and essential services while still allowing for varying degrees of success and personal achievement.

From a technical standpoint, within the framework of this PhD, we translate socialism to be construed as the aspiration for all individuals to attain the same outcome, irrespective of their individual efforts and that factors beyond an individual's control

[1] Socialism, throughout its history, has featured a multitude of perspectives and theories, frequently characterized by disparities in their conceptual, empirical, and normative principle. For example, Rappoport (1924) in his book of dictionary of socialism distinguished ~ 40 definitions of socialism.

should not be translated into social inequalities. This conceptualization is operationalized by employing a random estimator, a tool that assigns outcomes without regard to any of the input. This strategic use of a random estimator not only aligns with socialist principles but also facilitates a meaningful comparison with other fairness metrics.

This concept of a random estimator, as a means to uphold the principles of socialism, has also been articulated as a form of bias preservation in a study by Wachter et al. (2020).

2.3.1.3 Liberalism and Equal Treatment

Liberalism is a political and philosophical tradition that places a strong emphasis on individual rights, personal freedom, and limited government intervention in the lives of citizens (Friedman et al., 1990; Nozick, 1974). Liberals argue that individuals should be treated based on their merits and character rather than arbitrary factors such as race, gender, or socioeconomic status[2].

Liberalism pursues the idea that societies should strive to eliminate discriminatory practices and institutional barriers that prevent certain individuals or groups from enjoying the same rights and opportunities as others. This commitment to equal treatment is often reflected in liberal democracies through anti-discrimination laws, anti-affirmative action policies, and efforts to ensure equal access to education and employment.

Equal treatment has also been defined as "equal treatment-as-blindness" or neutrality (Sunstein, 1992; Miller and Howell, 1959). This expression is often used in the context of policies or practices that aim to ensure fairness and impartiality by treating individuals without regard to certain characteristics, such as race, gender, or other potential sources of bias. This concept is closely related to the idea of treating everyone with equality, regardless of their inherent attributes or background.

When it is said that equal treatment is "blind" to certain characteristics, it means that those characteristics should not influence how individuals are treated in a particular context [3]. The goal is to create a system or environment where decisions and actions are based on merit, qualifications, or relevant criteria rather than on factors that could

[2]We use the term *liberalism* to refer to the perspective exemplified by Friedman et al. (1990). This perspective can also be referred to as *neoliberalism* or *libertarianism* (Lamont and Favor, 2017; Friedman, 2022).

[3]Rawls (1971) introduces the concept of the "veil of ignorance." The veil of ignorance is a thought experiment suggesting that to achieve fairness and justice, individuals should design societal structures without knowing their own characteristics, such as their race, gender, or socioeconomic status. In this thesis, we understand that it also implies proxy discrimination, discrimination based on features that correlate with the protected attribute. This concept aligns with the idea of making decisions "blind" to certain characteristics to ensure impartiality and fairness.

lead to unfair discrimination. Achieving completely equal treatment has been argued to be a social desiderata that is impossible to achieve.

2.3.2 Moral Philosophy Notions

Moral philosophy explores what constitutes ethical behaviour and has given rise to established frameworks, two prominent among them being deontology and utilitarianism. Ethical considerations are fundamental to how individuals and societies make decisions, in this thesis we limit our review to these two perspectives.

2.3.2.1 Utilitarianism

Utilitarianism is one of the most powerful and persuasive approaches to normative ethics in the history of philosophy [4]. Utilitarianism, traditionally understood through the likes of Jeremy Bentham or John Stuart Mill, focuses on the consequences of an action to determine its ethical value, considering the utility or benefit brought about by the action (Macleod, 2020; Mill, 1966; Bentham, 1996). In traditional utilitarian accounts, utility is identified as pleasure, well-being, or happiness (Driver, 2022). The goal of moral action is therefore generally construed as increasing the amount of pleasure or happiness in our communities, with various proposed methods to measure these outcomes.

The Principle of Utility. At the core of utilitarianism lies the principle of utility, which serves as the guiding criterion for evaluating actions. This principle states:

"*the greatest amount of good for the greatest number*"[5].

Even if the sentence can provide a synthetic summary of utilitarianism, it is not perfectly accurate. While an action may enhance happiness for the majority (the greatest number of people), it can fall short of maximizing the overall good if the smaller group, whose happiness remains unchanged or decreases, experiences significant losses that outweigh the gains of the larger group. The principle of utility prohibits sacrificing the well-being of the smaller group for the benefit of the greater number unless the net increase in overall good surpasses any available alternative. The ethical worth of an

[4]Consequentialism is a broader ethical theory that asserts that the morality of an action is determined solely by the outcomes they produce, considering the overall good or bad resulting from those consequences. Utilitarianism, in the other hand, is a form of consequentialism where the right action is understood entirely in terms of consequences produced. (Sinnott-Armstrong, 2023)

[5]This sentence is a concise summary of the principle of utility. While it's not attributed to a specific individual, the concept is closely associated with utilitarian philosophers such as Jeremy Bentham and John Stuart Mill. Bentham (1996), in Chapter 1, titled "Of the Principle of Utility," Bentham discusses the foundation of his utilitarian philosophy. While the exact wording varies, the essence of the principle aligns with the idea of maximizing happiness or pleasure for the greatest number of people.

action is contingent upon its contribution to the general well-being and happiness of the affected individuals. One key aspect of the principle of utility is its emphasis on impartiality. It considers the well-being of all individuals equally, without giving special preference to oneself or a particular group. This universality contributes to the idea that the morally right action is the one that maximizes collective happiness.

Consequentialist Evaluation. Utilitarianism prompts us to assess actions based on their consequences. The moral calculus involves weighing the positive and negative outcomes, aiming for the greatest net happiness. This principle requires assessing not just the increase in happiness for the majority but also considering potential sacrifices made by a smaller group (Sinnott-Armstrong, 2023). Classic utilitarianism is consequentialist, denying that moral rightness depends on anything other than consequences. It contends that breaking promises or considering the circumstances of an act is only relevant insofar as they affect future consequences. Despite its seeming simplicity, classic utilitarianism comprises various interconnected claims, ranging from its consequentialist nature to specific views on evaluating consequences and considering the well-being of all individuals equally[6].

Guided by the principle of utility, this ethical framework seeks to maximize overall happiness, embracing a consequentialist evaluation that adapts to the nuances of individual situations.

2.3.2.2 Deontology

Deontological ethics, with Immanuel Kant as its central figure and in particular through his seminal work *Groundwork of the Metaphysics of Morals* Kant (1785), emphasizes that it is through duties, or universal set of rules, that the ethical value of actions is determined. Generally, deontology maintains that these duties flow from the recognition that all humans are intrinsically valuable—it is our rational nature that gives us reason to act in one way or another. This duty, or rule-based, moral theory thus aligns well with various declarations, such as the *Universal Declaration of Human Rights*, that recognize a set of inalienable rights based on the individual, such as the right to life and the right to equal treatment under the law; with deontology, moral action involves

[6]Classic utilitarianism, while seemingly straightforward in its emphasis on consequences, is, in fact, a complex amalgamation of distinct claims about the moral rightness of actions. These encompass various dimensions, from the overarching consequentialist stance to specific perspectives on evaluating consequences. Classic utilitarianism includes actual consequentialism, direct consequentialism, evaluative consequentialism, hedonism, maximizing consequentialism, aggregative consequentialism, total consequentialism, universal consequentialism, equal consideration, and agent-neutrality. Each of these facets adds layers of intricacy to the theory, demonstrating that classic utilitarianism entails a multifaceted framework that goes beyond a simplistic focus on consequences(Sinnott-Armstrong, 2023).

respecting inescapable duties that lay the fundamental foundation of the rights of others. While not all duties entail rights, all rights suppose a corresponding duty (Watson, 2021).

2.3.2.3 Kantian Critique of Utilitarianism

After providing a concise outline of Kantian deontology and utilitarianism, we now consider the Kantian critique of utilitarianism[7].

Intrinsic value and the pursuit of the 'good'. Utilitarianism believes that all moral actions ought to be aimed at obtaining the 'good' as an impersonal absolute value, such as happiness or pleasure, that can be maximized. Peter Singer writes that the classical utilitarian approach,

"*regards sentient beings as valuable only insofar as they make possible the existence of intrinsically valuable experiences like pleasure.*" (Singer, 2011).

Singer, himself a utilitarian, continues this notion that it is not the individuals that matter but the defined utility through the replaceability argument. While he appreciates that all individuals have inherent worth, this value does not extend far enough to guarantee its protection. Singer has thus argued that there are circumstances (albeit rare) where killing humans is the right thing to do if it brings about more happiness in the community Singer (2011).

This, of course, contrasts the second formula of the categorical imperative that stresses that each individual is intrinsically valuable and is not instrumentally subordinated to the well-being, happiness, or pleasure of others. An individual cannot be replaced by another individual even if their existence brings about more happiness. This is because Kantians disagree that the 'good' can be understood as something abstract and impersonal, something that is simply just 'good'. Rather, the good is a relational concept because, as Korsgaard argues, all value is tethered. Happiness or pleasure matters because they are *good-for* individuals (Korsgaard, 2018). Therefore, crucially, something can only be considered as universally good when it is *good-for* everyone, from every individual's perspective.

The issue of aggregation. Because utilitarians do not consider individuals intrinsically valuable, the moral decision procedure thus allows aggregating the good. There is no prohibition against using individuals, even if they are innocent, as mere means, as

[7]Kant, however, did not directly address the utilitarian of his time, Jeremy Bentham, so it is helpful to look to contemporary Kantians such as Onora O'Neill and Christine Korsgaard, who provide a direct critique of utilitarian ethics. The following provides a foundational, but certainly not exhaustive, list of critiques.

long as the loss of their dignity is for maximizing a sufficient amount of "good" in the community. Yet, this aggregation not only may violate the second formula of the categorical imperative but also consider the tethered notion of the "good". As Korsgaard contends:

"*what is good-for me plus what is good-for you is not necessarily good-for* anyone. (page 157 Korsgaard (2018))"

There is a problem when maximizing the good involves taking something away from an individual: if giving something to person A brings him more pleasure than it would to person B, the utilitarian would say we ought to give it to person A. But this action would just be good for person A, and it would be worse for person B. Because it is tethered, the good is not something we can aggregate (Korsgaard, 2018). This particularly is more concerning when those on the receiving end of aggregation are already marginalized individuals.

Optimizing outcomes. The utilitarian approach involves calculating outcomes, whether it be maximizing benefit or reducing harm, and is not a robust approach to identifying which actions should be prohibited. As Onora O'Neill points out, circumstances can change the outcome. Take the act of lying as an example: typically, lying is considered to be immoral and, thus, generally prohibited. Yet not all lies, such as 'white lies', cause harm. The same goes for telling the truth, which often provides a benefit, but it can also inflict emotional harm. Focusing on consequences is unpredictable and, therefore, provides a less robust standard to measure the rightness and wrongness of actions. The link between harmful consequences and actions is unstable. The connection between norms, or duties, and actions is a more robust approach to morally evaluating actions. Duties, such as the duties of civility or to respect the freedom of others, provide a more useful and robust way of evaluating actions(O'Neill, 2022).

Utilitarianism thus has a limitless scope of acceptable actions as it does not depend on the type of action but rather the consequence, or utility, it brings about. Being dishonest for the utilitarian, for example, is not *categorically* condemned; for this reason, it lacks precision as the rightness or wrongness of an action can only be determined until after it is carried out and the consequences are known. The Kantian approach, while it has a more restricted scope given its categorical stance on duties, provides a precise and fundamental standard of action, where moral authority persists irrespective of any potential harmful consequence. Being dishonest (to some Kantians) is always wrong irrespective of unforeseen consequences.

Later, in the discussion section of AI alignment, we will map the Kantian criticism of current utilitarianism to our proposed Equal Treatment.

2.4 Technical Types of Fairness

In this section, we compare our proposed notion and model for equal treatment with respect to other metrics found in the literature. In alignment with regulatory frameworks like U.S. Title VII of the Civil Rights Act of 1964, prohibiting discrimination in employment based on race, sex, and other protected characteristics (Commission, 1964) and prior work by Kuppler et al. (2021); we distinguish between two discrimination types: *disparate treatment*, aiming to prevent intentional model discrimination, and *disparate impact*, addressing unintentional effects across subpopulations. The former examines the system's intentions, as realized in model predictions $f_\theta(X)$, while the latter considers effects, interpreted as the error $f_\theta(X) - Y$.

2.4.1 Disparate Impact Fairness Metrics: Approaches that Rely on Labeled Data

2.4.1.1 Equal Opportunity

Two formulations of the Equal Opportunity fairness metric are defined in the literature. First, it is the difference in the True Positive Rate (TPR) between the protected group and the reference group Hardt et al. (2016); Podesta et al. (2014); Munoz et al. (2016):

$$\text{TPR} = \frac{TP}{TP + FN}$$
$$\text{EOF}_z = \text{TPR}_z - \text{TPR}$$

Second – a formulation proposed by Heidari et al. (2019) and that does not rely on the false negatives – a model f_θ achieves equal opportunity if $P(f_\theta(X) = 1|Y = 1, Z = z) = P(f_\theta(X) = 1|Y = 1)$.

Equal Opportunity comes with a technical limitation, as it necessitates the availability of labeled data for true positives. For instance, in a loan lending scenario, true positives represent users who are granted a loan and successfully repay it. Obtaining such data can be time-consuming, requiring the completion of the entire loan cycle. This reliance on true positive rates in formulating Equal Opportunity presents an additional challenge involving the calculation of False Negatives. In the context of the loan scenario, a False Negative occurs when a loan is denied to someone capable of repayment, but acquiring this data may not be feasible. An alternative approach involves maintaining a holdout set of randomly selected users to whom loans are uniformly granted, allowing for monitoring in that controlled environment. Nevertheless, the implementation of a holdout set comes with potential economic and societal implications. Careful consideration is needed, as this approach may impact both financial outcomes and broader societal dynamics.

2.4.1.2 Treatment Equality

Despite its similar name, Treatment Equality does not necessarily align with the same distributive justice values as equal treatment, and the mathematical metric it employs is notably distinct (Berk et al., 2021).

In the original formulation by Berk et al. (2021), the concept is defined as follows:

> *Treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories. The term "treatment" is used to convey that such ratios can be a policy lever with which to achieve other kinds of fairness. For example, if false negatives are taken to be more costly for men than women so that conditional procedure accuracy equality can be achieved, men and women are being treated differently by the algorithm. Incorrectly classifying a failure on parole as a success, say, is a bigger mistake for men. The relative numbers of false negatives and false positives across protected group categories also can by itself be viewed as a problem in criminal justice risk assessments addresses similar issues, but through the false negative rate and the false positive rate: our b/(a+b) and c/(c + d), respectively.*

$$\frac{FP_z}{FN_z} = \frac{FP}{FN}$$

The authors do not specify the principle or value with which the notion should align, and this clarity is absent from their definition. An illustrative sentence is "*men and women are being treated differently by the algorithm*". However, this statement may not be entirely accurate, as the algorithm may treat men and women equally while exhibiting different error rates. For instance, a constant classifier that grants loans to everyone treats everyone equally. Still, if males and females have distinct patterns of loan repayment, the classifier will yield different errors, resulting in distinct $\frac{FP}{FN}$ values.

2.4.2 Disparate Treatment Fairness Metrics: Approaches that Determine Unfairness without Labelled Data

In this section we compare against metrics that rely only on X and f_θ rather than on disparities of model errors.

2.4.2.1 Fairness Through Unawareness.

This bias mitigation method was initially proposed by Grgic-Hlaca et al. (2018) and is often used as a baseline.

Definition 2.1. *(Fairness Through Unawareness.* An algorithm is fair if no protected attributes Z is explicitly used as an input of the algorithm.

2.4.2.2 Demographic Parity.

Definition 2.2. *(Demographic Parity (DP)).* A model f_θ achieves demographic parity if $f_\theta(X) \perp Z$.

Thus, demographic parity holds if $\forall z.\, P(f_\theta(X)|Z = z) = P(f_\theta(X))$. For binary Z's, we can derive an unfairness metric as $d(P(f_\theta(X)|Z = 1), P(f_\theta(X))$, where $d(\cdot)$ is a distance between probability distributions.

A refined metric introduced by legal scholars aiming to align with European Court of Justice "gold standard" (Wachter et al., 2021) is Conditional Demographic Parity, that extends Demographic Parity by fixing one or more attributes.

Definition 2.3. *(Conditional Demographic Parity (CDP)).* A model f_θ achieves conditional demographic parity if $P(f_\theta(X)|X_i = \tau, Z = z) = P(f_\theta(X)|X_i = \tau)$, where τ is any fixed value.

CDP differs from DP only in the sense that one or more additional covariate conditions are added.

2.4.2.3 Individual Fairness.

Individual fairness principle is that any two individuals who are similar concerning a particular task should be classified similarly, a topic with a broad literature on fairness, notably in social choice theory, game theory, economics, and law as well, for example, John Rawls' in *Justice as Fairness* states"citizens who are similarly endowed and motivated should have similar opportunities to hold office" Rawls (1958). Within the computer science field, to accomplish this individual-based fairness, a distance metric defines the similarity between the individuals. Several individual fairness metrics have been proposed, such as the Lipschitz condition or the earth-mover distance between two distributions Dwork et al. (2012a).

We say that a model f achieves individual fairness if similar individuals received similar predictions:

$$\forall X, X'\, d(f(X), f(X')) \leq \mathcal{L} \cdot d(X, X')$$

Where $\mathcal{L} > 0$ is a constant and the two $d(\cdot, \cdot)$'s are distance functiosn between models' outputs and between instances respectively (Dwork et al., 2012a).

2.4.2.4 Counterfactual Fairness.

Counterfactual fairness, as defined by Kusner et al. (2017) captures the intuition that a decision is fair towards an individual if it is the same in *(i)* the actual case and *(ii)* in a counterfactual case where the individual belonged to a different protected attribute group.

Definition 2.4. *Counterfactual Fairness* We say that the model f_θ is counterfactually fair if:

$$\forall X \in X \, \forall z, z' \in Z \; P(f_\theta(X_{Z=z})|X = X, Z = z) = P(f_\theta(X_{Z=z'})|X = X, Z = z)$$

where $X_{Z=z}$ and $X_{Z=z'}$ are the instances X in the (counterfactual) worlds where $Z = z$ and $Z = z'$ respectively.

2.4.2.5 Causal Feature Selection

Another avenue of research in defining fairness operates within the causal fairness paradigm (Salimi et al., 2019). The following definition proposes a method for identifying a subset of features that ensures fairness in the data by conducting conditional independence tests between various feature subsets.

Definition 2.5. *(Causal Fairness).* For a given set of admissible variables, A, a model f_θ is considered fair if $P(f_\theta(X) = y|\texttt{do}(S) = s', \texttt{do}(A = a) = P(f_\theta(X) = y|\texttt{do}(S) = s, \texttt{do}(A = a)$ for all values of A,S.

Here, the `do-operator` is defined as an intervention or modification of initial attributes to a specific value, observing its effect (Pearl, 2009).

TABLE 2.2: Summary of Technical Definitions of Fairness Metrics

Fairness Metric	**Definition**
Equal Opportunity (EOF)	$\text{EOF}_z = \text{TPR}_z - \text{TPR}$ $P(f_\theta(X) = 1 \mid Y = 1, Z = z) = P(f_\theta(X) = 1 \mid Y = 1)$
Treatment Equality	${}^{FP_z}/_{FN_z} = {}^{FP}/_{FN}$
Fairness Through Unawareness	An algorithm is fair if no protected attributes Z is explicitly used as an input of the algorithm $f : X \to Y$ for which $Z \notin X$
Demographic Parity (DP)	$f_\theta(X) \perp Z$
Conditional Demographic Parity (CDP)	$P(f_\theta(X) \mid X_i = \tau, Z = z) = P(f_\theta(X) \mid X_i = \tau)$
Individual Fairness	$\forall X, X' \; d(f(X), f(X')) \leq \mathcal{L} \cdot d(X, X')$
Counterfactual Fairness	$\forall X \in X \; \forall z, z' \in Z$ $P(f_\theta(X_{Z=z}) \mid X = X, Z = z)$ $= P(f_\theta(X_{Z=z'}) \mid X = X, Z = z)$
Causal Fairness	For a given set of admissible variables, A, model f_θ is fair if $P(f_\theta(X) = y \mid \texttt{do}(S) = s', \texttt{do}(A = a)$ $= P(f_\theta(X) = y \mid \texttt{do}(S) = s, \texttt{do}(A = a)$ $\forall$ values of A, S

2.5 Model Monitoring

Model monitoring techniques help to detect unwanted changes in the performance of machine learning applications in a production environment. One of the biggest challenges in model monitoring is distribution shift, which is also one of the primary sources of model degradation (Quiñonero-Candela et al., 2009; Diethe et al., 2019).

Diverse types of model monitoring scenarios require different supervision techniques. We can distinguish two main groups: Supervised learning and unsupervised learning. Supervised monitoring, is the appealing one since the label is available and performance metrics can easily be tracked. Acquiring labelled data post-deployment poses a considerable challenge and is often impractical, exhibiting well-known biases, such as confounding, selection, and missingness (Feng et al., 2023).[8] While attractive, these techniques are often unfeasible as they rely either on having ground truth labelled data available or maintaining a hold outset, which leaves the challenge of how to monitor ML models to the realm of unsupervised learning (Diethe et al., 2019). Popular unsupervised methods that are used in this respect are the Population Stability Index (PSI) and the Kolmogorov-Smirnov test (K-S), all of which measure how much the distribution of the covariates in the new samples differs from the covariate distribution within the training samples (Rabanser et al., 2019). These methods are often limited to real-valued data, low dimensions, and require certain probabilistic assumptions (Diethe et al., 2019; Malinin et al., 2021c).

Monitoring machine learning models in production is not an easy task. There are situations when the true label of the deployment data is available, and performance metrics can be monitored. However, there are cases where it is not, and performance metrics are not so trivial to calculate once the model has been deployed. Model monitoring ensures that a machine learning application in a production environment displays consistent behavior over time.

Being able to explain or remain accountable for the performance or the deterioration of a deployed model is crucial, as a drop-in model performance can affect the whole business process (Mougan et al., 2021a), potentially having catastrophic consequences[9]. Once a deployed model has deteriorated, models are retrained using previous and new input data in order to maintain high performance. This process is called continual learning (Diethe et al., 2019), and it can be computationally expensive and put high demands on the software engineering system. Deciding when to retrain machine learning models is paramount in many situations (Mougan et al., 2021a).

[8] *Confounding Bias:* Misleading results due to an unconsidered external variable. *Selection Bias:* Skewed results from non-representative data samples. *Missingness Bias:* Distortion from systematically absent data (Good and Hardin, 2012).

[9] The Zillow case is an example of consequences of model performance degradation in an unsupervised monitoring scenario, see `https://edition.cnn.com/2021/11/09/tech/zillow-ibuying-home-zestimate/index.html` (Online accessed January 26, 2022).

2.5.1 Distribution Shift Formalization

The core machine learning assumption is that training data $\mathcal{D}^{tr}$ and novel data $\mathcal{D}^{new}$ are sampled from the same underlying distribution $\mathbf{P}(X, Y)$. The twin problems of *model monitoring* and recognizing that new data is *out-of-distribution* can now be described as predicting an absolute or relative performance drop between $\texttt{perf}(\mathcal{D}^{tr})$ and $\texttt{perf}(\mathcal{D}^{new})$, where $\texttt{perf}(\mathcal{D}) = \sum_{(x,y)\in\mathcal{D}} \ell_{\text{eval}}(f_\theta(x), y)$, ℓ_{eval} is a metric like 0-1-loss (accuracy), but $\mathcal{D}_Y^{new}$ is unknown and cannot be used for such judgment.

Therefore related work analyses distribution shifts between training and newly occurring data. Let two datasets $\mathcal{D}, \mathcal{D}'$ define two empirical distributions $\mathbf{P}(\mathcal{D}), \mathbf{P}(\mathcal{D}')$, then we write $\mathbf{P}(\mathcal{D}) \nsim \mathbf{P}(\mathcal{D}')$ to express that $\mathbf{P}(\mathcal{D})$ is sampled from a different underlying distribution than $\mathbf{P}(\mathcal{D}')$ with high probability $p > 1 - \epsilon$ allowing us to formalize various types of distribution shifts.

2.5.2 Types of Distribution Shift

This section aims to provide an illustrative introduction to distribution shift; for more in-depth surveys see Quiñonero-Candela et al. (2009).

2.5.2.1 Covariate Shift

The most classical and straightforward type of distribution shift is covariate shift, where the distribution from the training/source set $P(\mathcal{D}_X^{tr})$ differs from the test $P(\mathcal{D}_X^{te})$, meaning that the input distribution changes, but the conditional probability of a label given an input remains the same. Shimodaira (2000); Sugiyama et al. (2007); Masashi and Klaus-Robert (2005)

Definition 2.6 (Univariate data shift)**.** There is a univariate data shift between $\mathbf{P}(\mathcal{D}_X^{tr}) = \mathbf{P}(\mathcal{D}_{X_1}^{tr}, \ldots, \mathcal{D}_{X_p}^{tr})$ and $\mathbf{P}(\mathcal{D}_X^{new}) = \mathbf{P}(\mathcal{D}_{X_1}^{new}, \ldots, \mathcal{D}_{X_p}^{new})$, if $\exists i \in \{1 \ldots p\} : \mathbf{P}(\mathcal{D}_{X_i}^{tr}) \nsim \mathbf{P}(\mathcal{D}_{X_i}^{new})$.

Definition 2.7 (Covariate data shift)**.** There is a covariate data shift between $P(\mathcal{D}_X^{tr}) = \mathbf{P}(\mathcal{D}_{X_1}^{tr}, \ldots, \mathcal{D}_{X_p}^{tr})$ and $\mathbf{P}(\mathcal{D}_X^{new}) = \mathbf{P}(\mathcal{D}_{X_1}^{new}, \ldots, \mathcal{D}_{X_p}^{new})$ if $\mathbf{P}(\mathcal{D}_X^{tr}) \nsim \mathbf{P}(\mathcal{D}_X^{new})$, which cannot only be caused by univariate shift.

An example of this could be training a machine learning model to predict the gender of a person based on physiological characteristics. During the last century, population height and weight distribution have changed Max Roser and Ritchie (2013), but the proportion of males/females is not affected by this temporal evolution. It is important

to note that there is no causal relationship between our source data $P(\mathcal{D}_X^{tr})$ and our test data $P(\mathcal{D}_X^{te})$ at this stage.[10]

Another example of this distribution shift could be when one wants to build a machine-learning model to detect whether someone has COVID-19 by their coughing sound. In the data collection phase, recordings are collected from the hospitals. These recordings are recorded in controlled conditions: silent rooms with clear starting and ending times. However, when you deploy the model and provide the service to the population (for example, as a phone application), users can test coughing directly with the phone app microphones, inducing a covariate bias in the test data. The starting point could be before or in the middle of the cough. Also, the coughing can have some background noise due to not being performed in laboratory conditions.

Both of the above examples, might create a distribution shift that will jeopardize the model's performance.

2.5.2.2 Predictions or Label Shift

Only the prior class probabilities change between the training and test sets. In this case, this second type of distribution shift occurs when only the class probabilities $P(\mathcal{D}_Y^{te}) \neq P(\mathcal{D}_Y^{te})$ changes and everything else prevails $P(\mathcal{D}_X^{tr}) = P(\mathcal{D}_X^{te})$ Tasche (2017).This poses a risk towards monitoring the performance of a model since unwanted changes in the real test target distribution will make unfeasible performance metrics.

Definition 2.8 (Predictions Shift)**.** There is a predictions shift between distributions $\mathbf{P}(\mathcal{D}_X^{tr})$ and $\mathbf{P}(\mathcal{D}_X^{new})$ related to model f_θ if $\mathbf{P}(f_\theta(\mathcal{D}_X^{tr})) \nsim \mathbf{P}(f_\theta(\mathcal{D}_X^{new}))$.

An example of this shift could be economic inflation, where buying the same set of products (e.g., gas, tomatoes, and potatoes) in 2001 will cost a very different price than in 2022.

2.5.2.3 Concept or Posterior Shift

This type of shift is characterized by input data and predictions distribution that remains the same, and what changes is the conditional distribution of the output given an input changes:

[10]There can be confounding factors in the modelling or even statistical assumptions that do not hold in a real case scenario. But, for the sake of explanation, examples are overly simplistic

Definition 2.9 (Concept Shift)**.** There is a concept shift between $\mathbf{P}(\mathcal{D}^{tr}) = P(\mathcal{D}_X^{tr}, \mathcal{D}_Y^{tr})$ and $\mathbf{P}(\mathcal{D}^{new}) = P(\mathcal{D}_X^{new}, \mathcal{D}_Y^{new})$ if conditional distributions change, i.e. $\frac{\mathbf{P}(\mathcal{D}_Y^{tr})}{\mathbf{P}(\mathcal{D}_X^{tr})} \nsim \frac{\mathbf{P}(\mathcal{D}^{new})_Y}{\mathbf{P}(\mathcal{D}_X^{new})}$.

An example of this is the Zillow case, an example of the consequences of model performance degradation due to concept drift. The price of an apartment in the city centre before COVID-19 could have drastically dropped during/after COVID-19 due to people living in the city centre, even if the apartment stayed the same.

2.5.3 Model Monitoring Limitation

This thesis addresses a common scenario in machine learning applications, characterized by a specific setup: during training, both covariates, denoted as $\mathcal{D}_X^{tr}$, and labels, $\mathcal{D}_Y^{tr}$, are available. However, only covariates, $\mathcal{D}_X^{te}$, are known at the testing stage. This situation, where practitioners possess a labelled training set but must deal with unlabeled data at deployment, complicates model monitoring and evaluation.

In such a setup, detecting covariate shifts—changes in input data distribution—is theoretically and practically achievable, as evidenced by existing literature (Ramdas et al., 2015; Rabanser et al., 2019). Conversely, identifying concept shifts, which refer to changes in the relationship between the input data and the target labels, inherently depends on access to labelled data.

Consequently, accurately estimating the degradation of a model's performance when transitioning from the source data, $\texttt{perf}(\mathcal{D}^{tr})$, to new, unseen data, $\texttt{perf}(\mathcal{D}^{new})$, is recognized as an infeasible task without introducing additional assumptions or characterizing concept shift (Garg et al., 2022).

Chapter 3

Measuring Alignment with Equal Treament

In philosophy, debates surrounding the principles of a fair political system and the essence of ethical decision-making have given rise to well-defined frameworks addressing distributive justice (Lamont and Favor, 2017; Kymlicka, 2002) and moral decision-making processes (Sinnott-Armstrong, 2023; Kant, 1785)

The *liberalism* school of thought[1], put forward by scholars such as Friedman et al. (1990) and Nozick (1974), argues for *equal treatment* of individuals regardless of their protected characteristics. On the other hand, deontology, grounded in the philosophy of Immanuel Kant, emphasizes that actions must be judged based on their adherence to a set of moral principles or duties, regardless of the consequences.

Verdicts like the recent Supreme Court decision about college admissions,[2] may suggest that a proper formalization of *equal treatment* is required in software supporting these processes.

By analyzing the college admission use case[3] in Section 3.1.1 and the paper blind reviews case in Section 3.1.2, we observe that existing approaches fail to meet expected requirements. For example, in college admissions, *fairness-by-unawareness* (Grgic-Hlaca et al., 2018) fails to recognize discrimination through proxy variables, and *individual fairness* (Dwork et al., 2012a) cannot identify group discrimination. *Equal opportunity* Hardt et al. (2016) and *causal fairness* Pearl (2009) are good candidates for *equal treatment*, but require labelled data and background knowledge, respectively, neither of which is available in most practical scenarios. *Equal treatment* has been translated by researchers

[1] We use the term *liberalism* to refer to the perspective exemplified by Friedman et al. (1990). This perspective can also be referred to as *neoliberalism* or *libertarianism* (Lamont and Favor, 2017; Friedman, 2022).

[2] `https://www.jdsupra.com/legalnews/the-impact-of-the-supreme-court-s-7330075/`

[3] `https://en.wikipedia.org/wiki/Students_for_Fair_Admissions_v._Harvard`

(Simons et al., 2021; Heidari et al., 2019; Wachter et al., 2020) into Demographic Parity (DP), also called Statistical Parity, which compares the distributions of predicted outcomes of a model f for different social groups. However, we show that Demographic Parity may indicate fairness, even when groups are discriminated against according to the *equal treatment* principle.

We propose the novel fairness notion of Equal Treatment[4] for supervised machine learning models, which satisfies the requirements for a formalization of *equal treatment*.

The notion of Equal Treatment considers the contribution of non-protected features to the output of the machine learning model as explained by Shapley values. If two social groups are treated the same, the distributions of feature contributions, which we call *explanation distributions*, will not be distinguishable. Thus, Equal Treatment tests independence of Shapley values with the protected feature. We introduce a decision tool, the "Equal Treatment Inspector", that implements this idea. When detecting unequal treatment, it explains the features involved in such inequality, supporting the understanding of the roots of un-equal treatment in the machine learning model.

In summary, this chapter contributions are:

1. The definition of Equal Treatment, based on explanation distributions, as a formalization of *equal treatment*.
2. The definition of an "Equal Treatment Inspector" workflow, based on a classifier two sample test, for recognizing and explaining un-equal treatment.
3. The study of the formal relationships between Demographic Parity and Equal Treatment.
4. Extensive experiments, both on synthetic and natural datasets, to demonstrate our method and compare it with related work.
5. An open-source Python package `explanationspace` implementing the "Equal Treatment Inspector", which is `scikit-learn` compatible, and includes documentation and tutorials.

[4]We distinguish the philosophical principle of "*equal treatment*" from our proposed fairness notion, 'Equal Treatment" (ET), by printing the first in italics and capitalizing the initials of the latter.

3.1 Illustrative Examples

3.1.1 College Admissions

To illustrate the difference between equal opportunity, equal outcomes, and equal treatment, we consider the hypothetical use case, the admission process of a university:

In June 2023, the U.S. Supreme Court struck down race-conscious admission programs at some universities (Killenbeck and Killenbeck, 2022)[5], ruling these as discriminatory against certain racial groups. This decision marked a significant shift from previous policies that used race as one of several factors in a holistic admissions process to promote a diverse student body[6]. While the ruling does not completely forbid universities from considering applicants' racial features, it emphasizes a more colorblind approach to admissions. This move towards equal treatment focuses less on achieving equal outcomes (diverse representation) and more on not considering race as an admissions factor. Now we discuss these notions within the context of fair machine learning.

Equal Opportunity. In this approach, the university focuses on ensuring that students from different backgrounds have an equal chance of being admitted if they have high potential or qualifications. For instance, students from under-resourced high schools are given the same opportunity as those from well-funded schools if they demonstrate the same exceptional talent or potential. From a technical perspective, equal opportunity has been operationalized by the true positive rate. The university may implement this strategy by adjusting admission criteria based on educational opportunities tied to race. However, this can lead to challenges like overcompensating for disadvantages or unintentionally lowering standards for certain groups.

Equal Outcomes. This measure requires that the distribution of acceptance rates is similar, independently of the student (cf. Definition 2.2). This could mean setting targets or quotas to ensure representation from various groups, regardless of their individual academic credentials. For instance, if 20% of applicants are from a certain ethnicity, the university might aim to have 20% of their admits from that ethnicity as well. While this promotes a diverse student body, it can overlook individual merit and potentially lead to tension between equity and academic standards. Particularly given the recent ruling of the US Supreme Court, considering ethnicity to match the proportion of applicants from that ethnicity would likely be considered illegal[7]. It's also important to note that outcomes can have similar rates due to random chance.

Equal Treatment. In this scenario, the university ensures that every application is evaluated based on the same criteria, such as academic achievement, extracurricular involvement, and personal essays, without bias towards the applicant's background. This

[5] https://www.scotusblog.com/2023/06/supreme-court-strikes-down-affirmative-action-programs-in-college-admissions/
[6] https://www.politico.com/news/2023/06/29/supreme-court-ends-affirmative-action-in-college-admissions-00104179
[7] https://www.jdsupra.com/legalnews/the-impact-of-the-supreme-court-s-7330075/

means that a student's socioeconomic status, high school's reputation, or geographic location (protected attributes) would not influence their likelihood of admission. The challenge here is to ensure the following in the admissions process: *(R1) Group Discrimination* different decisions are made based on inherent or protected characteristics of individuals, such as race, gender, ethnicity, or other attributes that are present at birth and beyond an individual's choice or control; *(R2) Unlabeled Data*, focus on the disparate decisions, metrics should evaluate the decisions of the model instead of statistical differences on the errors; *(R3) No Background Knowledge*, metrics should not necessitate an understanding of causal or structural aspects, ensuring practical applicability in real-world scenarios; *(R4) Proxy discrimination*: metrics should be capable of detecting whether the model behaviour is genuinely free from discriminatory features, arising from proxy discrimination. This involves scrutinizing the differential impact that various features may have on the predictions, ensuring that no indirect bias is influencing the outcome; *(R5) Explanation Capabilities* focuses on the necessity for fairness metrics to not only detect biases or discriminations but also to explain them.

In the university admission case study, each approach to fairness in admissions – equal opportunity, equal outcomes, and equal treatment – has its merits and challenges. Equal opportunity aims to level the playing field based on potential, equal outcomes, strive for a representative student body, and equal treatment focuses on a uniform and unbiased evaluation process. We leave the normative discussion of which fairness paradigm should be pursued by policy to the discourse in the social sciences and the broad public. This case illustrates the complexity of implementing fair practices in college admissions, highlighting that there is no one-size-fits-all solution.

3.1.2 Paper Blind Reviews

To illustrate the difference between equal opportunity, equal outcomes, and equal treatment, based on the previously discussed framework and applied to AI, we consider the example of conference papers' blind reviews and focus on the protected attribute of the country of origin of the paper's author, comparing Germany and the United Kingdom.

For *equal opportunity*, we quantify fairness by the true positive rate. In other words, it is the acceptance ratio given that the quality of the paper is high. Achieving equal opportunity will imply that these ratios are similar between the two countries. In blind reviews, the purpose is to evaluate the paper's quality and the research's merit without being influenced by factors such as the author's identity, affiliations, background or country. If we were to enforce equal opportunity in this use case, we would aim for similar true positive rates for submissions from different countries. However, this approach could lead to unintended consequences, such as unintentionally favouring certain countries to meet some quota, overcorrection or quotas of affirmative action towards certain countries.

For *equal outcomes*, we require that the distribution of acceptance rates is similar, independently of the quality of the paper (cf. Definition 2.2). So that the ratio of accepted papers is similar for each country. Note that the outcomes can have similar rates due to random chance, even if there is a country bias in the acceptance procedure.

For *equal treatment*, we require that the contributions of the features used to make a decision on paper's acceptance has similar distributions (cf. Definition 3.4). Equality of treatment through blindness is more desirable than equal opportunity or equal outcomes because it ensures that all submissions are evaluated solely on the basis of their quality, without any bias or discrimination towards any particular country. By achieving equality of treatment through blindness, we can promote fairness and objectivity in the review process and ensure that all papers have an equal chance to be evaluated on their merits. We complement the previous use case recollecting the following requirements: *(R1) Group Discrimination* different decisions are made based on belonging to an institution,; *(R2) Unlabeled Data*, focus on the disparate decisions, metrics should evaluate the decisions of the model instead of statistical differences on the errors; *(R3) No Background Knowledge*, metrics should not necessitate an understanding of causal or structural aspects, ensuring practical applicability in real-world scenarios; *(R4) Proxy discrimination*: metrics should be capable of detecting whether the model behaviour is genuinely free from institutional discriminatory features, arising from proxy discrimination.*(R5) Explanation Capabilities* focuses on the necessity for fairness metrics to not only detect biases or discriminations but also to explain them.

3.2 Fairness Formalization

We assume a feature modelling protected social groups is denoted by Z, called *protected feature*, and assume it to be binary valued in the theoretical analysis. Z can either be included in the predictive features X used by a model or not. If not, we assume that it is still available for a test dataset. Even without the protected feature in training data, a model can discriminate against the protected groups by using correlated features as a proxy of the protected one (Pedreschi et al., 2008).

We write $A \perp B$ to denote statistical independence between the two sets of random variables A and B, or equivalently, between two multivariate proability distributions. We define two common fairness notions and corresponding fairness metrics that quantify a model's degree of discrimination or unfairness Mehrabi et al. (2021).

Definition 3.1. *(Demographic Parity (DP))*. A model f_θ achieves demographic parity if $f_\theta(X) \perp Z$

Thus, demographic parity holds if $\forall z.\, P(f_\theta(X)|Z = z) = P(f_\theta(X))$. For binary Z's, we can derive an unfairness metric as $d(P(f_\theta(X)|Z = 1), P(f_\theta(X))$, where $d(\cdot)$ is a distance between probability distributions.

Definition 3.2. *(Equal Opportunity (EO))* A model f_θ achieves equal opportunity if $\forall z.\, P(f_\theta(X)|Y = 1, Z = z) = P(f_\theta(X) = 1|Y = 1)$.

As before, we can measure unfairness for binary Z's as $d(P(f_\theta(X)|Y = 1, Z = 1), P(f_\theta(X) = 1|Y = 1))$. Equal opportunity comes with the problem that labels for correct outcomes are required.

3.3 A Model for Monitoring Equal Treatment

3.3.1 Formalizing Equal Treatment

To establish a criterion for equal treatment, we rely on the notion of explanation distributions.

Definition 3.3. *(Explanation Distribution)* An explanation function $\mathcal{S} : \mathcal{F} \times X \to \mathbb{R}^p$ maps a model $f_\theta \in \mathcal{F}$ and an input instance $x \in X$ into a vector of reals $\mathcal{S}(f_\theta, x) \in \mathbb{R}^p$. We extend it by mapping an input distribution $\mathcal{D}_X$ into an (empirical) *explanation distribution* as follows: $\mathcal{S}(f_\theta, \mathcal{D}_X) = \{\mathcal{S}(f_\theta, x) \mid x \in \mathcal{D}_X\} \subseteq \mathbb{R}^p$.

We will use Shapley values as an explanation function (cf. Appendix 2.2). We introduce next the new fairness notion of Equal Treatment, which considers independence over the explanations of the model's outputs.

Definition 3.4. *(Equal Treatment (ET))*. A model f_θ achieves equal treatment if $\mathcal{S}(f_\theta, X) \perp Z$.

As we will see later in Section 3.4, Equal Treatment is a stronger definition than Demographic Parity since it not only requires that the distributions of the predictions are similar but that the process of how predictions are made is also similar.

3.3.2 Equal Treatment Inspector

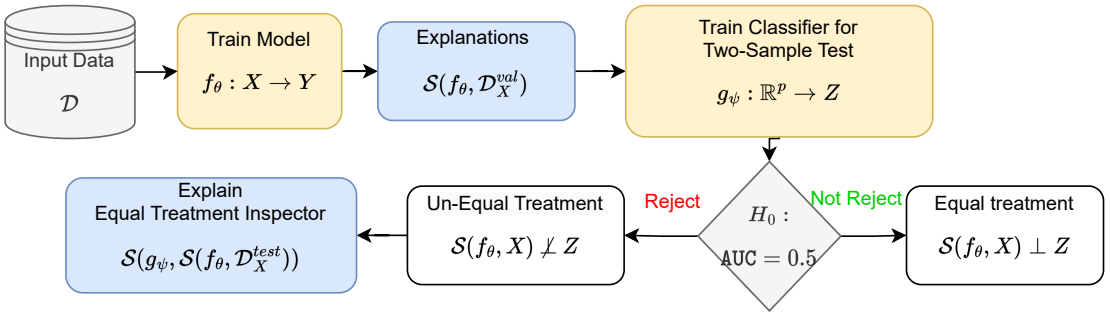

FIGURE 3.1: Equal Treatment Inspector workflow. The model f_θ is learned based on training data, $\mathcal{D} = \{(x_i, y_i)\}$, and outputs the explanations $\mathcal{S}(f_\theta, \mathcal{D}_X)$. The Classifier Two-Sample Test receives the explanations to predict the protected attribute, Z. The AUC of the two-sample test classifier g_ψ decides for or against *equal treatment*. We can interpret the driver for unequal treatment on g_ψ with explainable AI techniques

Our approach is based on the properties of the Shapley values (cf. Appendix 2.2) and the prior work of Lopez-Paz and Oquab (2017) of classifier two-sample tests. We split the available data into three parts $\mathcal{D}^{tr}, \mathcal{D}^{val}, \mathcal{D}^{te} \subseteq X \times Y$. Here $\mathcal{D}^{tr}$ is the training set of $f_\theta \in \mathcal{F}$ (not required if f_θ is already trained). Following the intuition above, $\mathcal{D}^{val}$ is used to train another model g_ψ on the distribution $\mathcal{S}(f_\theta, \mathcal{D}^{val}_{X\setminus Z}) \times \mathcal{D}^{val}_Z$ where the predictive features are in the explanation distribution $\mathcal{S}(f_\theta, \mathcal{D}^{val}_{X\setminus Z})$ (excluding Z) and the target feature is the protected feature Z. The model belongs to a family $\mathcal{G}$, possibly different from $\mathcal{F}$. The parameter ψ optimizes a loss function ℓ:

$$\psi = \arg\min_{\tilde{\psi}} \sum_{(x,z)\in\mathcal{D}^{val}} \ell(g_{\tilde{\psi}}(\mathcal{S}(f_\theta, x)), z) \tag{3.1}$$

Finally, we use $\mathcal{D}^{te}$ for testing the approach and for comparison with baselines. To evaluate whether there is a fairness violation or not, we perform a statistical test of independence based on the AUC. See the workflow of Figure 4.1 for a visualization of the process.

Besides detecting and measuring fairness violations in machine learning models, a common desideratum is to understand what are the specific features driving the discrimination. Using the "Equal Treatment Inspector" as an auditor method that aims to

depict and quantify possible fairness violations does not only report a metric but provides information on *which features are the cause of the un-equal treatment*. We propose to solve this issue by applying explainable AI techniques to the Inspector. The "Equal Treatment Inspector" can provide different types of explanations, re-purposing the goal of the traditional explainable AI field, from understanding the model predictions to accounting for the reasons of un-equal treatment.

3.4 Theoretical Analysis

Throughout this section, we assume an exact calculation of the Shapley values $\mathcal{S}(f_\theta, x)$ for instance x, possibly for the observational and interventional variants (see (2.4,2.5) in the foundations Chapter 2.2). In the experimental section, we will use non-IID data and non-linear models.

3.4.1 Equal Treatment Given Shapley Values of Protected Attribute

Can we measure Equal Treatment by looking only at the Shapley value of the protected feature? The following result considers a linear model (with unknown coefficients) over *independent* features. In such a very simple case, resorting to Shapley values leads to an exact test of both Demographic Parity and Equal Treatment, which turn out to coincide. In the following, we write $distinct(\mathcal{D}_X, i)$ for the number of distinct values in the i-th feature of dataset $\mathcal{D}_X$, and $\mathcal{S}(f_\beta, \mathcal{D}_X)_i \equiv 0$ if the Shapley values of the i-th feature are all 0's.

Lemma 3.5. *Consider a linear model $f_\beta(x) = \beta_0 + \sum_j \beta_j \cdot x_j$. Let Z be the i-th feature, i.e. $Z = X_i$, and let $\mathcal{D}_X$ be such that $distinct(\mathcal{D}_X, i) > 1$. If the features in X are independent, then $\mathcal{S}(f_\beta, \mathcal{D}_X)_i \equiv 0 \Leftrightarrow f_\beta(X) \perp Z \Leftrightarrow \mathcal{S}(f_\beta, X) \perp Z$.*

Proof. It turns out $\mathcal{S}(f_\beta, x)_i = \beta_i \cdot (x_i - E[X_i])$. This holds in general for the interventional variant (2.5), and assuming independent features, also for the observational variant (2.4) (Aas et al., 2021). Since $distinct(\mathcal{D}_X, i) > 1$, we have that $\mathcal{S}(f_\beta, \mathcal{D}_X)_i \equiv 0$ iff $\beta_i = 0$. By independence of X, this is equivalent to $f_\beta(X) \perp X_i$, i.e., $f_\beta(X) \perp Z$. Moreover, by the propagation of independence, this is also equivalent to $\mathcal{S}(f_\beta, X) \perp Z$. □

However, the result does not extend to the case of dependent features.

Example 3.1. *Consider $Z = X_2 = X_1^2$, and the linear model $f_\beta(x_1, x_2) = \beta_0 + \beta_1 \cdot x_1$ with $\beta_1 \neq 0$ and $\beta_2 = 0$, i.e., the protected feature is not used by the model. In the interventional variant, theuninformativeness property implies that $\mathcal{S}(f_\beta, x)_2 = 0$. However, this does not mean that $Z = X_2$ is independent of the output because $f_\beta(X_1, X_2) = \beta_0 + \beta_1 \cdot X_1 \not\perp X_2$. In the observational variant, Aas et al. (2021) show that:*

$$val(T) = \sum_{i \in N \setminus T} \beta_i \cdot E[X_i | X_T = x_T^\star] + \sum_{i \in T} \beta_i \cdot x_i^\star$$

from which, we calculate: $\mathcal{S}(f_\beta, x^\star)_2 = \frac{\beta_1}{2} E[X_1 | X_2 = x_2^\star]$. We have $\mathcal{S}(f_\beta, \mathcal{D}_X)_2 \equiv 0$ iff $E[X_1 | x_2 = x_2^\star] = 0$ for all $x^\star$ in $\mathcal{D}_X$. For the marginal distribution $P(X_1 = v) = 1/4$ for $v = 1, -1, 2, -2$, and considering that $X_2 = X_1^2$, it holds that $E[X_1 | x_2 = v] = 0$ for all v. Thus $\mathcal{S}(f, \mathcal{D}_X)_2 \equiv 0$. However, again $f_\beta(X_1, X_2) = \beta_0 + \beta_1 \cdot X_1 \not\perp X_2$.

The counterexample shows that focusing only on the Shapley values of the protected feature is not a viable way to prove Demographic Parity of a model – and, a fortiori, neither to prove Equal Treatmentof the model, as will show in Lemma 3.6.

3.4.2 Equal Treatment vs Demographic Parity vs Fairness of the Input

We start by observing that *equal treatment* (independence of the explanation distribution from the protected attribute) is a sufficient condition fordemographic parity (independence of the prediction distribution from the protected attribute).

Lemma 3.6. *If* $\mathcal{S}(f_\theta, X) \perp Z$ *then* $f_\theta(X) \perp Z$.

Proof. By the propagation of independence in probability distributions, the premise implies $(\sum_i \mathcal{S}_i(f_\theta, X) + c) \perp Z$ where c is any constant. By setting $c = E[f(X)]$ and by the efficiency property, we have the conclusion. □

Therefore, a Demographic Parity violation (on the prediction distribution) is also an Equal Treatment violation (in the explanation distribution). Equal Treatment accounts for a stricter notion of fairness. The other direction does not hold. We can have dependence of Z from the explanation features, but the sum of such features cancels out, resulting in perfect Demographic Parity on the prediction distribution. This issue is also known as Yule's effect (Ruggieri et al., 2023).

Example 3.2. *Consider the model* $f(x_1, x_2) = x_1 + x_2$. *Let* $Z \sim Ber(0.5)$, $A \sim U(-3, -1)$, *and* $B \sim N(2, 1)$ *be independent, and let us define:*

$$X_1 = A \cdot Z + B \cdot (1 - Z) \quad X_2 = B \cdot Z + A \cdot (1 - Z)$$

We have $f(X_1, X_2) = A + B \perp Z$ *since* A, B, Z *are independent. Let us calculate* $\mathcal{S}(f, X)$ *in the two cases* $Z = 0$ *and* $Z = 1$. *If* $Z = 0$, *we have* $f(X_1, X_2) = B + A$, *and then* $\mathcal{S}(f, X)_1 = B - E[B] = B - 2 \sim N(0, 1)$ *and* $\mathcal{S}(f, X)_2 = A - E[A] = A + 2 \sim U(-1, 1)$. *Similarly, for* $Z = 1$, *we have* $f(X_1, X_2) = A + B$, *and then* $\mathcal{S}(f, X)_1 = A - E[A] = A + 2 \sim U(-1, 1)$ *and* $\mathcal{S}(f, X)_2 = B - E[B] = B - 2 \sim N(0, 1)$. *This shows:*

$$P(\mathcal{S}(f, X)|Z = 0) \neq P(\mathcal{S}(f, X)|Z = 1)$$

and then $\mathcal{S}(f, X) \not\perp Z$. *Notice this example holds both for the interventional and the observational cases, as we exploited Shapley values of a linear model over independent features, namely* A, B, Z.

Statistical independence between the input X and the protected attribute Z, i.e., $X \perp Z$, is another fairness notion. It targets fairness of the (input) datasets, disregarding

the model f_θ. For fairness-aware training algorithms, which are able not to (directly or indirectly) rely on Z, violation of such a notion of fairness does not imply Equal Treatmentviolation nor Demographic Parity violation.

Example 3.3. *Let $X = X_1, X_2, X_3$ be independent features such that $E[X_1] = E[X_2] = E[X_3] = 0$, and $X_1, X_2 \perp Z$, and $X_3 \not\perp Z$. The target feature is defined as $Y = X_1 + X_2$, hence it is also independent from Z. Assume a linear regression model $f_\beta(x_1, x_2, x_3) = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$ trained from a sample data from (X, Y) with $\beta_1, \beta_2 \approx 1$ and $\beta_3 \approx 0$. Intuitively, this occurs when a number of features are collected to train a classifier without a clear understanding of which of them contributes to the prediction. It turns out that $X \not\perp Z$ but, for $\beta_3 = 0$ (which can be obtained by some fairness regularization method (Kamishima et al., 2011)), we have $f_\beta(X_1, X_2, X_3) = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 \perp Z$. By reasoning as in the proof of Lemma 3.5, we have $\mathcal{S}(f_\beta, X) = (\beta_1 \cdot X_1, \beta_2 \cdot X_2, 0)$ and then $\mathcal{S}(f_\beta, X) \perp Z$. This holds both in the interventional and in the observational variants.*

The above represents an example where the input data depends on the protected feature, but the model and the explanations are independent.

3.4.3 Statistical Independence Test via Classifier AUC Test

In this subsection, we introduce a statistical test of independence based on the AUC of a binary classifier. The test of $W \perp Z$ is stated in general form for multivariate random variables W and a binary random variable Z with $dom(Z) = \{0, 1\}$. In the next subsection, we will instantiate it to the case $W = \mathcal{S}(f_\theta, X)$.

Let $\mathcal{D} = \{(w_i, z_i)\}_{i=1}^n$ be a dataset of realizations of the random sample $(W, Z)^n \sim \mathcal{F}^n$ where $\mathcal{F}$ is unknown. The independence $W \perp Z$ can be tested via a two-sample test. In fact, we have $W \perp Z$ iff $P(W|Z) = P(W)$ iff $P(W|Z = 1) = P(W|Z = 0)$. We test whether the positives and negatives instances in $\mathcal{D}$ are drawn from the same distribution by a novel two-sample test, which does not require permutation of data nor equal proportion of positive and negatives as in (Lopez-Paz and Oquab, 2017, Sections 2 and 3). We rely on a probabilistic classifier $f : W \to [0, 1]$, for which $f(w)$ estimates $P(Z = 1|W = w)$, and on its AUC:

$$AUC(f) = E_{(W,Z),(W',Z')\sim\mathcal{F}}[I((Z - Z')(f(W) - f(W')) > 0) + 1/2 \cdot I(f(W) = f(W'))|Z \neq Z'] \tag{3.2}$$

Under the null hypothesis $H_0 : W \perp Z$, we have $AUC(f) = 1/2$.

Lemma 3.7. *If $W \perp Z$ then $AUC(f) = 1/2$ for any classifier f.*

Proof. Let us recall the definition of the Bayes Optimal classifier $f_{opt}(w) = P(Z = 1|W = w)$. For any classifier f, we have:

$$AUC(f_{opt}) \geq AUC(f) \geq 1 - AUC(f_{opt}) \tag{3.3}$$

The first bound $AUC(f_{opt}) \geq AUC(f)$ follows because the Bayes Optimal classifier minimizes the Bayes risk (Gao and Zhou, 2015). Assume the second bound does not hold, i.e., for some f we have $AUC(f_{opt}) < 1 - AUC(f)$. Consider the classifier $\bar{f}(w) = 1 - f(w)$. We have $AUC(\bar{f}) \geq 1 - AUC(f)$, and then $\bar{f}$ would contradict the first bound because $AUC(f_{opt}) < AUC(\bar{f})$.

If $W \perp Z$, then $P(Z = 1|W = w) = P(Z = 1)$, and then $f_{opt}(w)$ is constant. By (3.2), this implies $AUC(f_{opt}) = 1/2$. By (3.3), this implies $AUC(f) = 1/2$ for any classifier. □

As a consequence, any statistics to test $AUC(f) = 1/2$ can be used for testing $W \perp Z$. A classical choice is to resort to the Wilcoxon–Mann–Whitney test, which, however, assumes that the distributions of scores for positives and negatives have the same shape. Better alternatives include the Brunner–Munzel test (Neubert and Brunner, 2007) and the Fligner–Policello test (Fligner and Policello, 1981). The former is preferable, as the latter assumes that the distributions are symmetric.

3.4.4 Testing for Equal Treatment via an Inspector

We instantiate the previous AUC-based method for testing independence to the case of testing for Equal Treatment via an Equal TreatmentInspector.

Theorem 3.8. *Let $g_\psi : \mathcal{S}(f_\theta, X) \to [0, 1]$ be an "Equal Treatment Inspector" for the model f_θ, and α a significance level. We can test the null hypothesis $H_0 : \mathcal{S}(f_\theta, X) \perp Z$ at $100 \cdot (1 - \alpha)\%$ confidence level using a test statistics of $AUC(g_\psi) = 1/2$.*

Proof. Under H_0, by Lemma 3.2 with $W = \mathcal{S}(f_\theta, X)$ and $f = g_\psi$, we have $AUC(g_\psi) = 1/2$. □

Results of such a test can include p-values of the adopted test for $AUC(g_\psi) = 1/2$. Alternatively, confidence intervals for $AUC(g_\psi)$ can be reported, as returned by the Brunner–Munzel test or by the methods (DeLong et al., 1988; Cortes and Mohri, 2004; Gonçalves et al., 2014).

3.4.5 Statistical Independence Test via Classifier AUC Test

We complement the experiments of Section 4.3 by reporting in Table 3.1 the results of the C2ST for group pair-wise comparisons. As discussed in Section 3.4.3, we perform

the statistical test $H_0 : AUC = 1/2$ of the "Equal Treatment Inspector" using a Brunner-Munzel one tailed test against $H_1 : AUC > 1/2$ as implemented by Virtanen et al. (2020). Table 3.1 reports the empirical AUC on the test set, the confidence intervals at 95% confidence level (columns "Low" and "High"), and the p-value of the test. The "Random" row regards a randomly assigned group and represents a baseline for comparison. The statistical tests clearly show that the AUC is significantly different from $1/2$, also when correcting for multiple comparison tests.

TABLE 3.1: Results of the C2ST on the "Equal Treatment Inspector".

Pair	**AUC**	**Low**	**High**	**pvalue**	**Test Statistic**
Random	0.501	0.494	0.507	0.813	0.236
White-Other	0.735	0.731	0.739	$<$ 2.2*e*-16	97.342
White-Black	0.62	0.612	0.627	$<$ 2.2*e*-16	27.581
White-Mixed	0.615	0.607	0.624	$<$ 2.2*e*-16	23.978
Asian-Other	0.795	0.79	0.8	$<$ 2.2*e*-16	107.784
Asian-Black	0.667	0.659	0.676	$<$ 2.2*e*-16	38.848
Asian-Mixed	0.644	0.634	0.653	$<$ 2.2*e*-16	28.235
Other-Black	0.717	0.708	0.725	$<$ 2.2*e*-16	48.967
Other-Mixed	0.697	0.688	0.707	$<$ 2.2*e*-16	39.925
Black-Mixed	0.598	0.586	0.61	$<$ 2.2*e*-16	15.451

We also compare the power of the C2ST based on the AUC using the Brunner-Munzel test against the two-sample test of Lopez-Paz and Oquab (2017), which is based on accuracy, and against the AUC test of Chakravarti et al. (2023), which is based on the asyntotic normal approximation of the Wilcoxon–Mann–Whitney statistics. We generate synthetic datasets from $\mathbf{X} \times Z$, where $Z \sim Ber(0.5)$ (balanced groups) or $Z \sim Ber(0.2)$ (unbalanced groups), and $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ with positives ($Z = 1$) distributed as $N(\begin{bmatrix}\mu & \mu\end{bmatrix}, \mathbf{\Sigma})$ and negatives ($Z = 0$) distributed as $N(\begin{bmatrix}-\mu & -\mu\end{bmatrix}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \begin{bmatrix}1 & 0.5 \\ 0.5 & 1\end{bmatrix}$. Thus, the larger the parameter μ, the easier is to distinguish the distributions of positives and negatives. Figure 3.2 reports the power[8] of the three tests using in all cases a logistic regression classifier. The power is estimated by 1,000 runs for each of the μ's ranging from 0.005 to 0.5. The figure highlights that, under such a setting, testing the AUC using the Brunner-Munzel test achieves better power than using accuracy or using an asymptotic test. Our approach exhibits the best power, and the difference is higher in the case groups are unbalanced.

3.4.6 Explaining Un-Equal Treatment

The following example showcases one of our main contributions: detecting *the sources* of un-equal treatment through interpretable by-design (linear) inspectors. Here, we

[8] Probability of rejecting H_0 when it does not hold.

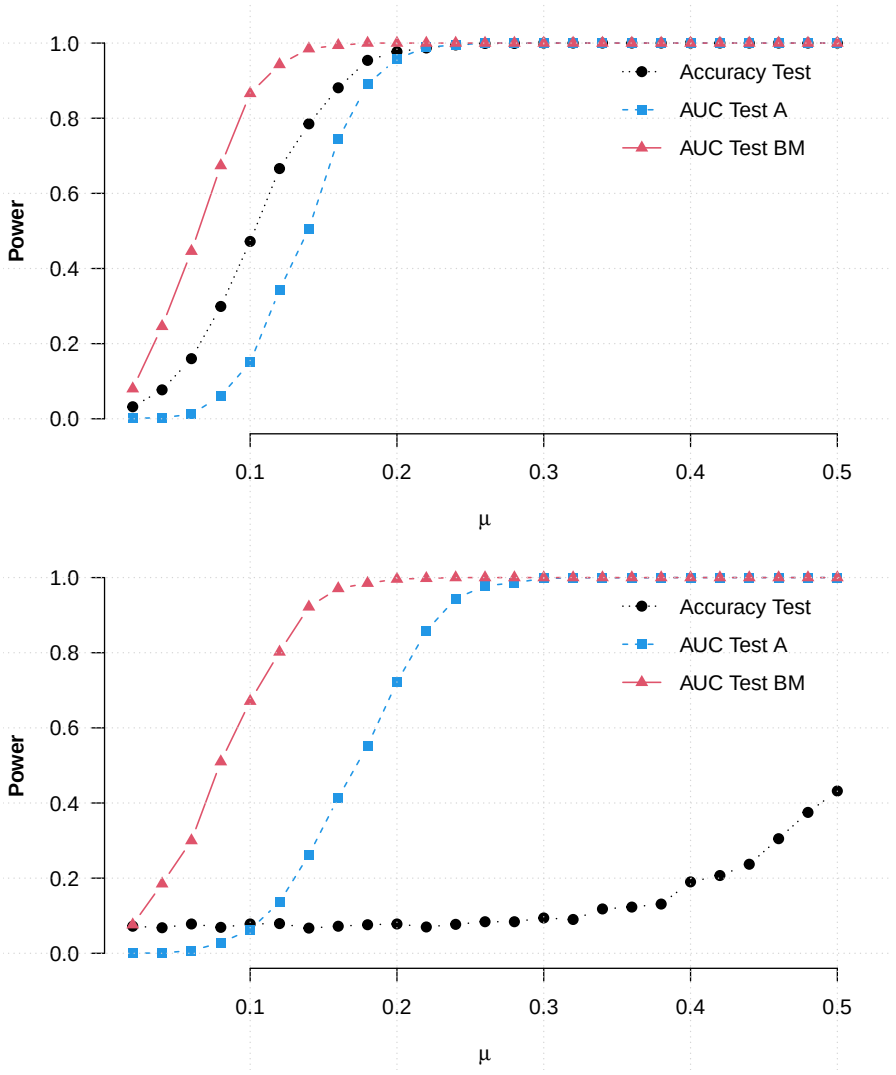

FIGURE 3.2: Comparing the power (the higher, the better) of C2ST based on AUC with Brunner-Munzel test (AUC Test BM) vs Accuracy vs AUC with an asymptotic normal approximation of the Wilcoxon–Mann–Whitney statistics (AUC Test A). Upper: balanced groups ($P(Z = 1) = 0.5$). Lower: unbalanced groups ($P(Z = 1) = 0.2$).

assume that the model is also linear. In the Appendix 3.4.6, we will experiment with non-linear models.

Example 3.4. *Let $X = X_1, X_2, X_3$ be independent features such that $E[X_1] = E[X_2] = E[X_3] = 0$, and $X_1, X_2 \perp Z$, and $X_3 \not\perp Z$. Given a random sample of i.i.d. observations from (X, Y), a linear model $f_\beta(x_1, x_2, x_3) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$ can be built by OLS (Ordinary Least Square) estimation, possibly with $\beta_1, \beta_2, \beta_3 \neq 0$. By reasoning as in the proof of Lemma 3.5, $\mathcal{S}(f_\beta, x)_i = \beta_i \cdot x_i$. Consider now a linear Equal TreatmentInspector $g_\psi(s) = \psi_0 + \psi_1 \cdot s_1 + \psi_2 \cdot s_2 + \psi_3 \cdot s_3$, which can be written in terms of the x's as: $g_\psi(x) = \psi_0 + \psi_1 \cdot \beta_1 \cdot x_1 + \psi_2 \cdot \beta_2 \cdot x_2 + \psi_3 \cdot \beta_3 \cdot x_3$. By OLS estimation properties, we have $\psi_1 \approx cov(\beta_1 \cdot X_1, Z)/var(\beta_1 \cdot X_1) = cov(X_1, Z)/(\beta_1 \cdot var(X_1)) = 0$ and analogously $\psi_2 \approx 0$. Finally, $\psi_3 \approx cov(X_3, Z)/(\beta_3 \cdot var(X_3)) \neq 0$. In summary, the coefficients of g_ψ provide information about which feature contributes (and how much it contributes) to the dependence*

between the explanation $\mathcal{S}(f_\beta, X)$ and the protected feature Z. Notice that also $f_\beta(X) \not\perp Z$, but we cannot explain which features contribute to such a dependence by looking at $f_\beta(X)$, since $\beta_i \approx cov(X_i, Y)/var(X_i)$ can be non-zero also for $i = 1, 2$.

3.4.7 Using AUC as Classifier Two-Sample Test evaluation metric

We complement the above results of the experimental with a further experiment relating the correlation hyperparameter γ to the coefficients of an explainable Equal Treatmentinspector. We consider a synthetic dataset with one more feature, by drawing $10,000$ samples from a $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, and (X_3, X_5) and (X_4, X_5) following bivariate normal distributions $N\left(\begin{bmatrix}0\,0\end{bmatrix}, \begin{bmatrix}1 & \gamma\,\gamma & 1\end{bmatrix}\right)$ and $N\left(\begin{bmatrix}0\,0\end{bmatrix}, \begin{bmatrix}1 & \gamma 0.5\,\gamma 0.5 & 1\end{bmatrix}\right)$, respectively. We define the binary protected feature Z with values $Z = 1$ if $X_5 > 0$ and $Z = 0$ otherwise. As in Section 4.3.2, we consider two experimental scenarios. In the first scenario, the indirect case, we have unfairness in the data and in the model. The targe feature is $Y = \sigma(X_1 + X_2 + X_3 + X_4)$, where σ is the logistic function. In the second scenario, the uninformative case, we have unfairness in the data and fairness in the model. The target feature is $Y = \sigma(X_1 + X_2)$.

Figure 3.3 shows how the coefficients of the inspector g_ψ vary with correlation γ in both scenarios. In the indirect case, coefficients for $\mathcal{S}(f_\theta, X_1)_1$ and $\mathcal{S}(f_\theta, X_1)_2$ correctly attributes zero importance to such variables, while coefficients for $\mathcal{S}(f_\theta, X_1)_3$ and $\mathcal{S}(f_\theta, X_1)_4$ grow linearly with γ, and with the one for $\mathcal{S}(f_\theta, X_1)_3$ with higher slope as expected. In the uninformative case, coefficients are correctly zero for all variables.

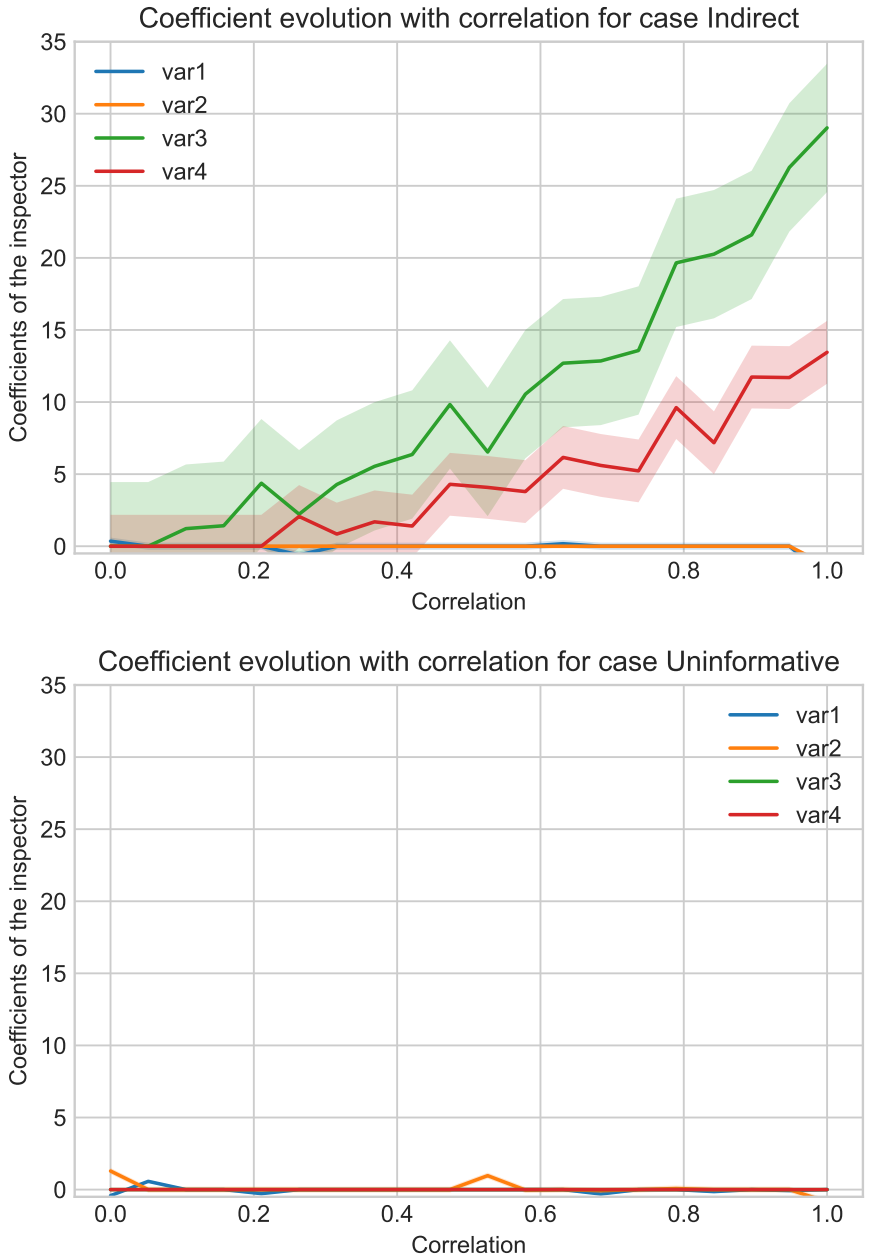

FIGURE 3.3: Coefficient of g_ψ over γ for synthetic datasets in two experimental scenarios.

3.5 Experiments

In our experiments, we aim to rigorously assess equal treatment by employing a methodology that includes:

- Experiments on synthetic data in Section 3.5.2
- Different types of Shapley values estimation in Section 3.5.2.1
- Using LIME as Local Feature Attribution Method 3.5.2.2
- Compare AUC vs accuracy for the Classifier Two Sample independence test 3.4.7
- Experiments natural datasets in Section 3.6
- Systematically varying the model f in Section 3.6.3
- Systematically varying the parameters θ in Section 3.6.3
- Extend the comparison against Demographic Parity in Section 3.6.2

3.5.1 Comparison methods and baselines:

We evaluate the "Equal Treatment Inspector" that does Classifier Two-Sample Test on the explanation, g_ψ(eq. 3.1),against the same test on other distributions: input data distributions g_Y (also known as bias on the data), prediction distributions g_v(also known as demographic parity) and a combination of both g_ϕ (see Equation 3.5.1). These experiments are grounded on previously discussed theory (Section 3.4):

$$Y = \arg\min_{\tilde{Y}} \sum_{(x,z)\in\mathcal{D}^{val}} \ell(g_{\tilde{Y}}(x), z) \tag{3.4}$$

$$v = \arg\min_{\tilde{v}} \sum_{(x,z)\in\mathcal{D}^{val}} \ell(g_{\tilde{v}}(f_\theta(x)), z) \tag{3.5}$$

$$\phi = \arg\min_{\tilde{\phi}} \sum_{(x,z)\in\mathcal{D}^{val}} \ell(g_{\tilde{\phi}}(f_\theta(x), x), z) \tag{3.6}$$

3.5.2 Experiments with Synthetic Data

Dataset: To generate a synthetic dataset for both cases, we first draw $10,000$ samples from a normal distribution $X_1 \sim N(0,1), X_2 \sim N(0,1), (X_3, X_4) \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \gamma\\ \gamma & 1\end{bmatrix}\right)$. We then define a binary feature Z with values 1 `if` $X_4 > 0$, `else` 0. We compare the fairness auditing methods while increasing the correlation $\gamma = r(X_3, Z)$ from 0 to 1. In both experimental scenarios below, our model f_β, is a function over the domain of the features X_1, X_2, X_3.

Experimental scenarios:
Indirect Case: *Bias in the data and un-equal treatment model.* There is a demographic parity violation in the input data that is learned by the model. The predictor's domain is (X_1, X_2, X_3), and the features are independent of each other. The protected attribute is Z (binary-valued) and its correlation with the predictor's domain (X_3, Z) is parameterized by $\gamma(X_3, Z)$, allowing to adjust for discrimination. To generate the synthetic demographic parity violation in the model we create the target $Y = \sigma(X_1 + X_2 + X_3)$, where σ is the logistic function.

Uninformative Case: *Bias on the data but equal treatment model.* The demographic parity violation on the input data remains the same but the relationship of the target variable changes, it is now independent of the protected attribute. The target function is defined as $Y = \sigma(X_1 + X_2)$, and, the γ parameter allows adjusting for discrimination in the training data even if the model does not capture it. The target is then independent of the protected attribute, $Y \perp X_3 \Rightarrow Y \perp Z \Rightarrow f_\beta \perp Z$.

In Figure 3.4 we present and compare the different experiments on synthetic data. Overall, we find that learning on the explanations distributions captures un-equal treatment of the model in both situations. We say that a method is *Accountable* if the feature attributions identified are the ones that indeed contribute towards the synthetically generated discrimination for both methods, see Appendix 3.4.6 for the experiment.

3.5.2.1 Alternative Feature Attribution Explanation Methods: Shapley Value Calculations True to the Model or True to the Data?

The "Equal Treatment Inspector" proposed in this work relies on the explanation distributions that satisfy efficiency and uninformative theoretical properties. We have used the Shapley values as an explainable AI method that satisfies these properties. A variety of (current) papers discusses the application of Shapley values, for feature attribution in machine learning models Strumbelj and Kononenko (2014); Lundberg et al. (2020b, 2018). However, the correct way to connect a model to a coalitional game, which is the central concept of Shapley values, is a source of controversy, with two main approaches (i) an interventional (Aas et al., 2021; Frye et al., 2020; Zern et al., 2023) or (ii)

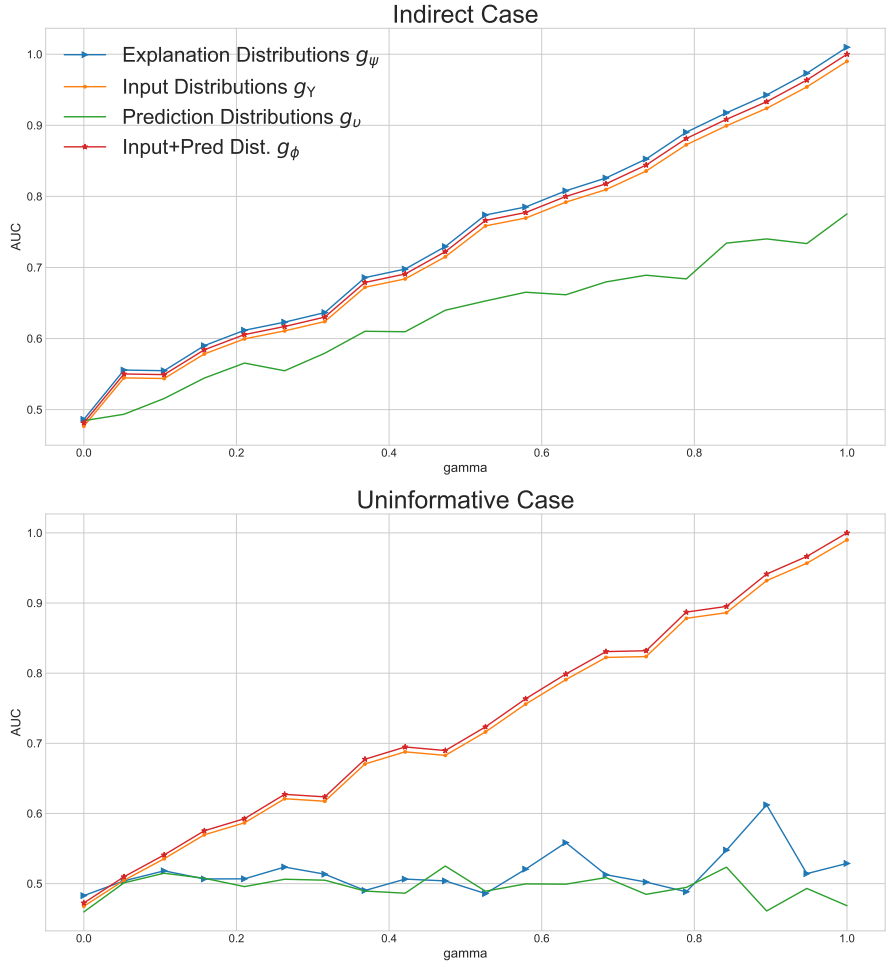

FIGURE 3.4: In the figure above, "Indirect case": All distributions capture this fairness violation; only the prediction distributions are less sensitive due to their low dimensionality. In the figure below, "Uninformative case", only the explanation distributions and prediction distributions detect that the model is non-discriminant, input data flags a false positive detection.

an observational formulation of the conditional expectation (Sundararajan and Najmi, 2020; Datta et al., 2016; Mase et al., 2019).

In the following experiment, we compare what are the differences between estimating the Shapley values using one or the other approach. We benchmark this experiment on the four prediction tasks based on the US census data Ding et al. (2021a) and using the "Equal Treatment Inspector", where both the model $f_\theta(X)$ and $g_\psi(\mathcal{S}(f_\theta, X))$ are linear models. We will calculate the Shapley values using the SHAP linear explainer. [9]

The comparison depends on a feature perturbation hyperparameter: whether the approach to compute the SHAP values is either *interventional* or *correlation dependent*. The interventional SHAP values break the dependence structure between features in the model to uncover how the model would behave if the inputs are changed (as it was an

[9] https://shap.readthedocs.io/en/latest/generated/shap.explainers.Linear.html

intervention). This option is said to stay "true to the model" meaning it will only give allocation credit to the features that are actually used by the model.

On the other hand, the full conditional approximation of the SHAP values respects the correlations of the input features. If the model depends on one input that is correlated with another input, then both get some credit for the model's behaviour. This option is said to say "true to the data", meaning that it only considers how the model would behave when respecting the correlations in the input data Chen et al. (2020).

In our case, we will measure the difference between the two approaches by looking at the linear coefficients of the model g_ψ and comparing the performance of predicting protected attributes, for this case only between White-Other.

TABLE 3.2: AUC comparison of the "Explanation Shift Detector" between estimating the Shapley values between the interventional and the correlation-dependent approaches for the four prediction tasks based on the US census dataset Ding et al. (2021a). The % character represents the relative difference. The performance differences are negligible.

	Interventional	Observational	%
Income	0.736438	0.736439	1.1e-06
Employment	0.747923	0.747923	4.44e-07
Mobility	0.690734	0.690735	8.2e-07
Travel Time	0.790512	0.790512	3.0e-07

TABLE 3.3: Linear regression coefficients comparison of the "Explanation Shift Detector" between estimating the Shapley values between the interventional and the correlation-dependent approaches for one of the US census based prediction tasks (ACS Income). The % character represents the relative difference. The coefficients show negligible differences between the calculation methods

	Interventional	Observational	%
Marital	0.348170	0.348190	2.0e-05
Worked Hours	0.103258	-0.103254	3.5e-06
Class of worker	0.579126	0.579119	6.6e-06
Sex	0.003494	0.003497	3.4e-06
Occupation	0.195736	0.195744	8.2e-06
Age	-0.018958	-0.018954	4.2e-06
Education	-0.006840	-0.006840	5.9e-07
Relationship	0.034209	0.034212	2.5e-06

Tables 3.2 and 3.3 compare the effects of using interventional and correlation-dependent approaches to train the "Explanation Shift Detector". Table 3.2 presents the AUC values for the four prediction tasks in the US Census dataset, showing negligible performance differences between the two methods. Table 3.3 provides a comparison of linear regression coefficients for one prediction task (ACS Income), illustrating similarly minimal variation between the approaches. While the two methods differ theoretically, the observed differences are negligible both in the resulting AUC scores and in the linear regression coefficients for explaining the protected characteristic. This experiment

underscores that, although estimation methods for Shapley values may diverge at the individual sample level, these discrepancies largely vanish when evaluating distributions of SHAP values, highlighting the robustness of the proposed approach.

3.5.2.2 Alternative Feature Attribution Explanation Methods: LIME as an Alternative to Shapley Values

The definition of ET (Def. 3.4) is parametric in the explanation function. We used Shapley values for their theoretical advantages (see Appendix 2.2). Another widely used feature attribution technique is LIME (Local Interpretable Model-Agnostic Explanations). The intuition behind LIME is to create a local linear model that approximates the behavior of the original model in a small neighbourhood of the instance to explain (Ribeiro et al., 2016b,a), whose mathematical intuition is very similar to the Taylor/Maclaurin series. This section discusses the differences in our approach when adopting LIME instead of the SHAP implementation of Shapley values. First of all, LIME has certain drawbacks:

- **Computationally Expensive:** Its current implementation is more computationally expensive than current SHAP implementations such as TreeSHAP (Lundberg et al., 2020b), Data SHAP (Kwon et al., 2021; Ghorbani and Zou, 2019), or Local and Connected SHAP (Chen et al., 2019). This problem is exacerbated when producing explanations for multiple instances (as in our case). In fact, LIME requires sampling data and fitting a linear model, which is a computationally more expensive approach than the aforementioned model-specific approaches to SHAP. A comparison of the execution time is reported in the next sub-section.
- **Local Neighborhood:** The randomness in the calculation of local neighbourhoods can lead to instability of the LIME explanations. Works including (Slack et al., 2020a; Alvarez-Melis and Jaakkola, 2018; Adebayo et al., 2018a) highlight that several types of feature attributions explanations, including LIME, can vary greatly.
- **Dimensionality:** LIME requires, as a hyperparameter, the number of features to use for the local linear model. For our method, all the features in the explanation distribution should be used. However, linear models suffer from the curse of dimensionality. In our experiments, this is not apparent, since our synthetic and real datasets are low-dimensional.

Figure 4.12 compares the AUC of the Equal Treatment inspector using SHAP and LIME as explanation functions over the synthetic dataset of Section 3.5.2. In both scenarios (indirect case and uninformative case), the two approaches have similar results.

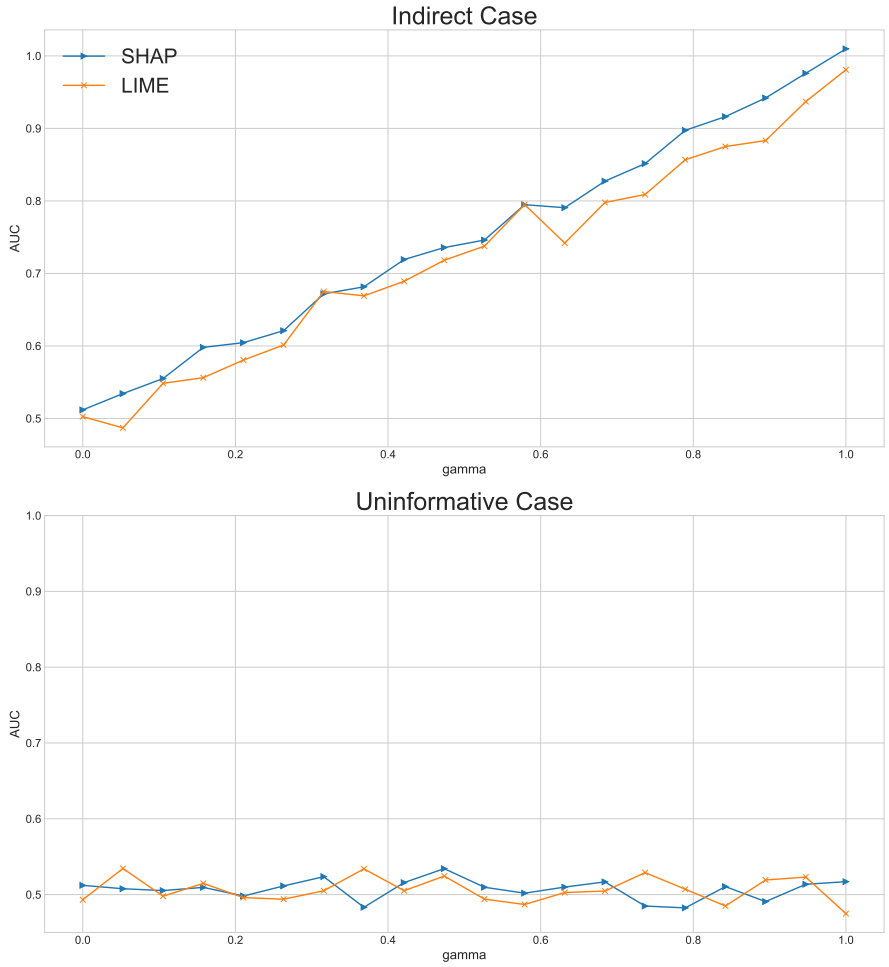

FIGURE 3.5: AUC of the ET inspect using SHAP vs using LIME.

3.6 Experiments on Real Data

We experiment here with datasets derived from the ACS data[10] (Ding et al., 2021a),. The fairness notions are tested against all pairs of groups from the protected attribute "Race".

We divide a dataset into three equal splits $\{\mathcal{D}^{tr}, \mathcal{D}^{val}, \mathcal{D}^{te}\} \subseteq \mathcal{D}$ and select our protected attribute, Z, to be a feature that indicates the ethnicity of an individual. We train our model f_β on $\{X^{tr}, Y^{tr}\}$ and, the "Equal Treatment Inspector" g_ψ on $\{\mathcal{S}(f_\beta, X^{val}), Z^{val}\}$. Both methods are evaluated on $\{X^{te}, Z^{te}, y^{te}\}$. For the type of models, f_β, as we are in tabular data we focus on we choose f_θ to be a `xgboost`Chen and Guestrin (2016a) that achieve state-of-the-art model performance Grinsztajn et al. (2022a); Elsayed et al. (2021b); Borisov et al. (2021),and for the inspector g_ψ a logistic regression. The final explanations are given by the coefficients of the logistic regression. We compare the AUC performances of several inspectors: g_ψ (see Eq. 3.1) for Equal Treatment (see Def. 3.4), g_v for Demographic Parity (see Def. 2.2), g_Y for fairness of the input (i.e., $X \perp Z$ as discussed in Section 3.4.2), and a combination g_ϕ of the last two inspectors to test $f_\theta(X), X \perp Z$.

3.6.1 Equal Treatment vs Demographic Parity: ACS Census

3.6.1.1 ACS Income

Figure 3.6 (above) shows the AUC performances of the Equal Treatment inspector g_ψ and the DT inspector g_v. The standard deviation of the AUC is calculated over 30 bootstrap runs, each one splitting the data into $1/3$ for training the model, $1/3$ for training the inspectors and $1/3$ for testing them. In the Secion 3.4.5, the results of the C2ST test of Section 3.4.1 are reported. The AUCs for the EP inspectors are greater than for the Demographic Parity inspectors, as expected due to Lemmma 3.6.

Figure 3.6 (below) shows the Wasserstein distance between the coefficients of the linear regressor g_ψ compared to a baseline where groups are assigned at random in the input dataset. This feature importance post-hoc explanation method provides insights into the impact of different features as sources of unfairness. We observe "Education" as a highly discriminatory proxy while the role of the feature "Worked Hours Per Week" is less relevant. This allows us to identify areas where adjustments or interventions may be needed to move closer to the ideal of equal treatment.

[10] ACS PUMS documentation: https://www.census.gov/programs-surveys/acs/microdata/documentation.html

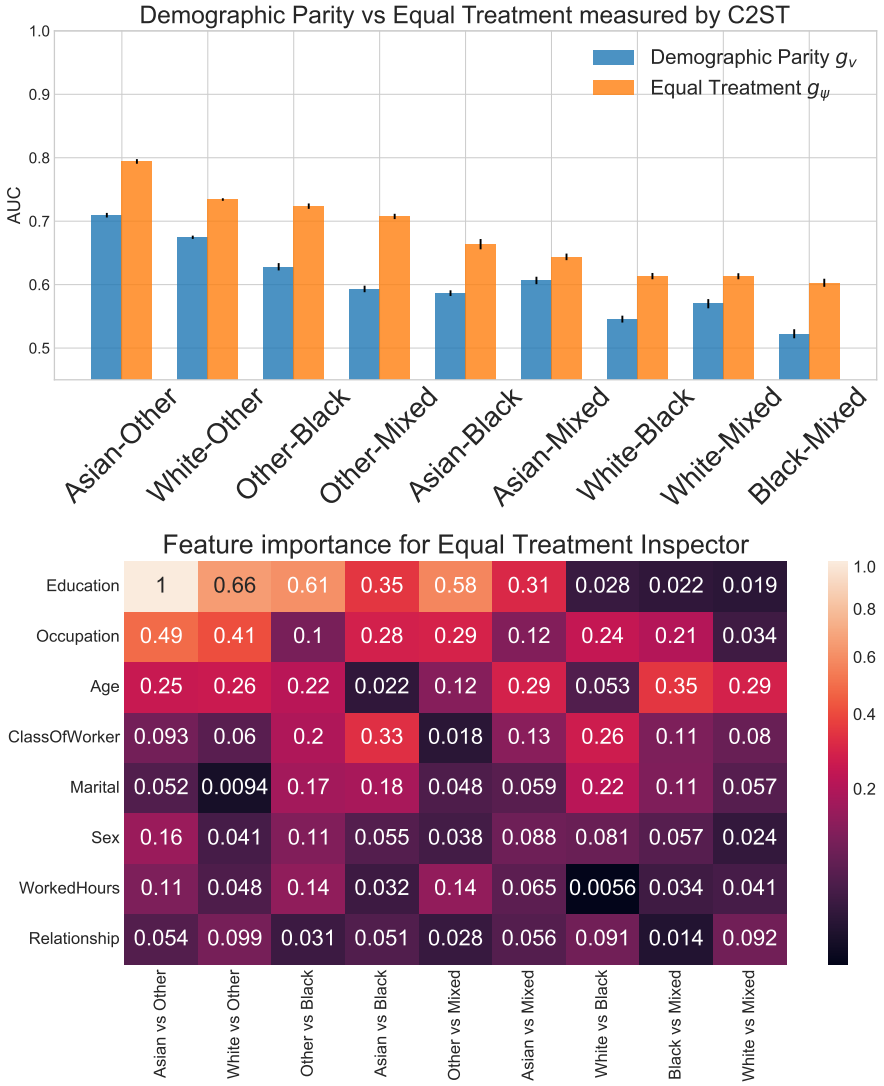

FIGURE 3.6: In the figure abobe, a comparison of Equal Treatment and Demographic Parity measures on the US Income data. The AUC range for Equal Treatment is notably wider, and aligning with the theoretical section, there are indeed instances where Demographic Parity fails to identify discrimination that Equal Treatment successfully detects. For a detailed statistical analysis, please refer to Appendix 3.6.2. The figure below provides insight into the influential features contributing to unequal treatment. Higher feature values correspond to a greater likelihood of these features being the underlying causes of unequal treatment.

3.6.1.2 ACS Employment

The goal of this task is to predict whether an individual is employed. Figure 3.7 shows a low DP violation, compared to the other prediction tasks based on the US census dataset. The AUC of the "Equal Treatment Inspector" is ranging from 0.55 to 0.70. For Asian vs Black un-equal treatment, we see a significant variation of the AUC, indicating that the method achieves different values on the bootstrapping folds. Looking at the features driving the Equal Treatment violation, we see particularly high values when

comparing "Asian" and "Black" populations, and for features "Citizenship" and "Employment". On average, the most important features across all group comparisons are also "Education" and "Area". Interestingly, features such as "difficulties on the hearing or seeing", do not play a role.

FIGURE 3.7: Above: AUC of the inspector for Equal Treatment and DP, over the district of California 2014 for the ACS Employment dataset. Below: contribution of features to the Equal Treatment inspector performance.

3.6.1.3 ACS Travel Time

The goal of this task is to predict whether an individual has a commute to work that is longer than 20 minutes. The threshold of 20 minutes was chosen as it is the US-wide median travel time to work based on 2018 data. Figure 3.8 shows an AUC for the Equal Treatment inspector in the range of 0.50 to 0.60. By looking at the features, they highlight different Equal Treatment drivers depending on the pair-wise comparison made.

In general, the features "Education", "Citizenship" and "Area" are the those with the highest difference. Even though for Asian-Black pairwise comparison "Employment" is also one of the most relevant features.

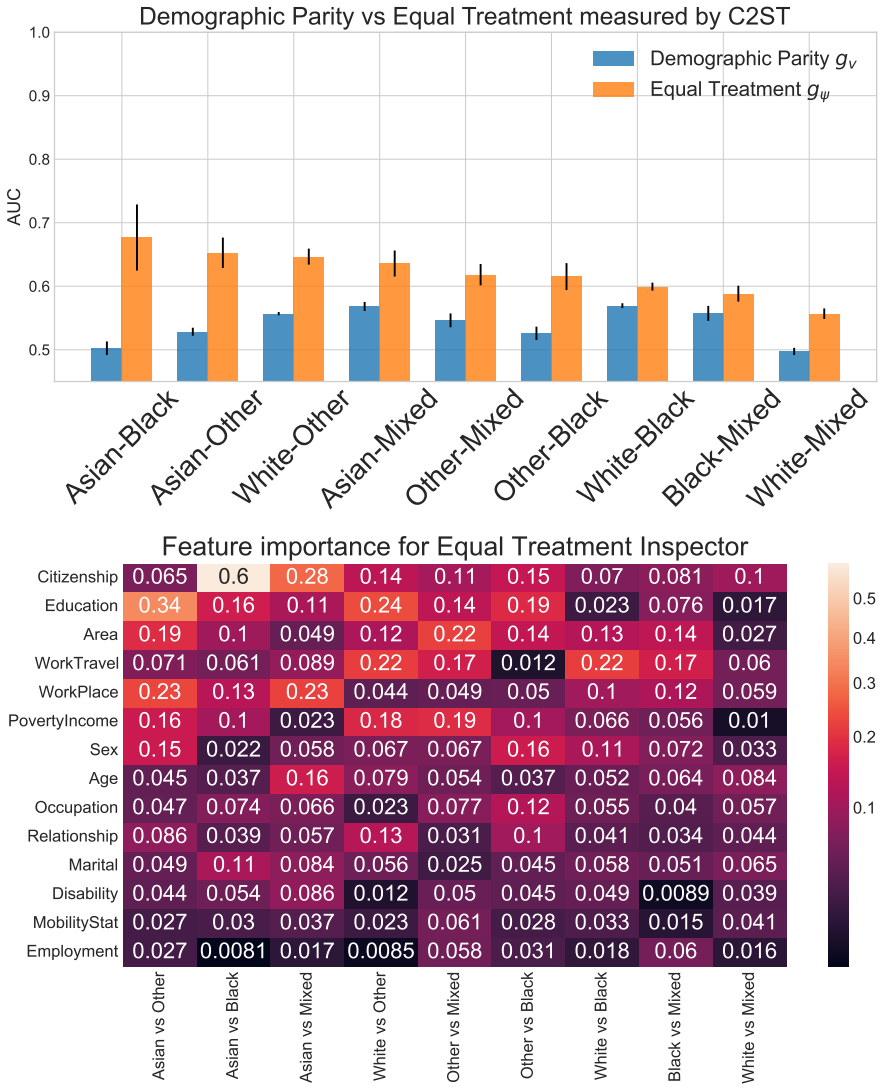

FIGURE 3.8: Above AUC of the inspector for Equal Treatment and DP, over the district of California 2014 for the ACS Travel Time dataset. Below: contribution of features to the Equal Treatment inspector performance.

3.6.1.4 ACS Mobility

The goal of this task is to predict whether an individual had the same residential address one year ago, only including individuals between the ages of 18 and 35. This filtering increases the difficulty of the prediction task, as the base rate of staying at the same address is above 90% for the general population (Ding et al., 2021a). Figure 3.9 show an AUC of the Equal Treatment inspector in the range of 0.55 to 0.80. By looking at the features, they highlight different sources of the Equal Treatment violation

depending on the group pair-wise comparison. In general, the feature "Ancestry", i.e. "ancestors' lives with details like where they lived, who they lived with, and what they did for a living", plays a high relevance when predicting Equal Treatment violation.

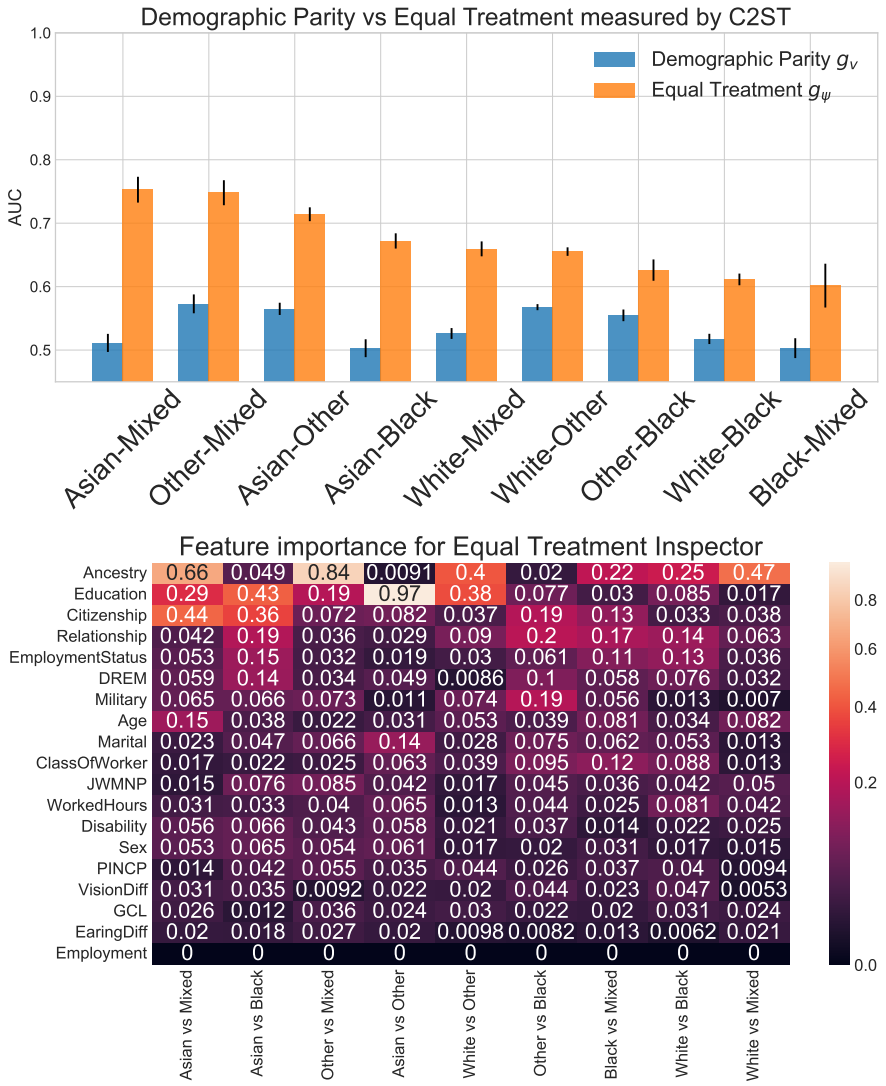

FIGURE 3.9: Above: AUC of the inspector for Equal Treatment and DP, over the district of California 2014 for the ACS Mobility dataset. Below: contribution of features to the Equal Treatment inspector performance.

3.6.2 Quantitative Analysis with Demographic Parity

So far, we measured Equal Treatment and Demographic Parity fairness using the AUC of an inspector, g_ψ and g_v, respectively. For DP, however, other probability density distance metrics can be considered, including the p-value of the Kolmogorov–Smirnov (KS) test and the Wasserstein distance. Table 3.4 reports all such distances in the format "mean $\pm$ stdev" calculated over 100 random sampled datasets. The pairs of group comparisons are sorted by descending AUC values. We highlight in red values below

the threshold of 0.05 for the KS test, of 0.55 for the AUC of the C2ST, and of 0.05 for the Wasserstein distance. They represent cases where Equal Treatment violation occurs, but no Demographic Parity violation is measured (with different metrics).

TABLE 3.4: Comparison ofEqual Treatment and Demographic Parity measured in different ways. The cases of equal Treatment violation but no demographic parity violation are highlighted in red.

Pair	Data	Equal treatment	Demographic Parity		
		C2ST(AUC)	C2ST(AUC)	KS(pvalue)	Wasserstein
Asian-Other	Income	0.794 ± 0.004	0.709 ± 0.004	0.338 ± 0.007	0.256 ± 0.004
White-Other	Income	0.734 ± 0.002	0.675 ± 0.003	0.282 ± 0.003	0.209 ± 0.002
Other-Black	Income	0.724 ± 0.004	0.628 ± 0.006	0.216 ± 0.007	0.143 ± 0.004
Other-Mixed	Income	0.707 ± 0.005	0.593 ± 0.005	0.169 ± 0.006	0.117 ± 0.004
Asian-Black	Income	0.664 ± 0.008	0.587 ± 0.004	0.142 ± 0.005	0.111 ± 0.004
Asian-Mixed	Income	0.644 ± 0.005	0.607 ± 0.006	0.159 ± 0.008	0.128 ± 0.006
White-Mixed	Income	0.613 ± 0.005	0.546 ± 0.005	0.082 ± 0.004	0.058 ± 0.002
White-Black	Income	0.613 ± 0.005	0.57 ± 0.007	0.113 ± 0.008	0.08 ± 0.006
Black-Mixed	Income	0.603 ± 0.006	0.523 ± 0.007	0.055 ± 0.007	0.023 ± 0.004
Asian-Black	TravelTime	0.677 ± 0.052	0.502 ± 0.011	0.021 ± 0.009	0.01 ± 0.003
Asian-Other	TravelTime	0.653 ± 0.024	0.528 ± 0.006	0.053 ± 0.011	0.027 ± 0.004
Asian-Mixed	TravelTime	0.647 ± 0.013	0.557 ± 0.003	0.096 ± 0.004	0.045 ± 0.002
White-Other	TravelTime	0.636 ± 0.02	0.568 ± 0.007	0.107 ± 0.01	0.06 ± 0.005
Other-Mixed	TravelTime	0.618 ± 0.017	0.546 ± 0.011	0.079 ± 0.012	0.043 ± 0.006
Other-Black	TravelTime	0.615 ± 0.021	0.526 ± 0.011	0.049 ± 0.014	0.026 ± 0.006
White-Black	TravelTime	0.599 ± 0.006	0.569 ± 0.004	0.12 ± 0.006	0.057 ± 0.003
Black-Mixed	TravelTime	0.588 ± 0.012	0.557 ± 0.012	0.098 ± 0.015	0.0557 ± 0.001
White-Mixed	TravelTime	0.557 ± 0.008	0.497 ± 0.006	0.016 ± 0.004	0.006 ± 0.002
Other-Black	Employment	0.744 ± 0.008	0.524 ± 0.005	0.036 ± 0.005	0.036 ± 0.004
Asian-Other	Employment	0.711 ± 0.011	0.557 ± 0.003	0.066 ± 0.004	0.066 ± 0.003
White-Other	Employment	0.695 ± 0.007	0.524 ± 0.003	0.019 ± 0.005	0.019 ± 0.002
Other-Mixed	Employment	0.683 ± 0.022	0.557 ± 0.008	0.083 ± 0.005	0.083 ± 0.003
Black-Mixed	Employment	0.678 ± 0.028	0.534 ± 0.007	0.049 ± 0.007	0.048 ± 0.004
Asian-Mixed	Employment	0.671 ± 0.019	0.61 ± 0.006	0.0144 ± 0.006	0.145 ± 0.004
Asian-Black	Employment	0.655 ± 0.021	0.587 ± 0.004	0.106 ± 0.006	0.106 ± 0.004
White-Mixed	Employment	0.651 ± 0.009	0.581 ± 0.006	0.095 ± 0.004	0.095 ± 0.003
White-Black	Employment	0.619 ± 0.011	0.544 ± 0.004	0.049 ± 0.003	0.049 ± 0.002
Asian-Mixed	Mobility	0.753 ± 0.02	0.511 ± 0.014	0.04 ± 0.012	0.014 ± 0.006
Other-Mixed	Mobility	0.748 ± 0.02	0.573 ± 0.015	0.113 ± 0.017	0.062 ± 0.009
Asian-Other	Mobility	0.714 ± 0.011	0.565 ± 0.01	0.114 ± 0.011	0.054 ± 0.005
Asian-Black	Mobility	0.672 ± 0.012	0.503 ± 0.014	0.032 ± 0.011	0.012 ± 0.004
Other-Black	Mobility	0.66 ± 0.012	0.526 ± 0.009	0.044 ± 0.009	0.02 ± 0.004
White-Mixed	Mobility	0.655 ± 0.007	0.568 ± 0.005	0.105 ± 0.007	0.044 ± 0.003
White-Other	Mobility	0.626 ± 0.017	0.555 ± 0.009	0.091 ± 0.01	0.046 ± 0.005
White-Black	Mobility	0.611 ± 0.009	0.518 ± 0.008	0.043 ± 0.008	0.017 ± 0.004
Black-Mixed	Mobility	0.602 ± 0.035	0.503 ± 0.016	0.031 ± 0.013	0.012 ± 0.006

3.6.3 Varying Estimator and Inspector

We vary here the model f_θ and the inspector g_ψ over a wide range of well-known classification algorithms. Table 4.7 shows that the choice of model and inspector impacts on the measure of Equal Treatment, namely the AUC of the inspector.

By Theorem 3.8, the larger the AUC of any inspector the smaller is the p-value of the null hypothesis $\mathcal{S}(f_\theta, X) \perp Z$. Therefore, inspectors able to achive the best AUC should be considered. Weak inspectors have lower probability of rejecting the null hypothesis when it does not hold.

	Model f_θ				
Inspector g_ψ	**DecisionTree**	**SVC**	**Logistic Reg.**	**RF**	**XGB**
DecisionTree	0.631	0.644	0.644	0.664	0.634
KNN	0.737	0.754	0.75	0.744	0.751
Logistic Reg.	0.767	0.812	0.812	0.812	0.821
MLP	0.786	0.795	0.795	0.813	0.804
RF	0.776	0.782	0.781	0.758	0.795
SVC	0.743	0.807	0.807	0.790	0.810
XGB	0.775	0.780	0.780	0.789	0.790

TABLE 3.5: AUC of theEqual Treatmentinspector for different combinations of models and inspectors.

3.6.4 Hyperparameters Evaluation

This section presents an extension to our experimental setup, where we increase the model complexity by varying the model hyperparameters. We use the US Income dataset for the population of the CA14 district. We consider three models for f_θ: Decision Trees, Gradient Boosting, and Random Forest. For the Decision Tree models, we vary the depth of the tree, while for the Gradient Boosting and Random Forest models, we vary the number of estimators. Shapley values are calculated by means of the TreeExplainer algorithm (Lundberg et al., 2020b). For theEqual Treatmentinspector g_ψ, we consider logistic regession, and XGB.

Figure 3.10 shows that less complex models, such as Decision Trees with maximum depth 1 or 2, are also less unfair. However, as we increase the model complexity, the unequal treatment of the model becomes more pronounced, achieving a plateau when the model has enough complexity. Furthermore, when we compare the results for different Equal Treatment inspectors, we observe minimal differences (note that the y-axis takes different ranges).

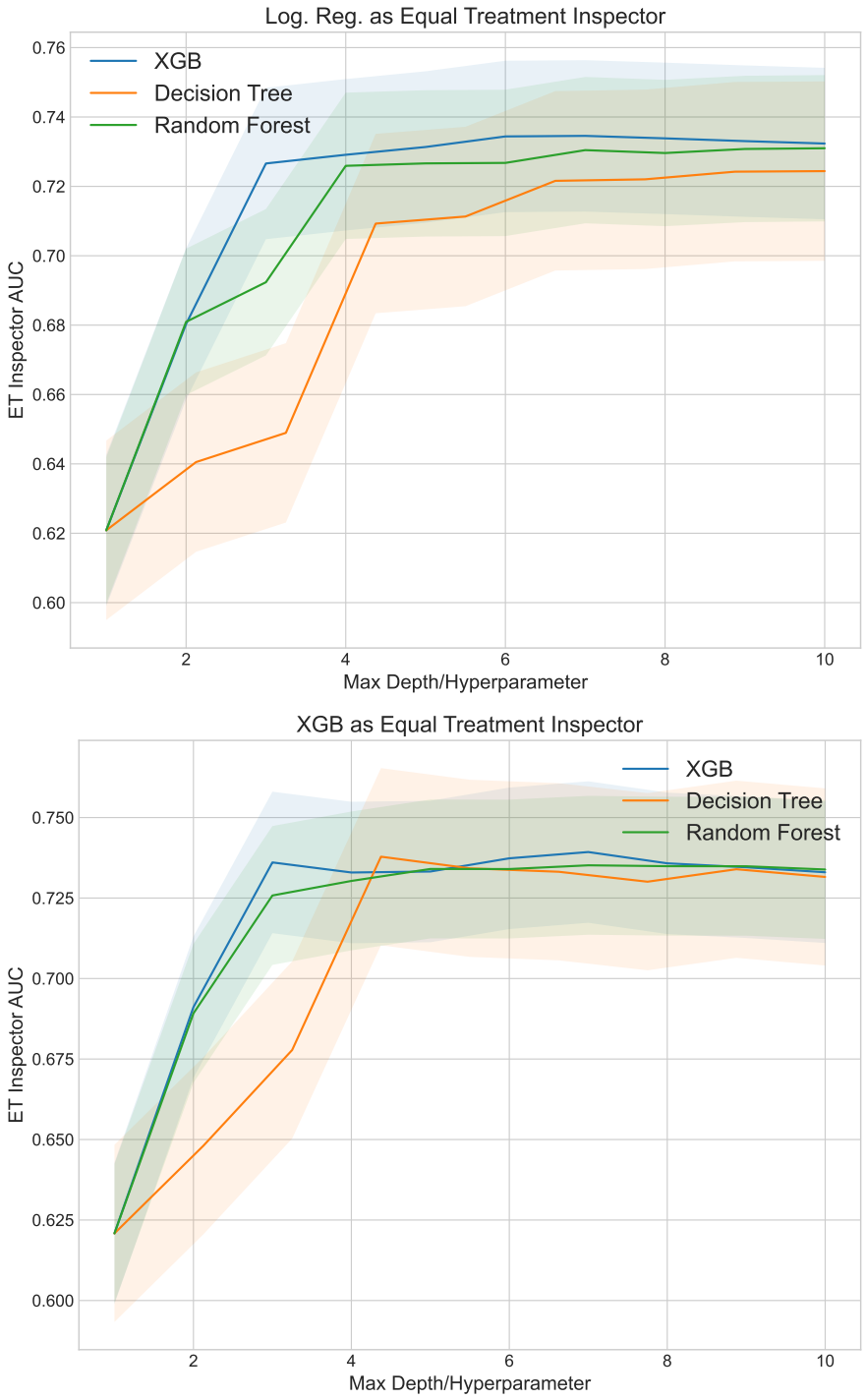

FIGURE 3.10: AUC of the inspector for ET, over the district of CA14 for the US Income dataset. The model f_θ is a decision tree based algorithm, that we modify the maximum depth (x-axis), while the model g_ψ varies in each of the images. The image above shows the results using a logistic regression detector, while the image below uses an XGBoost detector. By changing the model, we can see that simpler models (decision tree with depth 1 or 2) are less discriminative, while when increasing the model complexity, the unequal treatment of the model starts taking higher values.

3.7 Related Work on Measuring Equal Treatment

3.7.1 Fairness and Auditing

"Audits are tools for interrogating complex processes, often to determine whether they comply with company policy, industry standards or regulations" (Raji et al., 2020). Algorithmic fairness audits are closely linked to audits studies as understood in the social sciences, with a strong emphasis on social justice (Vecchione et al., 2021). A recent survey on public algorithmic audit identified four categories of "problematic machine behaviour" that can be unveiled by audit studies: discrimination, distortion, exploitation and misjudgement (Bandy, 2021; Liu et al., 2012). This taxonomy is highly helpful when putting forward auditing studies and also relevant for this work: our "Equal Treatment Inspector" can be seen as a tool to help understand discrimination in machine learning models, thus, falling into the first category.

Selecting a measure to compare fairness between two sensitive groups has been a highly discussed topic, where results such as (Chouldechova, 2017; Hardt et al., 2016; Kleinberg et al., 2017), have highlighted the impossibility to satisfy simultaneously three type of fairness measures: demographic parity (Dwork et al., 2012b), equalized odds (Hardt et al., 2016), and predictive parity (Corbett-Davies et al., 2017; Ruf and Detyniecki, 2021; Wachter et al., 2020). Previous work has relied on the notion of equal outcomes by measuring and calculating demographic and statistical parity on the model predictions (Kearns et al., 2018). The arguments of Simons et al. (2021) provide background and motivation towards interpretations of equal treatment as blindness and discuss its importance in practical policy applications.

3.7.2 Explainability for Fair Supervised Learning

The intersection of fairness and explainable AI has been an active topic in recent years. The most similar work is (Lundberg, 2020) where they apply Shapley values to statistical parity. Our work continues in-depth on this research line by formalizing the explanation distribution, deriving theoretical guarantees, and proposing novel methodology. Specifically, our approach allows for comparison across different protected groups. We also introduce the concept of accountability, which refers to the ability of the algorithm to provide insights into the sources of *equal treatment* violation of the model. Additionally, we evaluate our method on multiple datasets and synthetic examples, demonstrating its effectiveness in identifying and mitigating equal treatment disparities in machine learning models.

A recent line of work assumes knowledge about causal relationships between random variables. For example, Grabowicz et al. (2022) presents a post-processing method

based on Shapley values aiming to detect and nullify the influence of a protected attribute on the output of the system. They assume known direct causal links from the data to the protected attribute and no measured confounders. Our work does not rely on causal graphs knowledge but exploits the Shapley values' theoretical properties to obtain fairness model auditing guarantees.

Stevens et al. (2020) presents an approach to explaining fairness based on adapting the Shapley value function to explain model unfairness. They also introduce a new meta-algorithm that considers the problem of learning an additive perturbation to an existing model in order to impose fairness. In our work, we do not adopt the Shapley value function. Instead, we use the theoretical Shapley properties to provide fairness auditing guarantees. Our *Equal Treatment Inspector* is not perturbation-based but uses Shapley values to project the model to the explanation distribution, and then measures *un-equal treatment*. It also allows us to pinpoint what are the specific features driving this violation.

In Grabowicz et al. (2022), the authors present a post-processing method based on Shapley values aiming to detect and nullify the influence of a protected attribute on the output of the system. For this, they assumed there are direct causal links from the data to the protected attribute and that there are no measured confounders. Our work does not use causal graphs but exploits the theoretical properties of the Shapley values to obtain fairness model auditing guarantees.

Instead of using feature attribution explanation, other works have researched fairness using other explainability techniques such as counterfactual explanations (Kusner et al., 2017; Manerba and Guidotti, 2021; Mutlu et al., 2022). We don't focus on counterfactual explanations but on feature attribution methods that allow us to measure unequal feature contribution to the prediction. Further work can be envisioned by applying explainable AI techniques to the "Equal Treatment Inspector" or constructing the explanation distribution out of other techniques.

3.7.3 Classifier Two-Sample Test

The use of classifiers to obtain statistical tests or measure independence between two distributions has been previously explored in the literature (Lopez-Paz and Oquab, 2017). Specifically, the first authors introduced the use of "Classifier Two-Sample Tests (C2ST)" to represent the data that returns an interpretable unit test statistic, allowing it to measure how two distributions differ from each other. Their approach establishes the main theoretical properties, compares performance against state-of-the-art methods, and outlines applications of C2ST. Similar work of Liu et al. (2020a) propose a kernel-based to two-sample tests classification, aiming to determine whether two samples are

drawn from the same underlying distribution. Alike work has also been used in Kaggle competitions under the name of "Adversarial Validation" (Ellis, 2023; Guschin et al., 2018), a technique which aims to detect which features are distinct between train and leaderboard datasets to avoid possible leaderboard shakes. Our work builds on previous approaches by adopting their interpretable power test analysis and theoretical properties. Still, we focus on testing for fairness notions of equal treatment in the explanation distribution. As the test statistic for our approach, we use the Area Under the Curve (AUC) of the "Equal Treatment Inspector".

Another related work in the literature is by Edwards and Storkey (2016), who focuses on removing statistical parity from images by using an adversary that tries to predict the relevant sensitive variable from the model representation and censoring the learning of the representation of the model and data on images and neural networks. While methods for images or text data are often developed specifically for neural networks and cannot be directly applied to traditional machine learning techniques, we focus on tabular data where techniques such as gradient boosting decision trees achieve state-of-the-art model performance (Grinsztajn et al., 2022a; Elsayed et al., 2021b; Borisov et al., 2021). Furthermore, our model and data projection into the explanation distributions leverages Shapley value theory to provide fairness auditing guarantees. In this sense, our work can be viewed as an extension of their work, both in theoretical and practical applications.

3.8 Comparison with Related Fairness Metrics

In the main body of the paper, we have made an explicit comparison against demographic parity. In this section, we compare our proposed notion and model for equal treatment with respect to other metrics proposed (see Chapter 2).

TABLE 3.6: Fairness metrics alignment with *Equal Treatment* criteria. Requirements are derived from a use case explained in Section 3.1.1, and metrics are further discussed and defined in Section 2.4

Metric	Group Discrimination *(R1)*	Unlabelled Data *(R2)*	No Background Knowledge *(R3)*	Proxy Discrimination *(R4)*	Explanation Capabilities *(R5)*
Equal Treatment	✓	✓	✓	✓	✓
Demographic Parity	✓	✓	✓	✗	✗
Equal Opportunity	✓	✗	✓	✗	✗
Treatment Equality	✓	✗	✓	✗	✗
Feature Importance Disparity	✓	✓	✓	✗	✓
Counterfactual Fairness	✓	✓	✗	✓	✓
Fairness-Unawareness	✓	✓	✓	✗	✗
Individual Fairness	✗	✓	✓	✓	✓

3.8.1 Equal Opportunity/Equalized Odds

$$\text{TPR} = \frac{TP}{TP + FN} \tag{3.7}$$

$$\text{EOF}_z = \text{TPR}_z - \text{TPR} \tag{3.8}$$

A negative value in Equal Opportunity is due to the worse ability of a ML model to find actual True Positives for the protected group in comparison with the reference group. The reliance on true positive rates in formulating Equal Opportunity presents an additional challenge involving the calculation of False Negatives. In the context of the loan scenario, a False Negative occurs when a loan is denied to someone capable of repayment, but acquiring this data may not be feasible. An alternative approach involves maintaining a holdout set of randomly selected users to whom loans are uniformly granted, allowing for monitoring in that controlled environment.

Nevertheless, the implementation of a holdout set comes with potential economic and societal implications. Careful consideration is needed, as this approach may impact both financial outcomes and broader societal dynamics.

A similar logic can be applied to the other Equalized Odds, that also relies on having access to labeled data,hence it does not meet *(R2)*.

From the AI Alignment perspective, the alignment of both metrics—Equal Opportunity and Equalized Odds—with egalitarian ideals is evident, as they prioritize access to good outcomes through error ratios. These metrics also demonstrate a parallel alignment with utilitarian principles by focusing on minimising disparate errors, emphasizing a utilitarian approach to optimizing overall societal welfare.

In contrast, our conceptualization of equal treatment diverges in its alignment, finding resonance with distributive justice principles rooted in liberal ideology and deontological ethical considerations. This divergence reflects multiple perspectives and values and the nature of ethical foundations and justice paradigms. While Equal Opportunity and Equalized Odds lean towards utilitarian and egalitarian perspectives, our equal treatment notion is grounded in a more liberal-oriented distributive justice framework, emphasizing fairness and individual rights.

This diversity in alignment underscores, as expressed in the AI alignment foundation in Chapter 2 the multiple values of society and the challenge of defining them mathematically.

3.8.2 Treament Equality

$$\frac{FP_z}{FN_z} = \frac{FP}{FN} \tag{3.9}$$

From a technical perspective the challenge relies on obtaining false positive and false negative data, specially the later one as we have discussed in the previous section,hence it does not meet *(R2)*.

Regarding AI alignment, the original authors do not specify the principle or value with which the notion should align, and this clarity is absent from their definition. An illustrative sentence is "*men and women are being treated differently by the algorithm*". However, this statement may not be entirely accurate, as the algorithm may treat men and women equally while exhibiting different error rates. For instance, a constant classifier that grants loans to everyone treats everyone equally. Still, if males and females have distinct patterns of loan repayment, the classifier will yield different errors, resulting in distinct $\frac{FP}{FN}$ values.

Then this metric of "treament equality" is very distinct from our proposed notion due to not belonging to a clear distributive justice perspective and the technical implementation. Identically to equal opportunity, it's a utility measure; therefore, it is utilitarian rather than our notion that is based on principles rather than consequences, deontological.

3.8.3 Demographic Parity

Demographic parity, is not able to detect proxy discrimination *(R3)* nor capable to account for the sources behind discrimination *(R4)*, as we have discussed in the main body of the chapter.

Thus, our notion of equal treatment is an improvement with respect to demographic parity in the two dimensions that we have discussed in the foundation following Gabriel (2020): *(i)* better AI alignment with philosophical notions and *(ii)* translated as a stronger and more sensitive metric.

3.8.4 Individual Fairness

We say that a model achieves individual fairness w.r.t. two individual samples if:

$$d(f(x), f(x')) \leq \mathcal{L} \cdot d(x, x') \tag{3.10}$$
$$\forall x, x' \in X \tag{3.11}$$

From a philosophical perspective, individual fairness is closest to the liberal and deontological point of view. However, it fails the requirements of liberalist arguments. Liberalism argues for meritocracy, i.e. disparate treatment may be considered fair if it is based on varying efforts or preferences of individuals but unfair if it is based on protected characteristics that individuals did not choose. E.g. it's fair to hire someone because of better grades, but it is unfair if these grades depend on ethnicity. The definition of individual fairness does not consider protected characteristics or proxies thereof failing research requirement *(R1)* of group discrimination.. While individual fairness may be easily combined with blindness or adjusted definitions of similarities (Fleisher, 2021; Yurochkin et al., 2020), the treatment of proxies to protected characteristics is hard to avoid, rendering the alignment of an AI with distributive justice values difficult. Therefore, individual fairness is not a popular metric — unlike group fairness metrics (Fleisher, 2021).

3.8.5 Counterfactual Fairness

Counterfactual fairness, as defined by Kusner et al. (2017):

Definition 3.9. *Counterfactual Fairness* We say that the model f_θ is counterfactually fair if under any context $X = x$ and $A = a$

$$P(f_\theta(x_{Z=z}) = y | X = x, Z = z) = P(f_\theta(x_{Z=z'}) = y | X = x, Z = z) \tag{3.12}$$
$$\forall y \in Y \quad \forall z \in Z \quad \forall x \in X \tag{3.13}$$

Rosenblatt and Witter (2023) have shown that: (i) an algorithm that satisfies counterfactual fairness also satisfies demographic parity; and (ii) all algorithms that satisfy demographic parity can be modifed to satisfy counterfactual fairness. These results conclude that counterfactual fairness is equivalent to demographic parity, therefore failing the same requirements *(R4)*, proxy discrimination and *(R5)* explanation capabilities. Moreover, this fairness definition fails on our requirement *(R3)*, of not needing background knowledge – since counterfactuals can only be computed for a given causal graph.

3.8.6 Fairness Through Unawareness

This metric was initially proposed as a baseline by Grgic-Hlaca et al. (2018)

Definition 3.10. *(Fairness Through Unawareness.* An algorithm is fair if any protected attributes Z are not explicitly used in decision-making.

Any mapping $f : \mathbf{X} \to Y$ for which $Z \notin \mathbf{X}$ satisfies such a definition. It has a clear shortcoming as features in $\mathbf{X}$ can contain discriminatory information, known as proxy features for discrimination. Further, the research requirement of detecting *(R4) Proxy discrimination* is not met. One of the main contributions of our notion of equal treatment is that it captures statistical relations of all the features with the protected attribute. In our approach, the model does not consider the protected attribute to be present in the covariates X, therefore implying fairness through unawareness. Furthermore, in Section 3.4.1 of the main body, we discuss the limitations of analyzing the Shapley values of the protected attribute.

3.8.6.1 Decomposition Method Specific of Demographic Parity.

We formally compare our approach to the prior work of Lundberg (2020) and the related SHAP Python package documentation. The authors addressed DP using SHAP value estimation. This brief workshop paper emphasizes the importance of "decomposing a fairness metric among each of a model's inputs to reveal which input features may be driving any observed fairness disparities". In terms of statistical independence, the approach can be rephrased as decomposing $f_\theta(\mathbf{X}) \perp Z$ by examining $S(f_\theta, \mathbf{X})_i \perp Z$ individually for $i \in [1, p]$. Actually, the paper limits itself to consider only differences of means, namely testing for $E[\mathcal{S}(f_\theta, \mathbf{X})_i | Z = 1] \neq E[\mathcal{S}(f_\theta, \mathbf{X})_i | Z = 0]$. However, the method proposed by Lundberg (2020) is not sufficient nor necessary to prove DP, as shown next.

Lemma 3.11. *$f_\theta(\mathbf{X}) \perp Z$ is neither implied by nor it implies ($\mathcal{S}(f_\theta, \mathbf{X})_i \perp Z$ for $i \in [1, p]$).*

Proof. Consider $f_\theta(\mathbf{X}) = \mathbf{X}_1 - \mathbf{X}_2$ with $\mathbf{X}_1, \mathbf{X}_2 \sim \texttt{Ber}(0.5)$ and $Z = 1$ if $\mathbf{X}_1 = \mathbf{X}_2$, and $Z = 0$ otherwise. Hence $Z \sim \texttt{Ber}(0.5)$. We have $\mathcal{S}(f_\theta, \mathbf{X}_1) = \mathbf{X}_1 \perp Z$ and $\mathcal{S}(f, \mathbf{X}_2) = -\mathbf{X}_2 \perp Z$. However, $f_\theta(\mathbf{X}) \not\perp Z$, e.g., $P(Z = 0 | f_\theta(\mathbf{X}) = 0) = 1 \neq 0.5$. Example 3.2 illustrates a case where $f_\theta(\mathbf{X}) \perp Z$ yet $\mathcal{S}(f_\theta, \mathbf{X})_1$ and $\mathcal{S}(f_\theta, \mathbf{X})_2$ are not independent of Z. □

Our approach to ET considers the independence of the *multivariate* distribution of $\mathcal{S}(f, \mathbf{X})$ with respect to Z, rather than the independence of each marginal distribution $\mathcal{S}(f_\theta, \mathbf{X})_i \perp Z$. With our definition, we obtain a sufficient condition for DP, as shown in Lemma 3.6.

Chang et al. (2023) follows a similar approach but from the perspective of discovering which subgroup $z \in Z$ exhibits the largest importance disparity relative to a feature $\mathbf{X}_i$.

Definition 3.12. *(Feature Importance Disparity)* Assume Z is discrete, not necessarily binary. The feature importance disparity of $z \in Z$ relatively to a feature $\mathbf{X}_i$ is defined as:

$$\texttt{FID}(z, i) = |E[\mathcal{S}(f_\theta, \mathbf{X})_i | Z = z] - E[\mathcal{S}(f_\theta, \mathbf{X})_i]|$$

From an alignment perspective, the authors of this method don't clarify which equality philosophical criteria the proposed metric measures. Moreover, assuming DP as the reference notion, the method shares the same pitfalls illustrated by the lemma above for Lundberg (2020). From the requirements perspective, the definition does not meet *(R4)* proxy discrimination detection due to using the expected value to measure distribution differences.

3.9 Discussion

Alignment of AI Methodology with Equal Treatment Specifications

The methodology's alignment with equal treatment is structured to comply with the stipulated requirements from the identified Table 1.1. Each aspect of our method addresses mandates from this table, ensuring the alignment of our AI tools with the foundational principles of liberal political philosophy and deontological ethics.

- **ET-01 Unlabeled Data:** Our methodology analyses the model behaviour disparities without reliance on labelled data, thereby resonating with both the deontological and liberal perspectives of assessing fairness in the decisions over errors.
- **ET-02 Group Discrimination:** Our metric aligns with ethical standards and societal expectations by emphasizing equal treatment and addressing discrimination based on characteristics non-chosen at birth.
- **ET-03 Measuring Proxy Discrimination:** Our analytical tools are designed to identify and address proxy discrimination, where seemingly neutral features could indirectly generate bias against protected groups.
- **ET-04 Explanation Capability:** Through the implementation of explanation capabilities, our metric offers transparency regarding how decisions are made, providing theoretical and empirical insights into the factors driving any detected biases, hence empowering end-users with the knowledge to understand and rectify potential discrimination.
- **ET-05 Documentation:** Our software and documentation are Python-based hosted in `https://explanationspace.readthedocs.io/`. This aligns with the requirement for accessibility and clarity. Documentation supports the community of developers and researchers, facilitating understanding, deployment, and further innovation.
- **ET-06 Reproducibility:** Our metric is implemented in an open source python package `explanationspace` fosters the reproducibility of fairness evaluations, enabling comparisons across different models or iterations thereof.
- **ET-07 Evaluation:** Providing extensive tutorials and feedback mechanisms by `GitHub` in `https://github.com/cmougan/explanationspace` enriches the user experience, promoting active community participation.

3.10 Chapter Conclusions

In this chapter of the thesis, we have introduced a novel approach for fairness in machine learning by measuring *equal treatment*. While related work reasoned over model predictions to measure *equal outcomes*, our notion of equal treatment is more fine-grained, accounting for the usage of attributes by the model via explanation distributions. Consequently, equal treatment implies equal outcomes, but the converse is not necessarily true, which we confirmed both theoretically and experimentally.

This paper also seeks to improve the understanding of how theoretical concepts of fairness from liberalism-oriented political philosophy and moral deontology aligns with technical measurements. Rather than merely comparing one social group to another based on disparities within decision distributions, our concept of equal treatment takes into account differences through the explanation distribution of all non-protected attributes, which often act as proxies for protected characteristics. Implications warrant further techno-philosophical discussions. Implications warrant further techno-philosophical discussions.

Chapter 4

How Did Distribution Shift Impact the Model?

Machine Learning (ML) theory provides means to forecast the quality of ML models on unseen data, provided that this data is sampled from the same distribution as the data used to train and evaluate the model. If unseen data is sampled from a different distribution than the training data, model quality may deteriorate, making monitoring how the model's behavior changes crucial.

Recent research has highlighted the impossibility of reliably estimating the performance of machine learning models on unseen data sampled from a different distribution in the absence of further assumptions about the nature of the shift Ben-David et al. (2010); Lipton et al. (2018); Garg et al. (2022). State-of-the-art techniques attempt to model statistical distances between the distributions of the training and unseen data Diethe et al. (2019); Labs (2021) or the distributions of the model predictions Garg et al. (2022, 2021); Lu et al. (2023). However, these measures of *distribution shifts* only partially relate to changes of interaction between new data and trained models or they rely on the availability of a causal graph or types of shift assumptions, which limits their applicability. Thus, it is often necessary to go beyond detecting such changes and understand how the feature attribution changes (Kenthapadi et al., 2022; Haug et al., 2022; Mougan and Nielsen, 2023; Diethe et al., 2019).

The field of explainable AI has emerged as a way to understand model decisions Barredo Arrieta et al. (2020); Molnar (2019) and interpret the inner workings of ML models Guidotti et al. (2018b). The core idea of this paper is to go beyond the modeling of distribution shifts and monitor for *explanation shifts* to signal a change of interactions between learned models and dataset features in tabular data. We newly define explanation shift as the statistical comparison between how predictions from training data are explained and how predictions on new data are explained. In summary, our contributions are:

- We propose measures of explanation shifts as a key indicator for investigating the interaction between distribution shifts and learned models.
- We define an *Explanation Shift Detector* that operates on the explanation distributions allowing for more sensitive and explainable changes of interactions between distribution shifts and learned models.
- We compare our monitoring method that is based on explanation shifts with methods that are based on other kinds of distribution shifts. We find that monitoring for explanation shifts results in more sensitive indicators for varying model behavior.
- We release an open-source Python package `skshift`, which implements our "*Explanation Shift Detector*", along usage tutorials for reproducibility.

4.1 A Model for Explanation Shift Detection

Our model for explanation shift detection is sketched in Fig. 4.1. We define it step-by-step as follows:

Definition 4.1 (Explanation distribution)**.** An explanation function $\mathcal{S} : F \times \text{dom}(X) \to \mathbb{R}^p$ maps a model f_θ and data $x \in \mathbb{R}^p$ to a vector of attributions $\mathcal{S}(f_\theta, x) \in \mathbb{R}^p$. We call $\mathcal{S}(f_\theta, x)$ an explanation. We write $\mathcal{S}(f_\theta, \mathcal{D})$ to refer to the empirical *explanation distribution* generated by $\{\mathcal{S}(f_\theta, x) | x \in \mathcal{D}\}$.

We use local feature attribution methods SHAP and LIME as explanation functions $\mathcal{S}$.

Definition 4.2 (Explanation shift)**.** Given a model f_θ learned from $\mathcal{D}^{tr}$, explanation shift with respect to the model f_θ occurs if $\mathcal{S}(f_\theta, \mathcal{D}^{new}_X) \nsim \mathcal{S}(f_\theta, \mathcal{D}^{tr}_X)$.

Definition 4.3 (Explanation shift metrics)**.** Given a measure of statistical distances d, explanation shift is measured as the distance between two explanations of the model f_θ by $d(\mathcal{S}(f_\theta, \mathcal{D}^{tr}_X), \mathcal{S}(f_\theta, \mathcal{D}^{new}_X))$.

We follow Lopez et al. Lopez-Paz and Oquab (2017) to define an explanation shift metrics based on a two-sample test classifier. We proceed as depicted in Figure 4.1. To counter overfitting, given the model f_θ trained on $\mathcal{D}^{\text{tr}}$, we compute explanations $\{\mathcal{S}(f_\theta, x) | x \in \mathcal{D}^{\text{val}}_X\}$ on an in-distribution validation data set $\mathcal{D}^{\text{val}}_X$. Given a dataset $\mathcal{D}^{new}_X$, for which the status of in- or out-of-distribution is unknown, we compute its explanations $\{\mathcal{S}(f_\theta, x) | x \in \mathcal{D}^{\text{new}}_X\}$. Then, we construct a two-samples dataset $E = \{(S(f_\theta, x), a_x) | x \in \mathcal{D}^{\text{val}}_X, a_x = 0\} \cup \{(S(f_\theta, x), a_x) | x \in \mathcal{D}^{\text{new}}_X, a_x = 1\}$ and we train a discrimination model $g_\psi : R^p \to \{0, 1\}$ on E, to predict if an explanation should be classified as in-distribution (ID) or out-of-distribution (OOD):

$$\psi = \arg\min_{\tilde{\psi}} \sum_{x \in \mathcal{D}_X^{\text{val}} \cup \mathcal{D}_X^{\text{new}}} \ell(g_{\tilde{\psi}}(\mathcal{S}(f_\theta, x)), a_x), \tag{4.1}$$

where ℓ is a classification loss function (e.g. cross-entropy). g_ψ is our two-sample test classifier, based on which AUC yields a test statistic that measures the distance between the D_X^{tr} explanations and the explanations of new data D_X^{new}.

Explanation shift detection allows us to detect *that* a novel dataset D^{new} changes the model's behavior. Beyond recognizing explanation shift, using feature attributions for the model g_ψ, we can interpret *how* the features of the novel dataset D_X^{new} interact differently with model f_θ than the features of the validation dataset D_X^{val}. These features are to be considered for model monitoring and for classifying new data as out-of-distribution.

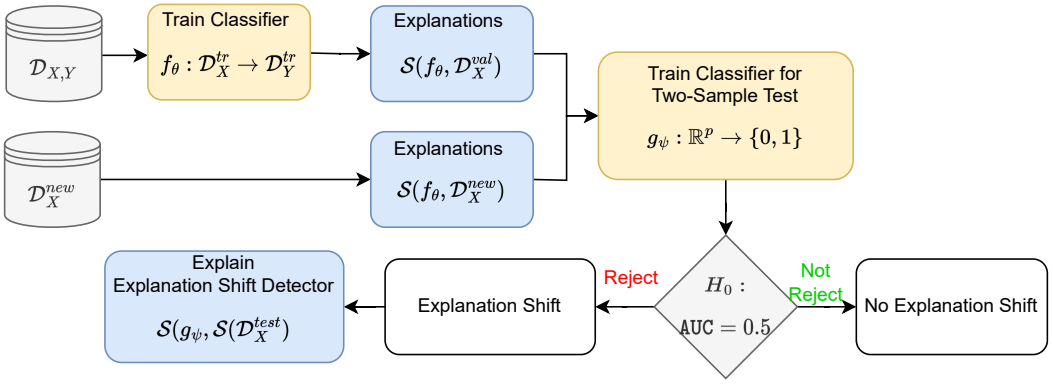

FIGURE 4.1: Our model for explanation shift detection. The model f_θ is trained on $\mathcal{D}^{tr}$ implying explanations for distributions $\mathcal{D}_X^{val}, \mathcal{D}_X^{new}$. The AUC of the two-sample test classifier g_ψ decides for or against explanation shift. If an explanation shift occurred, it could be explained which features of the $\mathcal{D}_X^{new}$ deviated in f_θ compared to $\mathcal{D}_X^{val}$.

4.2 Relationships between Common Distribution Shifts and Explanation Shifts

This section analyses and compares data shifts, prediction shifts, with explanation shifts.

4.2.1 Explanation Shift vs Data Shift

4.2.1.1 Covariate Shift

One type of distribution shift that is challenging to detect comprises cases where the univariate distributions for each feature j are equal between the source $\mathcal{D}_X^{tr}$ and the unseen dataset $\mathcal{D}_X^{new}$, but where interdependencies among different features change. Multi-covariance statistical testing is a hard taks with high sensitivity that can lead to false positives. The following example demonstrates that Shapley values account for co-variate interaction changes while a univariate statistical test will provide false negatives.

Example 4.1. *(**Covariate Shift**) Let* $D^{tr} \sim N\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma^2_{X_1} & 0\\ 0 & \sigma^2_{X_2}\end{bmatrix}\right) \times Y$. *We fit a linear model* $f_\theta(x_1, x_2) = \gamma + a \cdot x_1 + b \cdot x_2$. *If* $\mathcal{D}_X^{new} \sim N\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma^2_{X_1} & \rho\sigma_{X_1}\sigma_{X_2}\\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma^2_{X_2}\end{bmatrix}\right)$, *then* $\mathbf{P}(\mathcal{D}^{tr}_{X_1})$ *and* $\mathbf{P}(\mathcal{D}^{tr}_{X_2})$ *are identically distributed with* $\mathbf{P}(\mathcal{D}^{new}_{X_1})$ *and* $\mathbf{P}(\mathcal{D}^{new}_{X_2})$, *respectively, while this does not hold for the corresponding* $\mathcal{S}_j(f_\theta, \mathcal{D}^{tr}_X)$ *and* $\mathcal{S}_j(f_\theta, \mathcal{D}^{new}_X)$.

$$\mathcal{S}_1(f_\theta, x) = a(x_1 - \mu_1) \tag{4.2}$$

$$\mathcal{S}_1(f_\theta, x^{new}) = \tag{4.3}$$

$$= \frac{1}{2}[\text{val}(\{1,2\}) - \text{val}(\{2\})] + \frac{1}{2}[\text{val}(\{1\}) - \text{val}(\emptyset)] \tag{4.4}$$

$$\text{val}(\{1,2\}) = E[f_\theta | X_1 = x_1, X_2 = x_2] = ax_1 + bx_2 \tag{4.5}$$

$$\text{val}(\emptyset) = E[f_\theta] = a\mu_1 + b\mu_2 \tag{4.6}$$

$$\text{val}(\{1\}) = E[f_\theta(x) | X_1 = x_1] + b\mu_2 \tag{4.7}$$

$$\text{val}(\{1\}) = \mu_1 + \rho\frac{\rho_{x_1}}{\sigma_{x_2}}(x_1 - \sigma_1) + b\mu_2 \tag{4.8}$$

$$\text{val}(\{2\}) = \mu_2 + \rho\frac{\sigma_{x_2}}{\sigma_{x_1}}(x_2 - \mu_2) + a\mu_1 \tag{4.9}$$

$$\Rightarrow \mathcal{S}_1(f_\theta, x^{new}) \neq a(x_1 - \mu_1) \tag{4.10}$$

4.2.1.2 Uninformative Features

False positives frequently occur in out-of-distribution data detection when a statistical test recognizes differences between a source distribution and a new distribution, thought the differences do not affect the model behavior Grinsztajn et al. (2022b); Huyen (2022). Shapley values satisfy the *Uninformativeness* property, where a feature j that does not change the predicted value has a Shapley value of 0 (Property 2.2.1.2).

Example 4.2. ***Unused features*** *Let* $\mathcal{D}^{tr} = \{(x_0^{tr}, y_0^{tr}) \ldots, (x_n^{tr}, y_n^{tr})\}$ *with predictors* $\mathcal{D}_X^{tr} \sim X_1 \times X_2 \times X_3 = N(\mu, \text{diag}(c))$*, and create a synthetic target* $y_i = a_0 + a_1 \cdot x_{i,1} + a_2 \cdot x_{i,2} + \epsilon$ *that is independent of* X_3*. As new data we have* $\mathcal{D}_X^{new} \sim X_1^{new} \times X_2^{new} \times X_3^{new} = N(\mu, \text{diag}(c'))$*, where* c *and* c′ *are an identity matrix of order three and* $\mu = (\mu_1, \mu_2, \mu_3)$*. We train a linear regression* f_θ *on* $\mathcal{D}_{X,Y}^{tr}$*, with coefficients* a_0, a_1, a_2, a_3*. Then if* $\mu_3' \neq \mu_3$ *or* $c_3' \neq c_3$*, then* $\mathbf{P}(\mathcal{D}_{X_3})$ *can be different from* $\mathbf{P}(\mathcal{D}_{X_3}^{new})$ *but* $\mathcal{S}_3(f_\theta, \mathcal{D}_X^{tr}) = \mathcal{S}_3(f_\theta, \mathcal{D}_X^{new})$

$$\mathcal{D}_{X_3} \sim N(\mu_3, c_3), \mathcal{D}_{X_3}^{new} \sim N(\mu_3^{'}, c_3^{'}) \tag{4.11}$$

$$\text{If} \quad \mu_3^{'} \neq \mu_3 \quad \text{or} \quad c_3^{'} \neq c_3 \rightarrow P(X_3) \neq P(X_3^{new}) \tag{4.12}$$

$$\mathcal{S}(f_\theta, X) = \left(\begin{bmatrix} a_1(X_1 - \mu_1) \\ a_2(X_2 - \mu_2) \\ a_3(X_3 - \mu_3) \end{bmatrix} \right) = \left(\begin{bmatrix} a_1(X_1 - \mu_1) \\ a_2(X_2 - \mu_2) \\ 0 \end{bmatrix} \right) \tag{4.13}$$

$$\mathcal{S}_3(f_\theta, \mathcal{D}_X) = \mathcal{S}_3(f_\theta, \mathcal{D}_X^{new}) \tag{4.14}$$

4.2.2 Explanation Shift vs Prediction Shift

Analyses of the explanations detect distribution shifts that interact with the model. In particular, if a prediction shift occurs, the explanations produced are also shifted.

Proposition 1. Given a model $f_\theta : \mathcal{D}_X \rightarrow \mathcal{D}_Y$. If $f_\theta(x') \neq f_\theta(x)$, then $\mathcal{S}(f_\theta, x') \neq \mathcal{S}(f_\theta, x)$.

By efficiency property of the Shapley values Aas et al. (2021) (equation ((2.3))), if the prediction between two instances is different, then they differ in at least one component of their explanation vectors.

The opposite direction does not always hold:

Example 4.3. ***(Explanation shift not affecting prediction distribution)*** *Given* $\mathcal{D}^{tr}$ *is generated from* $(X_1 \times X_2 \times Y), X_1 \sim U(0,1), X_2 \sim U(1,2), Y = X_1 + X_2 + \epsilon$ *and thus the optimal model is* $f(x) = x_1 + x_2$*. If* $\mathcal{D}^{new}$ *is generated from* $X_1^{new} \sim U(1,2), X_2^{new} \sim U(0,1), \quad Y^{new} = X_1^{new} + X_2^{new} + \epsilon$*, the prediction distributions are identical* $f_\theta(\mathcal{D}_X^{tr}), f_\theta(\mathcal{D}_X^{new}) \sim U(1,3)$*, but explanation distributions are different* $S(f_\theta, \mathcal{D}_X^{tr}) \nsim S(f_\theta, \mathcal{D}_X^{new})$*, because* $\mathcal{S}_i(f_\theta, x) = \alpha_i \cdot x_i$*.*

Thus, an explanation shift does not always imply a prediction shift.

4.2.3 Explanation Shift vs Concept Shift

One of the most challenging types of distribution shift to detect are cases where distributions are equal between source and unseen data-set $\mathbf{P}(\mathcal{D}_X^{tr}) = \mathbf{P}(\mathcal{D}_X^{new})$ and the target

variable $\mathbf{P}(\mathcal{D}_Y^{tr}) = \mathbf{P}(\mathcal{D}_Y^{new})$ and what changes are the relationships that features have with the target $\mathbf{P}(\mathcal{D}_Y^{tr}|\mathcal{D}_X^{tr}) \neq \mathbf{P}(\mathcal{D}_Y^{new}|\mathcal{D}_X^{new})$, this kind of distribution shift is also known as concept drift or posterior shift (Huyen, 2022) and is especially difficult to notice, as it requires labeled data to detect. The following example compares how the explanations change for two models fed with the same input data and different target relations.

Concept shift comprises cases where the covariates retain a given distribution, but their relationship with the target variable changes (cf. Section 2.5.2). This example shows the negative result that concept shift cannot be indicated by the detection of explanation shift.

Example 4.4. ***Concept shift*** *Let $\mathcal{D}_X = (X_1, X_2) \sim N(\mu, I)$, and $\mathcal{D}_X^{new} = (X_1^{new}, X_2^{new}) \sim N(\mu, I)$, where I is an identity matrix of order two and $\mu = (\mu_1, \mu_2)$. We now create two synthetic targets $Y = a + \alpha \cdot X_1 + \beta \cdot X_2 + \epsilon$ and $Y^{new} = a + \beta \cdot X_1 + \alpha \cdot X_2 + \epsilon$. Let f_θ be a linear regression model trained on $f_\theta : \mathcal{D}_X \rightarrow \mathcal{D}_Y)$ and h_ϕ another linear model trained on $h_\phi : \mathcal{D}_X^{new} \rightarrow \mathcal{D}_Y^{new})$. Then $\mathbf{P}(f_\theta(X)) = \mathbf{P}(h_\phi(X^{new}))$, $P(X) = \mathbf{P}(X^{new})$ but $\mathcal{S}(f_\theta, X) \neq \mathcal{S}(h_\phi, X)$.*

$$X \sim N(\mu, \sigma^2 \cdot I), X^{new} \sim N(\mu, \sigma^2 \cdot I) \tag{4.15}$$

$$\rightarrow P(\mathcal{D}_X) = P(\mathcal{D}_X^{new}) \tag{4.16}$$

$$Y \sim a + \alpha N(\mu, \sigma^2) + \beta N(\mu, \sigma^2) + N(0, \sigma'^2) \tag{4.17}$$

$$Y^{new} \sim a + \beta N(\mu, \sigma^2) + \alpha N(\mu, \sigma^2) + N(0, \sigma'^2) \tag{4.18}$$

$$\rightarrow P(\mathcal{D}_Y) = P(\mathcal{D}_Y^{new}) \tag{4.19}$$

$$\mathcal{S}(f_\theta, \mathcal{D}_X) = \begin{pmatrix} \alpha(X_1 - \mu_1) \\ \beta(X_2 - \mu_2) \end{pmatrix} \sim \begin{pmatrix} N(\mu_1, \alpha^2\sigma^2) \\ N(\mu_2, \beta^2\sigma^2) \end{pmatrix} \tag{4.20}$$

$$\mathcal{S}(h_\phi, \mathcal{D}_X) = \begin{pmatrix} \beta(X_1 - \mu_1) \\ \alpha(X_2 - \mu_2) \end{pmatrix} \sim \begin{pmatrix} N(\mu_1, \beta^2\sigma^2) \\ N(\mu_2, \alpha^2\sigma^2) \end{pmatrix} \tag{4.21}$$

$$\text{If} \quad \alpha \neq \beta \rightarrow \mathcal{S}(f_\theta, \mathcal{D}_X) \neq \mathcal{S}(h_\phi, \mathcal{D}_X) \tag{4.22}$$

In general, concept shift cannot be detected because $\mathcal{D}_Y^{new}$ is unknown Garg et al. (2022). Some research studies have made specific assumptions about the conditional $\frac{P(\mathcal{D}^{new})}{P(\mathcal{D}_X^{new})}$ in order to monitor models and detect distribution shift Lu et al. (2023); Alvarez et al. (2023).

4.2.4 Analytical Experiments on Synthetic Data with Non-Linear Models

This experimental section explores the detection of distribution shifts in the previous synthetic examples but using non linear models.

4.2.4.1 Detecting multivariate shift

Given two bivariate normal distributions $\mathcal{D}_X = (X_1, X_2) \sim N\left(0, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$ and $\mathcal{D}_X^{new} = (X_1^{new}, X_2^{new}) \sim N\left(0, \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}\right)$, then, for each feature j the underlying distribution is equally distributed between $\mathcal{D}_X$ and $\mathcal{D}_X^{new}$, $\forall j \in \{1, 2\} : P(\mathcal{D}_{X_j}) = P(\mathcal{D}_{X_j}^{new})$, and what is different are the interaction terms between them. We now create a synthetic target $Y = X_1 \cdot X_2 + \epsilon$ with $\epsilon \sim N(0, 0.1)$ and fit a gradient boosting decision tree $f_\theta(\mathcal{D}_X)$. Then we compute the SHAP explanation values for $\mathcal{S}(f_\theta, \mathcal{D}_X)$ and $\mathcal{S}(f_\theta, \mathcal{D}_X^{new})$

TABLE 4.1: Displayed results are the one-tailed p-values of the Kolmogorov-Smirnov test comparison between two underlying distributions. Small p-values indicate that compared distributions would be very unlikely to be equally distributed. SHAP values correctly indicate the interaction changes that individual distribution comparisons cannot detect

Comparison	**p-value**	**Conclusions**
$\mathbf{P}(\mathcal{D}_{X_1})$, $\mathbf{P}(\mathcal{D}_{X_1}^{new})$	0.33	Not Distinct
$\mathbf{P}(\mathcal{D}_{X_2})$, $\mathbf{P}(\mathcal{D}_{X_2}^{new})$	0.60	Not Distinct
$\mathcal{S}_1(f_\theta, \mathcal{D}_X)$, $\mathcal{S}_1(f_\theta, \mathcal{D}_X^{new})$	3.9e−153	Distinct
$\mathcal{S}_2(f_\theta, \mathcal{D}_X)$, $\mathcal{S}_2(f_\theta, \mathcal{D}_X^{new})$	2.9e−148	Distinct

Having drawn 50,000 samples from both $\mathcal{D}_X$ and $\mathcal{D}_X^{new}$, in Table 4.1, we evaluate whether changes in the input data distribution or in the explanations are able to detect changes in covariate distribution. For this, we compare the one-tailed p-values of the Kolmogorov-Smirnov test between the input data distribution and the explanations distribution. Explanation shift correctly detects the multivariate distribution change that univariate statistical testing can not detect.

4.2.4.2 Detecting concept shift

As mentioned before, concept shift cannot be detected if new data comes without target labels. If new data is labelled, the explanation shift can still be a useful technique for detecting concept shifts.

Given a bivariate normal distribution $\mathcal{D}_X = (X_1, X_2) \sim N(1, I)$ where I is an identity matrix of order two. We now create two synthetic targets $Y = X_1^2 \cdot X_2 + \epsilon$ and $Y^{new} = X_1 \cdot X_2^2 + \epsilon$ and fit two machine learning models $f_\theta : \mathcal{D}_X \to \mathcal{D}_Y)$ and $h_Y : \mathcal{D}_X \to \mathcal{D}_Y^{new})$. Now we compute the SHAP values for $\mathcal{S}(f_\theta, \mathcal{D}_X)$ and $\mathcal{S}(h_Y, \mathcal{D}_X)$

In Table 4.2, we see how the distribution shifts are not able to capture the change in the model behavior while the SHAP values are different. The "Distinct/Not distinct" conclusion is based on the one-tailed p-value of the Kolmogorov-Smirnov test with a 0.05 threshold drawn out of 50,000 samples for both distributions. As in the synthetic

TABLE 4.2: Distribution comparison for synthetic concept shift. Displayed results are the one-tailed p-values of the Kolmogorov-Smirnov test comparison between two underlying distributions

Comparison	**Conclusions**
$\mathbf{P}(\mathcal{D}_X)$, $\mathbf{P}(\mathcal{D}_X^{new})$	Not Distinct
$\mathbf{P}(\mathcal{D}_Y)$, $\mathbf{P}(\mathcal{D}_Y^{new})$	Not Distinct
$\mathbf{P}(f_\theta(\mathcal{D}_X))$, $\mathbf{P}(h_Y(\mathcal{D}_X^{new}))$	Not Distinct
$\mathbf{P}(\mathcal{S}(f_\theta, \mathcal{D}_X))$, $\mathbf{P}(\mathcal{S}(h_Y, \mathcal{D}_X))$	Distinct

example, in table 4.2 SHAP values can detect a relational change between $\mathcal{D}_X$ and $\mathcal{D}_Y$, even if both distributions remain equivalent.

4.2.4.3 Uninformative features on synthetic data

To have an applied use case of the synthetic example from the methodology section, we create a three-variate normal distribution $\mathcal{D}_X = (X_1, X_2, X_3) \sim N(0, I_3)$, where I_3 is an identity matrix of order three. The target variable is generated $Y = X_1 \cdot X_2 + \epsilon$ being independent of X_3. For both, training and test data, $50,000$ samples are drawn. Then out-of-distribution data is created by shifting X_3, which is independent of the target, on test data $\mathcal{D}_{X_3}^{new} = \mathcal{D}_{X_3}^{te} + 1$.

TABLE 4.3: Distribution comparison when modifying a random noise variable on test data. The input data shifts while explanations and predictions do not.

Comparison	**Conclusions**
$\mathbf{P}(\mathcal{D}_{X_3}^{te})$, $\mathbf{P}(\mathcal{D}_{X_3}^{new})$	Distinct
$f_\theta(\mathcal{D}_X^{te})$, $f_\theta(\mathcal{D}_X^{new})$	Not Distinct
$\mathcal{S}(f_\theta, \mathcal{D}_X^{te})$, $\mathcal{S}(f_\theta, \mathcal{D}_X^{new})$	Not Distinct

In Table 4.3, we see how an unused feature has changed the input distribution, but the explanation distributions and performance evaluation metrics remain the same. The "Distinct/Not Distinct" conclusion is based on the one-tailed p-value of the Kolmogorov-Smirnov test drawn out of $50,000$ samples for both distributions.

4.2.4.4 Explanation shift that does not affect the prediction

In this case we provide a situation when we have changes in the input data distributions that affect the model explanations but do not affect the model predictions due to positive and negative associations between the model predictions and the distributions cancel out producing a vanishing correlation in the mixture of the distribution (Yule's effect 4.2.2).

We create a train and test data by drawing $50,000$ samples from a bi-uniform distribution $X_1 \sim U(0,1), \quad X_2 \sim U(1,2)$ the target variable is generated by $Y = X_1 + X_2$

where we train our model f_θ. Then if out-of-distribution data is sampled from $X_1^{new} \sim U(1,2)$, $X_2^{new} \sim U(0,1)$

TABLE 4.4: Distribution comparison over how the change on the contributions of each feature can cancel out to produce an equal prediction (cf. Section 4.2.2), while explanation shift will detect this behaviour changes on the predictions will not.

Comparison	**Conclusions**
$f(\mathcal{D}_X^{te}), f(\mathcal{D}_X^{new})$	Not Distinct
$\mathcal{S}(f_\theta, \mathcal{D}_{X_2}^{te}), \mathcal{S}(f_\theta, \mathcal{D}_{X_2}^{new})$	Distinct
$\mathcal{S}(f_\theta, \mathcal{D}_{X_1}^{te}), \mathcal{S}(f_\theta, \mathcal{D}_{X_1}^{new})$	Distinct

In Table 4.4, we see how an unused feature has changed the input distribution, but the explanation distributions and performance evaluation metrics remain the same. The "Distinct/Not Distinct" conclusion is based on the one-tailed p-value of the Kolmogorov-Smirnov test drawn out of $50,000$ samples for both distributions.

4.3 Empirical Evaluation

We evaluate the effectiveness of explanation shift detection on tabular data by comparing it against methods from the literature, which are all based on discovering distribution shifts. For this comparison, we systematically vary models f, model parametrizations θ, and input data distributions $\mathcal{D}_X$. We complement core experiments described in this section by adding further experimental results in the appendix that *(i)* add details on experiments with synthetic data (Appendix 4.2.4), *(ii)* add experiments on further natural datasets (Appendix 4.3.5.1), *(iii)* exhibit a larger range of modeling choices (Appendix 4.4),and *(iv)* contrast our SHAP-based method against the use of LIME, an alternative explanation approach (Appendix 4.5). Core observations made in this section will only be confirmed and refined but not countered in the appendix.

4.3.1 Baseline Methods and Datasets

Baseline Methods. We compare our method of explanation shift detection (Section 4.1) with several methods that aim to detect that input data is out-of-distribution: *(B1)* statistical Kolmogorov Smirnov test on input data (Rabanser et al., 2019), *(B2)* prediction shift detection by Wasserstein distance (Lu et al., 2023), *(B3)* NDCG-based test of feature importance between the two distributions (Nigenda et al., 2022), *(B4)* prediction shift detection by Kolmogorov-Smirnov test (Diethe et al., 2019), and *(B5)* model agnostic uncertainty estimation (Mougan and Nielsen, 2023; Kim et al., 2020). All Distribution Shift Metrics are scaled between 0 and 1. We also compare against Classifier Two-Sample Test (Lopez-Paz and Oquab, 2017) on different distributions as discussed

in Section 4.2, viz. *(B6)* classifier two-sample test on input distributions (g_ϕ) following Barrabés et al. (2023) and *(B7)* classifier two-sample test on the predictions distributions (g_Y):

$$\phi = \arg\min_{\tilde{\phi}} \sum_{x \in \mathcal{D}_X^{val} \cup \mathcal{D}_X^{new}} \ell(g_{\tilde{\phi}}(x)), a_x) \tag{4.23}$$

$$Y = \arg\min_{\tilde{Y}} \sum_{x \in \mathcal{D}_X^{val} \cup \mathcal{D}_X^{new}} \ell(g_{\tilde{Y}}(f_\theta(x)), a_x) \tag{4.24}$$

Datasets. In the paper's main body, we base our comparisons on the UCI Adult Income dataset Dua and Graff (2017a) and on synthetic data. In the Appendix, we extend experiments to several other datasets, which confirm our findings: ACS Travel Time (Ding et al., 2021b), ACS Employment, Stackoverflow dataset (Stackoverflow, 2019).

4.3.2 Synthetic Data

Our first experiment on synthetic data showcases the two main contributions of our method: (i) being more sensitive to changes in the model than prediction shift and input shift and (ii) accounting for its drivers. We first generate a synthetic dataset $\mathcal{D}^\rho$, with a parametrized multivariate shift between (X_1, X_2), where ρ is the correlation coefficient, and an extra variable $X_3 = N(0,1)$ and generate our target $Y = X_1 \cdot X_2 + X_3$. We train the f_θ on $\mathcal{D}^{tr,\rho=0}$ using a gradient boosting decision tree, while for $g_\psi : \mathcal{S}(f_\theta, \mathcal{D}_X^{val,\rho}) \rightarrow \{0,1\}$, we train on different datasests with different values of ρ. For g_ψ we use a logistic regression. In Appendix 4.4, we benchmark other models f_θ and detectors g_ψ.

The above image in Figure 4.2 compares our approach against C2ST on input data distribution*(B6)* and on the predictions distribution *(B7)* different data distributions, for detecting multi-covariate shifts on different distributions. In our covariate experiment, we observed that using the explanation shift led to higher sensitivity towards detecting distribution shift. We interpret the results with the efficiency property of the Shapley values, which decomposes the vector $f_\theta(\mathcal{D}_X)$ into the matrix $\mathcal{S}(f_\theta, \mathcal{D}_X)$. Moreover, we can identify the features that cause the drift by extracting the coefficients of g_ψ, providing global and local explainability.

The below image in Figure 4.2 features the same setup compared to the other out-of-distribution detection methods (B1-B5). Table 4.5 quantitatively evaluates how the baselines correlate with the covariate correlation coefficient (ρ). One can see how our method behaves favourably compared to the others.

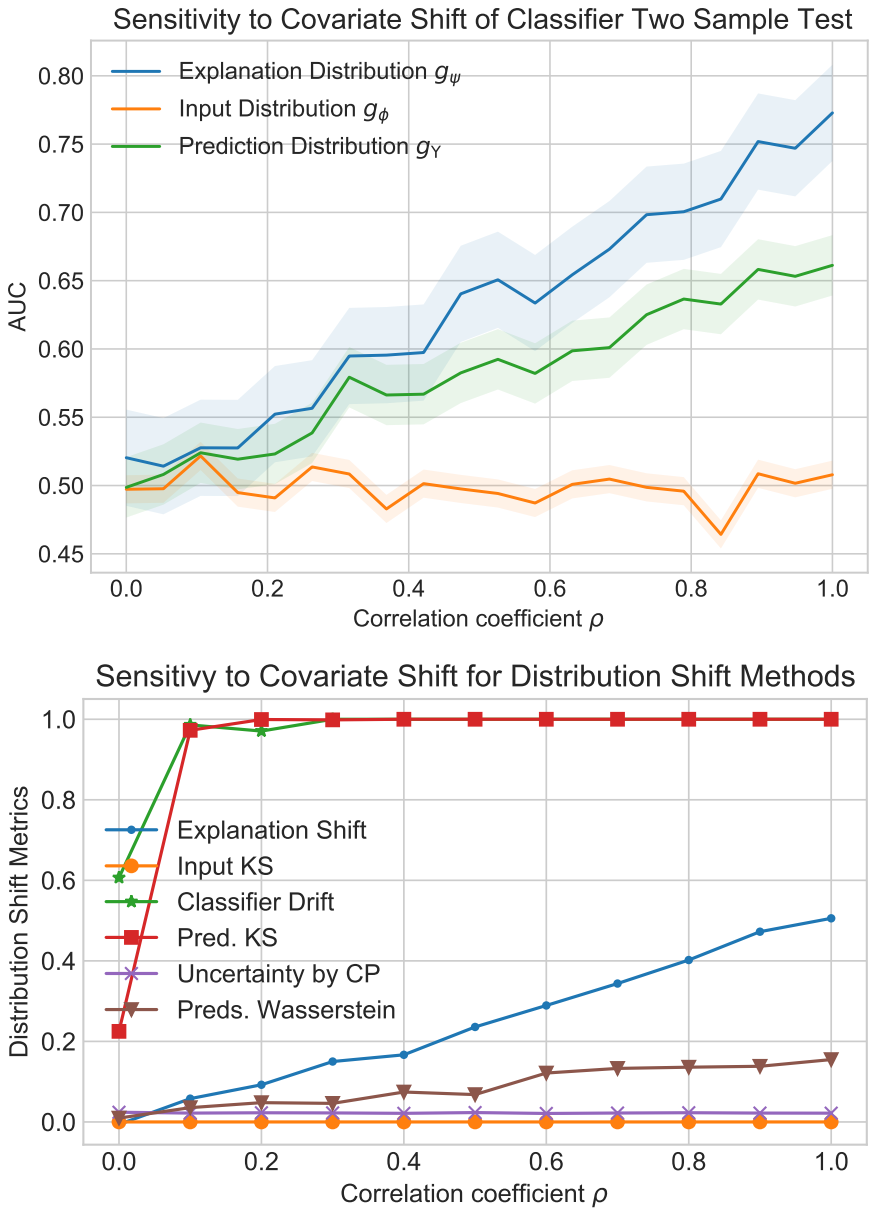

FIGURE 4.2: In the figure above, we compare the Classifier Two-Sample Test on explanation distribution (ours) versus input distribution *(B6)* and prediction distribution *(B7)*. Explanation distribution shows the highest sensitivity. The figure below, related work comparison of distribution shift methods*(B1-B5)*, as the experimental setup has a gradual distribution shift, good indicators should follow a progressive steady positive slope, following the correlation coefficient, as our method does. In Table 4.5 we provide a quantitative evaluation.

4.3.3 Novel Group Shift

The distribution shift in this experimental setup is constituted by the appearance of a hitherto unseen group at prediction time (the group information is not present in the training features). We vary the ratio of presence of this unseen group in $\mathcal{D}_X^{new}$ data. The experiment is done with two f_θ models: a gradient-boosting decision tree and a

TABLE 4.5: Pearson Correlation of the correlation coefficient ρ and baseline methods, extending Figure 4.2. Explanation Shift achieves better covariate shift detection on synthetic data.

Baseline	**Pearson Correlation with ρ**
B1 Input KS	0.01
B2 Prediction Wasserstein	0.97
B3 Explanation NDCG	0.52
B4 Prediction KS	0.70
B5 Uncertainty	0.26
B6 C2ST Input	0.18
B7 C2ST Output	0.96
(Ours) Explanation Shift	**0.99**

logistic regression; for g_ψ, we use a logistic regression. Results are presented in Figure 4.3, Figure 4.4 and Table 4.6. Confidence intervals are extracted out of 10 bootstraps. Furthermore, we compare the performance of different algorithms for f_θ and g_ψ in Appendix 4.4.1, and varying hyperparameters in Appendix 4.4.

TABLE 4.6: Pearson correlation between baselines and the ratio of presence of the unseen group. The *Accountable* column relates to how well a method, like the Explanation Shift Detector, is underpinned by theoretical guarantees (e.g., Shapley value properties) and supported by empirical evidence (e.g., synthetic tests and real-world validations), as discussed in Section 4.2.

Baseline	**Pearson Correlation**		**Accountable**
	$f_\theta =$ **Log**	$f_\theta =$ **XGB**	
B1 Input KS	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	✗
B2 Pred. Wass.	0.95 ± 0.02	$\mathbf{0.98 \pm 0.01}$	✗
B3 NDCG	0.37 ± 0.25	0.81 ± 0.10	✗
B4 Pred. KS	0.97 ± 0.02	0.96 ± 0.01	✗
B5 Uncertainty	0.73 ± 0.10	0.74 ± 0.12	✓
B6 C2ST Input	0.95 ± 0.03	0.95 ± 0.03	✗
B7 C2ST Output	0.67 ± 0.13	0.96 ± 0.02	✗
Explanation Shift	$\mathbf{0.98 \pm 0.01}$	$\mathbf{0.98 \pm 0.01}$	✓

This experiment further validates the effectiveness of the "Explanation Shift Detector" in addressing novel group changes in real-world datasets. The method demonstrates consistent performance across multiple datasets, providing valuable insights into the sensitivity of model behavior as previously unseen social groups constitute a larger portion of the prediction data. As shown in Figure 4.4, the proposed approach significantly outperforms Exp. NDCG '(B6)' across four datasets. Unlike Exp. NDCG, which exhibits instability and frequently reaches horizontal asymptotes, the Explanation Shift Detector remains robust and adaptable. These limitations in Exp. NDCG arise from its reliance on feature importance rankings, which lack direct information about the value changes induced by the novel group. In contrast, the Explanation Shift Detector's use of a Classifier Two-Sample Test on the distributions of explanations effectively captures these shifts, ensuring accurate and reliable detection of novel group effects on model behavior.

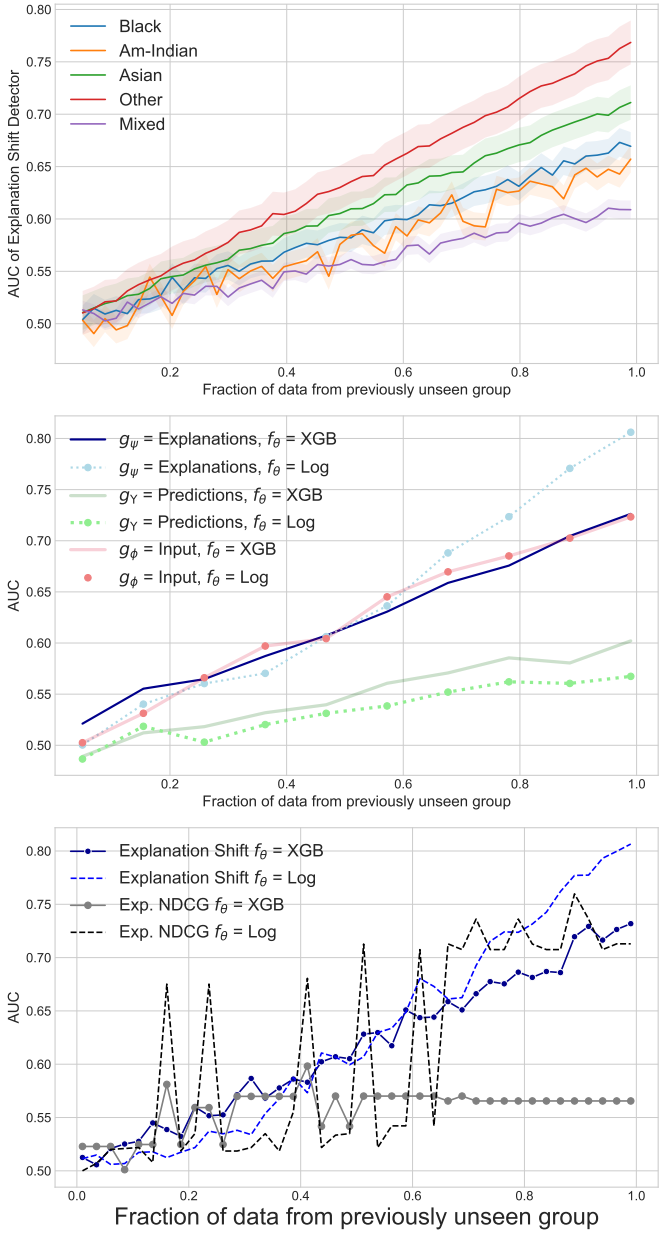

FIGURE 4.3: Novel group shift experiment on the US Income dataset. Sensitivity (AUC) increases with the growing fraction of previously unseen social groups. *Left figure*: The explanation shift indicates that different social groups exhibit varying deviations from the distribution on which the model was trained (White). *Middle Figure*: We vary the model f_θ by training it using both `XGBoost` (solid lines) and Logistic Regression (dots). The novel ethnicity group is Black. We compare Explanation Shift against C2ST on input *(B6)* and output *(B7)*. *Right figure*: Comparison of Explanation Shift against Exp. NDCG *(B4)*. We see how monitoring method*(B4)* is more unstable with a linear model, and with an `xgboost` it erroneously finds a horizontal asymptote. We don't compare against methods relying purely on input data such as *(B1)* as we are changing the model, which they don't take into consideration.

As a takeaway, the Explanation Shift Detector emerges as a highly reliable and accountable approach for detecting and adapting to novel group shifts. It outperforms traditional metrics like Exp. NDCG effectively captures distributional changes in explanations, making it an indispensable tool for robust model monitoring in diverse and

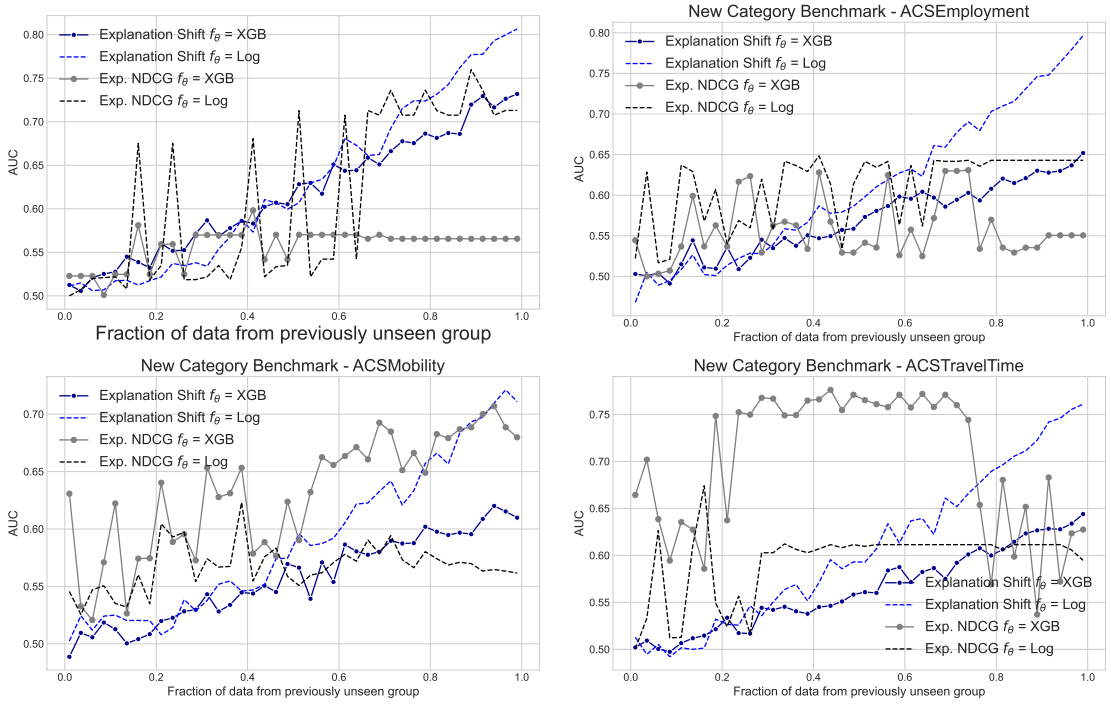

FIGURE 4.4: Novel group shift experiment conducted on the 4 Datasets Comparison to NDCG. Sensitivity (AUC) increases as the proportion of previously unseen social groups grows. As the experimental setup has a gradual distribution shift, ideal indicators should exhibit a steadily increasing slope. However, in all figures, NDCG exhibits saturation and instability. These observations align with the analysis presented in the synthetic experiment section, as discussed in Section 4.3.2 of the main paper

evolving datasets.

4.3.4 StackOverflow Survey Data

In this extended experiment from novel covariate group, we evaluate the new unseen group in the new data over the StackOverflow dataset. As a training data country, we use the United States. The model $f_\theta$used is a gradient-boosting decision tree or logistic regression, and logistic regression is used for the detector. The results show that the AUC of the Explanation Shift Detector varies depending on the quantification of OOD explanations, and it shows more sensitivity concerning model variations than other state-of-the-art techniques.

The dataset used is the StackOverflow annual developer survey, with over 70,000 responses from over 180 countries examining aspects of the developer experience (Stackoverflow, 2019). The data has high dimensionality, leaving it with +100 features after data cleansing and feature engineering. The goal of this task is to predict the total annual compensation.

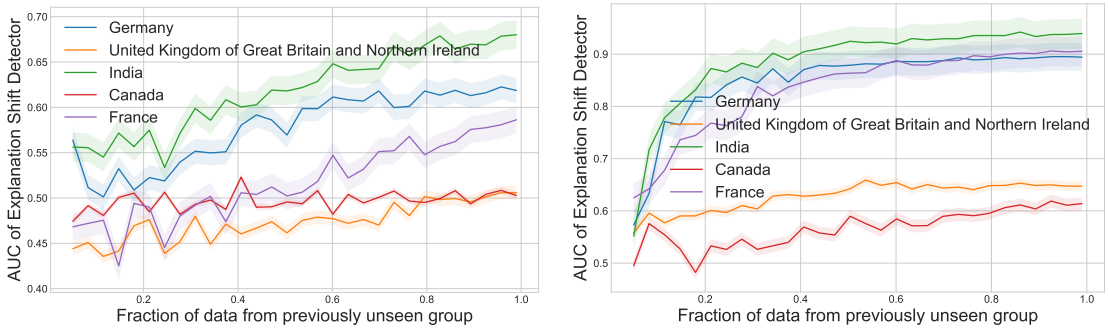

FIGURE 4.5: Both images represent the AUC of the *Explanation Shift Detector* for different countries on the StackOverflow survey dataset under novel group shift. In the left image, the estimator, f_θ, is a gradient boosting decision tree; in the right image, for both cases the detector, g_ψ, is a logistic regression. By changing the type of estimator model, we can see how different types of models are affected differently for the same distribution shift

4.3.5 Geopolitical and Temporal Shift

In this section, we tackle a geopolitical and temporal distribution shift; for this, the training data $\mathcal{D}^{tr}$ for the model f_θ is composed of data from California in 2014 and a $\mathcal{D}^{new}$ for each of the states in 2018. The model g_ψ is trained each time on each state using only the $\mathcal{D}_X^{new}$ in the absence of the label, and a 50/50 random train-test split evaluates its performance. As models, we use `xgboost` as f_θ and logistic regression for the *Explanation Shift Detector* (g_ψ).

We hypothesize that the AUC of the Explanation Shift Detector on new data will be distinct from that on in-distribution data, primarily owing to the distinctive nature of out-of-distribution model explanations. Figure 4.6 illustrates the performance of

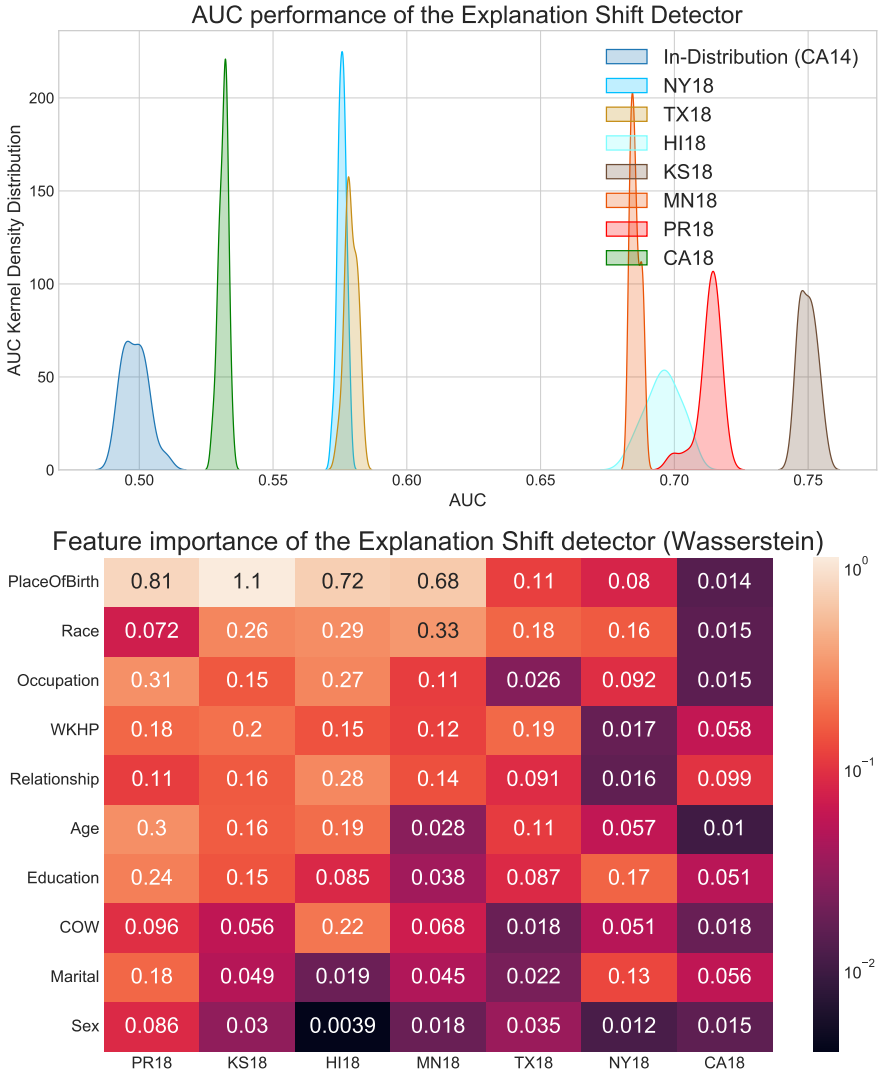

FIGURE 4.6: In the above figure, a comparison of the performance of *Explanation Shift Detector* in different states. In the below figure, the strength analysis of features driving the change in the model, on the y-axis are the features, and on the x-axis are the different states. Explanation shifts allow us to identify how the distribution shift of different features impacted the model.

our method on different data distributions, where the baseline is a ID hold-out set of CA14. The AUC for CA18, where there is only a temporal shift, is the closest to the baseline, and the OOD detection performance is better in the rest of the states. The most disparate state is Puerto Rico (PR18).

Our next objective is to identify the features where the explanations differ between $\mathcal{D}_X^{tr}$ and $\mathcal{D}_X^{new}$ data. To achieve this, we compare the distribution of linear coefficients of the detector between both distributions. We use the Wasserstein distance as a distance measure, generating 1000 in-distribution bootstraps using a 63.2% sampling fraction from California-14 and 1000 bootstraps from other states in 2018. In the below image of Figure 4.6, we observe that for PR18, the most crucial feature is the Place of Birth.

Furthermore, we conduct an across-task evaluation by comparing the performance of the "Explanation Shift Detector" on another prediction task in the Appendix 4.3.5.1. Although some features are present in both prediction tasks, the weights and importance order assigned by the "Explanation Shift Detector" differ. One of this method's advantages is that it identifies differences in distributions and how they relate to the model.

4.3.5.1 Further Experiments on Real Data

In this section, we extend the prediction task of the main body of the paper. The methodology used follows the same structure. We start by creating a distribution shift by training the model f_θ in California in 2014 and evaluating it in the rest of the states in 2018, creating a geopolitical and temporal shift. The model g_θ is trained each time on each state using only the X^{New} in the absence of the label, and its performance is evaluated by a 50/50 random train-test split. As models, we use a gradient boosting decision tree(Chen and Guestrin, 2016b; Prokhorenkova et al., 2018) for f_θ, approximating the Shapley values by TreeExplainer (Lundberg et al., 2020a), and using logistic regression for the *Explanation Shift Detector*.

For further understanding of the meaning of the features the ACS PUMS data dictionary contains a comprehensive list of available variables `https://www.census.gov/programs-surveys/acs/microdata/documentation.html`.

ACS Employment The objective of this task is to determine whether an individual aged between 16 and 90 years is employed or not. The model's performance was evaluated using the AUC metric in different states, except PR18, where the model showed an explanation shift. The explanation shift was observed to be influenced by features such as Citizenship and Military Service. The performance of the model was found to be consistent across most of the states, with an AUC below 0.60. The impact of features such as difficulties in hearing or seeing was negligible in the distribution shift impact on the model. The left figure in Figure 4.7 compares the performance of the Explanation Shift Detector in different states for the ACS Employment dataset.

Additionally, the feature importance analysis for the same dataset is presented in the below figure in Figure 4.7.

ACS Travel Time The goal of this task is to predict whether an individual has a commute to work that is longer than +20 minutes. For this prediction task, the results are different from the previous two cases; the state with the highest OOD score is *KS*18, with the "Explanation Shift Detector" highlighting features as Place of Birth, Race or

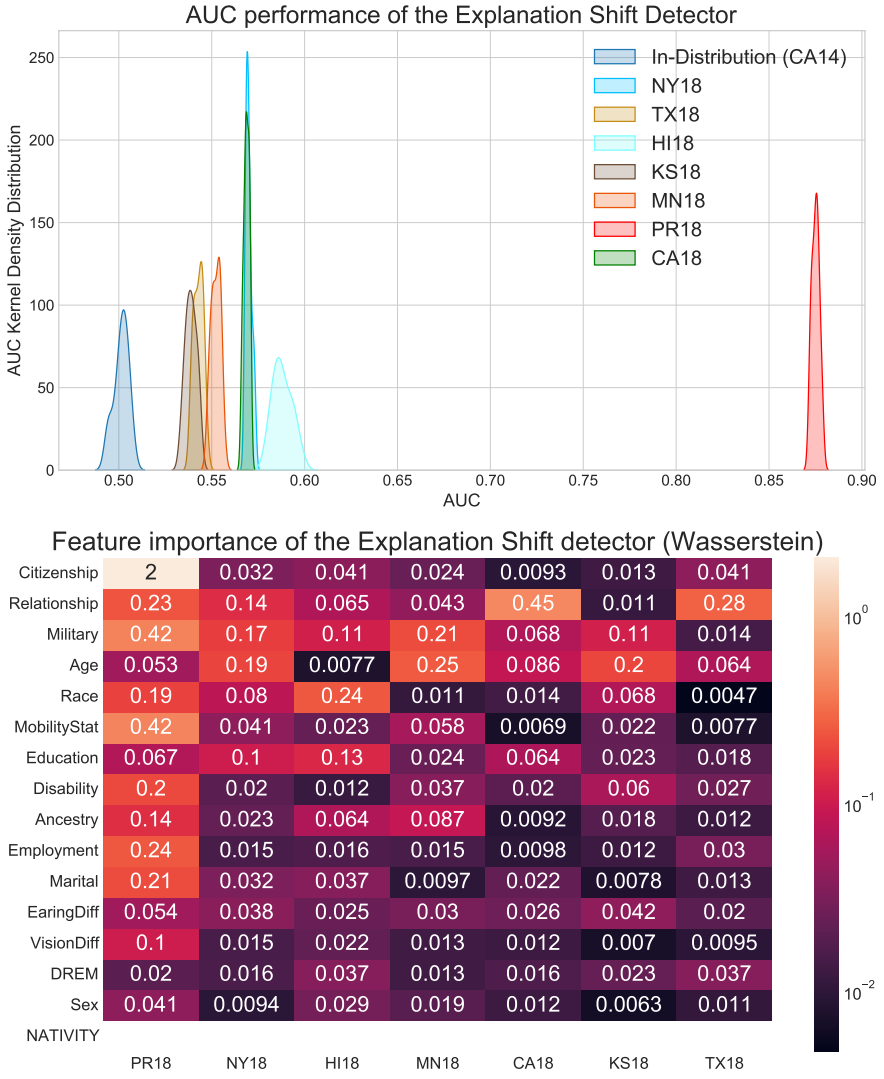

FIGURE 4.7: The above figure shows a comparison of the performance of the Explanation Shift Detector in different states for the ACS Employment dataset. The below figure shows the feature importance analysis for the same dataset.

Working Hours Per Week. The closest state to ID is CA18, where there is only a temporal shift without any geospatial distribution shift.

ACS Mobility The objective of this task is to predict whether an individual between the ages of 18 and 35 had the same residential address as a year ago. This filtering is intended to increase the difficulty of the prediction task, as the base rate for staying at the same address is above 90% for the population (Ding et al., 2021b).

The experiment shows a similar pattern to the ACS Income prediction task (cf. Section 4.6), where the inland US states have an AUC range of $0.55 - 0.70$, while the state of PR18 achieves a higher AUC. For PR18, the model has shifted due to features such as Citizenship, while for the other states, it is Ancestry (Census record of your ancestors'

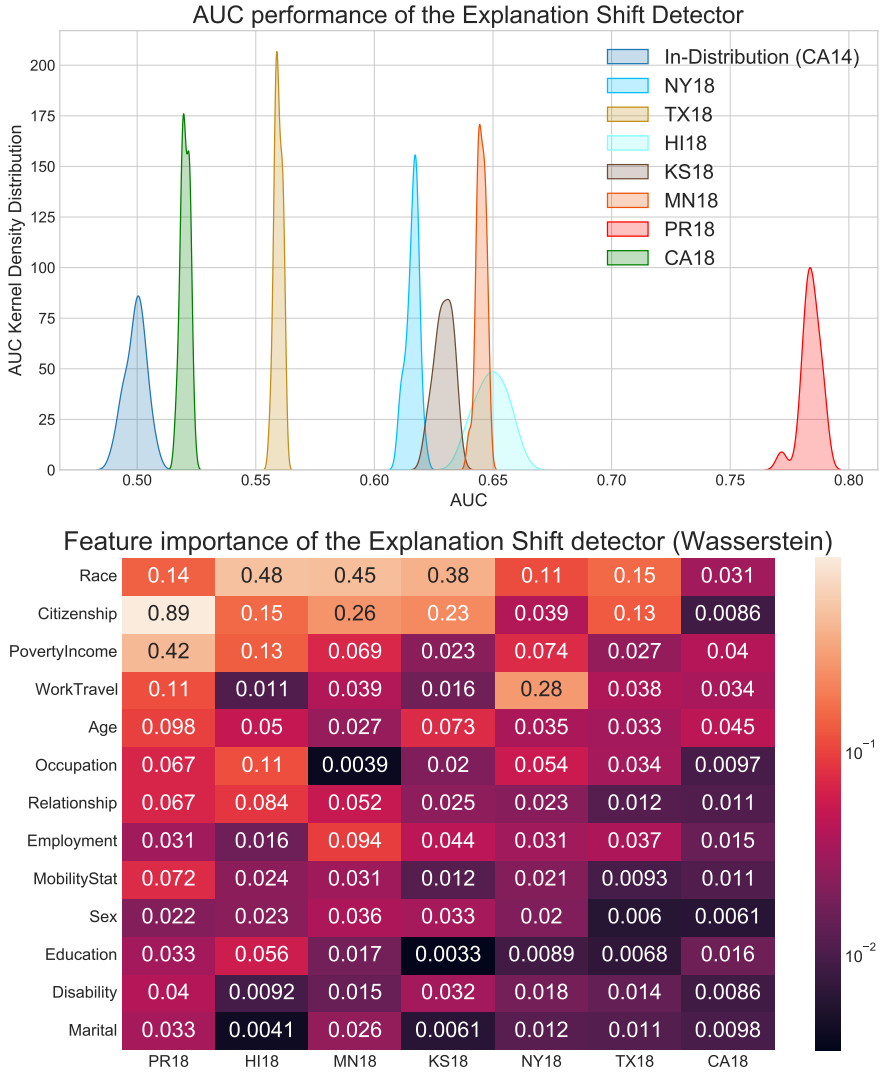

FIGURE 4.8: In the left figure, comparison of the performance of *Explanation Shift Detector*, in different states for the ACS TravelTime prediction task. In the left figure, we can see how the state with the highest OOD AUC detection is KS18 and not PR18 as in other prediction tasks; this difference with respect to the other prediction task can be attributed to "Place of Birth", whose feature attributions the model finds to be more different than in CA14.

lives with details like where they lived, who they lived with, and what they did for a living) that drives the change in the model.

As depicted in Figure 4.9, all states, except for PR18, fall below an AUC of explanation shift detection of 0.70. Protected social attributes, such as Race or Marital status, play an essential role for these states, whereas for PR18, Citizenship is a key feature driving the impact of distribution shift in model.

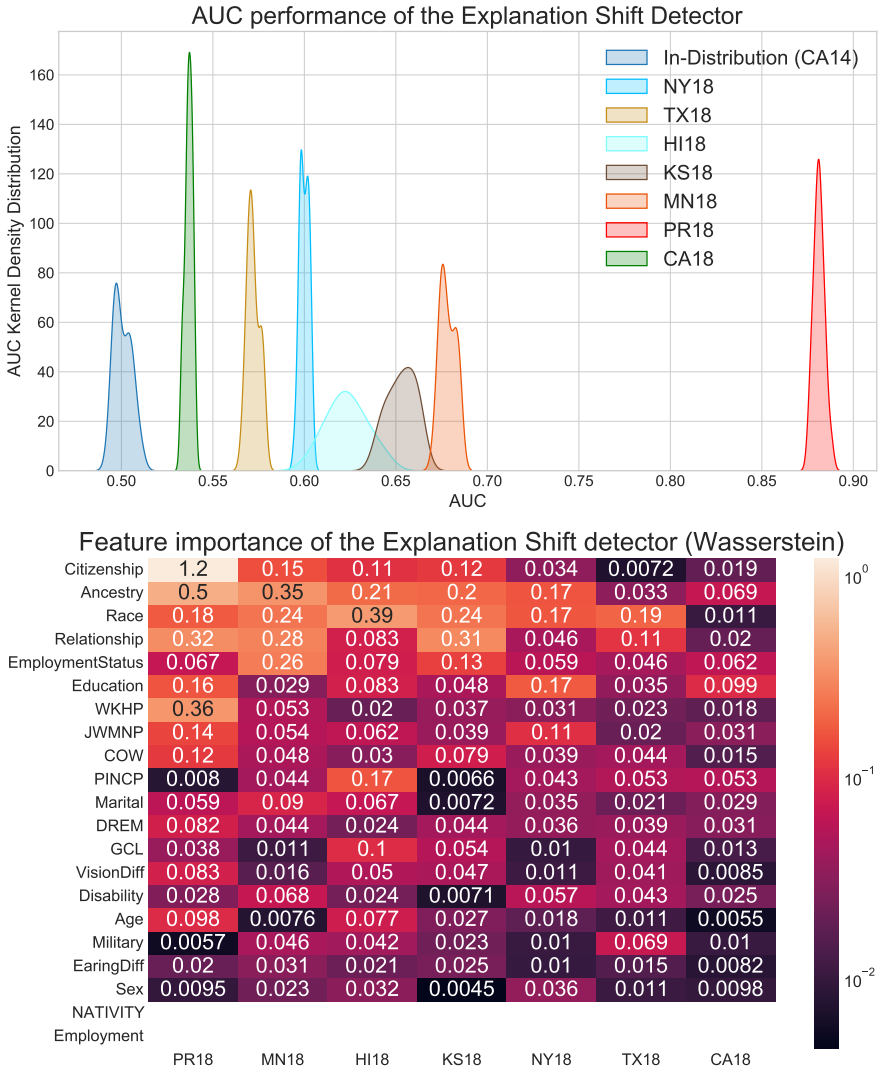

FIGURE 4.9: Left figure shows a comparison of the *Explanation Shift Detector*'s performance in different states for the ACS Mobility dataset. Except for PR18, all other states fall below an AUC of explanation shift detection of 0.70. The features driving this difference are Citizenship and Ancestry relationships. For the other states, protected social attributes, such as Race or Marital status, play an important role.

4.4 Experiments with Modeling Methods and Hyperparameters

In the next sections, we are going to show the sensitivity or our method to variations of the model f, the detector g, and the parameters of the estimator f_θ.

As an experimental setup, In the main body of the paper, we have focused on the UCI Adult Income dataset. The experimental setup has been using Gradient Boosting Decision Tree as the model f_θ and then as "Explanation Shift Detector" g_ψ a logistic regression. In this section, we extend the experimental setup by providing experiments by varying the types of algorithms for a given experimental set-up: the UCI Adult Income dataset using the Novel Covariate Group Shift for the "Asian" group with a fraction ratio of 0.5 (cf. Section 4.3).

4.4.1 Varying Models and Explanation Shift Detectors

OOD data detection methods based on input data distributions only depend on the type of detector used, being independent of the model f_θ. OOD Explanation methods rely on both the model and the data. Using explanations shifts as indicators for measuring distribution shifts impact on the model enables us to account for the influencing factors of the explanation shift. Therefore, in this section, we compare the performance of different types of algorithms for explanation shift detection using the same experimental setup. The results of our experiments show that using Explanation Shift enables us to see differences in the choice of the original model f_θ and the Explanation Shift Detector g_ϕ

	Estimator f_θ						
Detector g_ϕ	**XGB**	**Log.Reg**	**Lasso**	**Ridge**	**Rand.Forest**	**Dec.Tree**	**MLP**
XGB	0.583	0.619	0.596	0.586	0.558	0.522	0.597
LogisticReg.	0.605	0.609	0.583	0.625	0.578	0.551	0.605
Lasso	0.599	0.572	0.551	0.595	0.557	0.541	0.596
Ridge	0.606	0.61	0.588	0.624	0.564	0.549	0.616
RandomForest	0.586	0.607	0.574	0.612	0.566	0.537	0.611
DecisionTree	0.546	0.56	0.559	0.569	0.543	0.52	0.569

TABLE 4.7: Comparison of explanation shift detection performance, measured by AUC, for different combinations of explanation shift detectors and estimators on the UCI Adult Income dataset using the Novel Covariate Group Shift for the "Asian" group with a fraction ratio of 0.5 (cf. Section 4.3). The table shows that the algorithmic choice for f_θ and g_ψ can impact the OOD explanation performance. We can see how, for the same detector, different f_θ models flag different OOD explanations performance. On the other side, for the same f_θ model, different detectors achieve different results.

4.4.2 Hyperparameters Sensitivity Evaluation

This section presents an extension to our experimental setup where we vary the model complexity by varying the model hyperparameters $\mathcal{S}(f_\theta, X)$. We use the UCI Adult Income dataset with the Novel Covariate Group Shift for the "Asian" group with a fraction ratio of 0.5 as described in Section 4.3. And for the Stackoverflow as training data we use the United States of America and a novel covariate group France.

In this experiment, we changed the hyperparameters of the original model: for the decision tree, we varied the depth of the tree, while for the gradient-boosting decision, we changed the number of estimators, and for the random forest, both hyperparameters. We calculated the Shapley values using TreeExplainer (Lundberg et al., 2020a). For the Detector choice of model, we compare Logistic Regression and XGBoost models.

The results presented in Figure 4.10 show the AUC of the *Explanation Shift Detector* for the ACS Income dataset under novel group shift. We observe that the distribution shift does not affect very simplistic models, such as decision trees with depths 1 or 2. However, as we increase the model complexity, the out-of-distribution data impact on the model becomes more pronounced. Furthermore, when we compare the performance of the *Explanation Shift Detector* across different models, such as Logistic Regression and Gradient Boosting Decision Tree, we observe distinct differences(note that the y-axis takes different values).

In conclusion, the explanation distributions serve as a projection of the data and model sensitive to what the model has learned. The results demonstrate the importance of considering model complexity under distribution shifts.

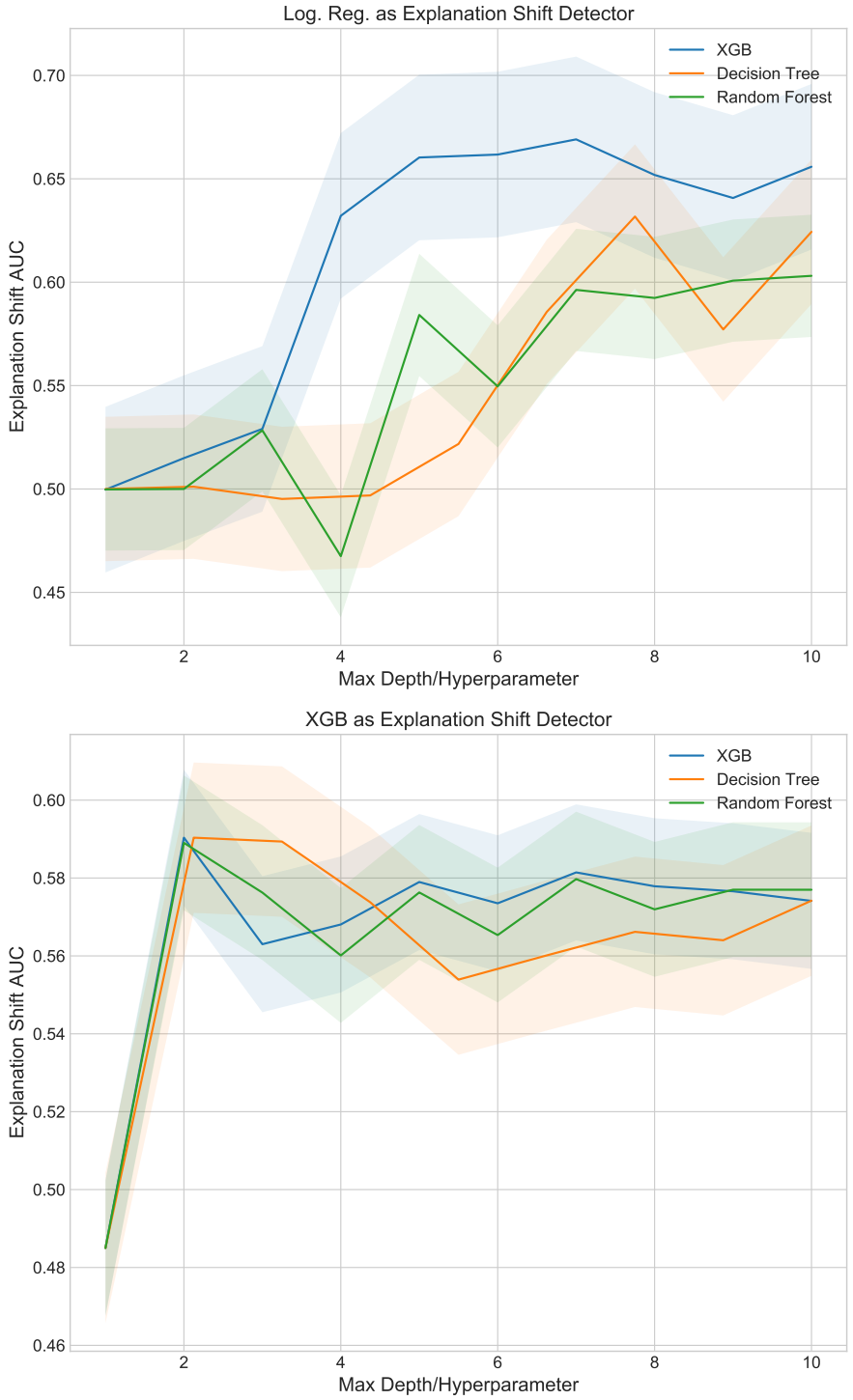

FIGURE 4.10: Images represent the AUC of the *Explanation Shift Detector*, on ACS Income and last two Stackoverflow under novel group shift. In the image above, the detector is a logistic regression, and in the images below, it is a gradient-boosting decision tree classifier. By changing the model, we can see that vanilla models (decision tree with depth 1 or 2) are unaffected by the distribution shift, while when increasing the model complexity, the out-of-distribution impact of the data in the model starts to be tangible

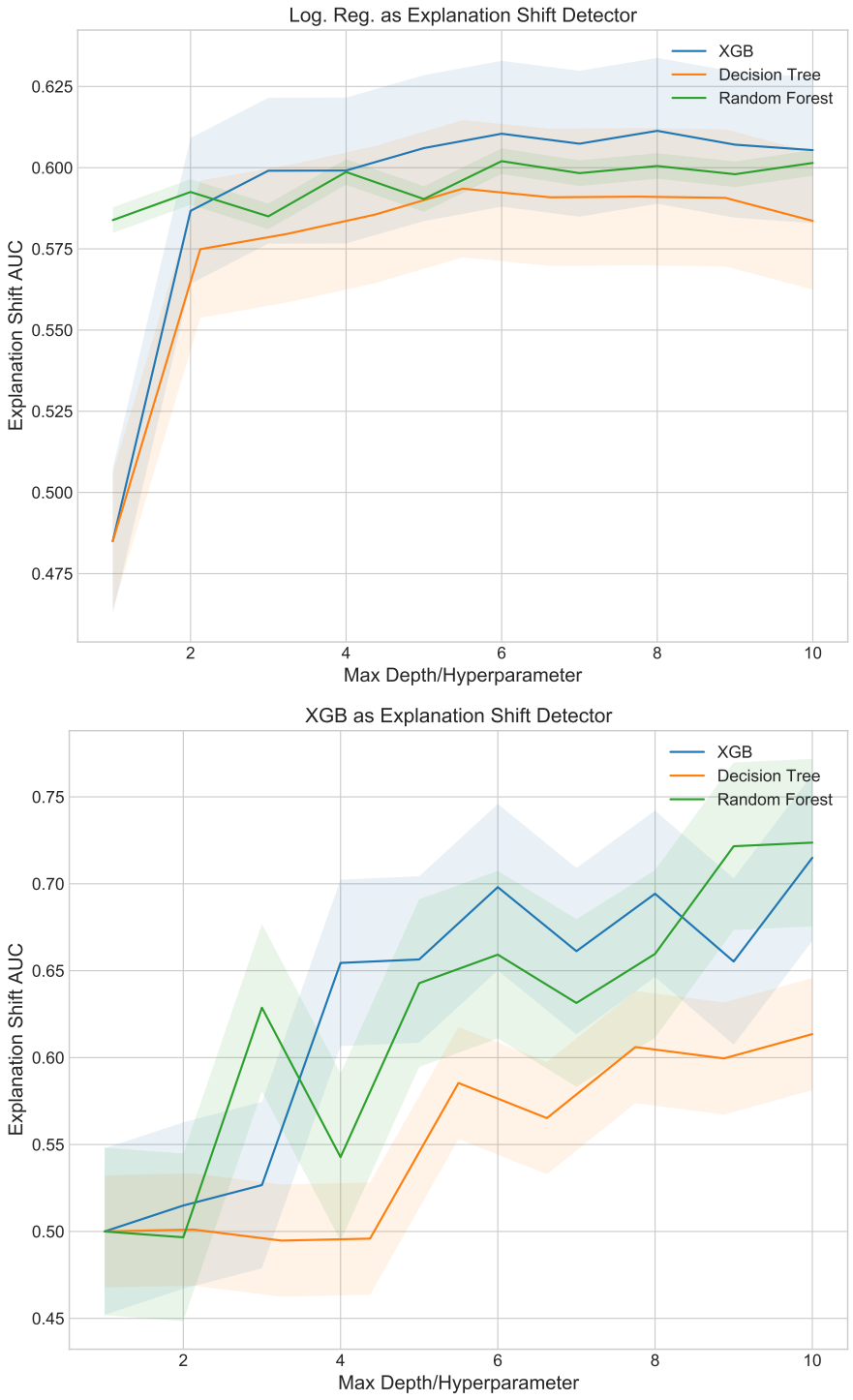

FIGURE 4.11: Images represent the AUC of the *Explanation Shift Detector*, on Stackoverflow under novel group shift. In the image above, the detector is a logistic regression, and in the image below, it is a gradient-boosting decision tree classifier. By changing the model, we can see that vanilla models (decision tree with depth 1 or 2) are unaffected by the distribution shift, while when increasing the model complexity, the out-of-distribution impact of the data in the model starts to be tangible

4.5 LIME as an Alternative Explanation Method

Another feature attribution technique that satisfies the aforementioned properties (efficiency and uninformative features Section 2.2) and can be used to create the explanation distributions is LIME (Local Interpretable Model-Agnostic Explanations). The similar discussion that we had in section 3.5.2.2, also applies here.

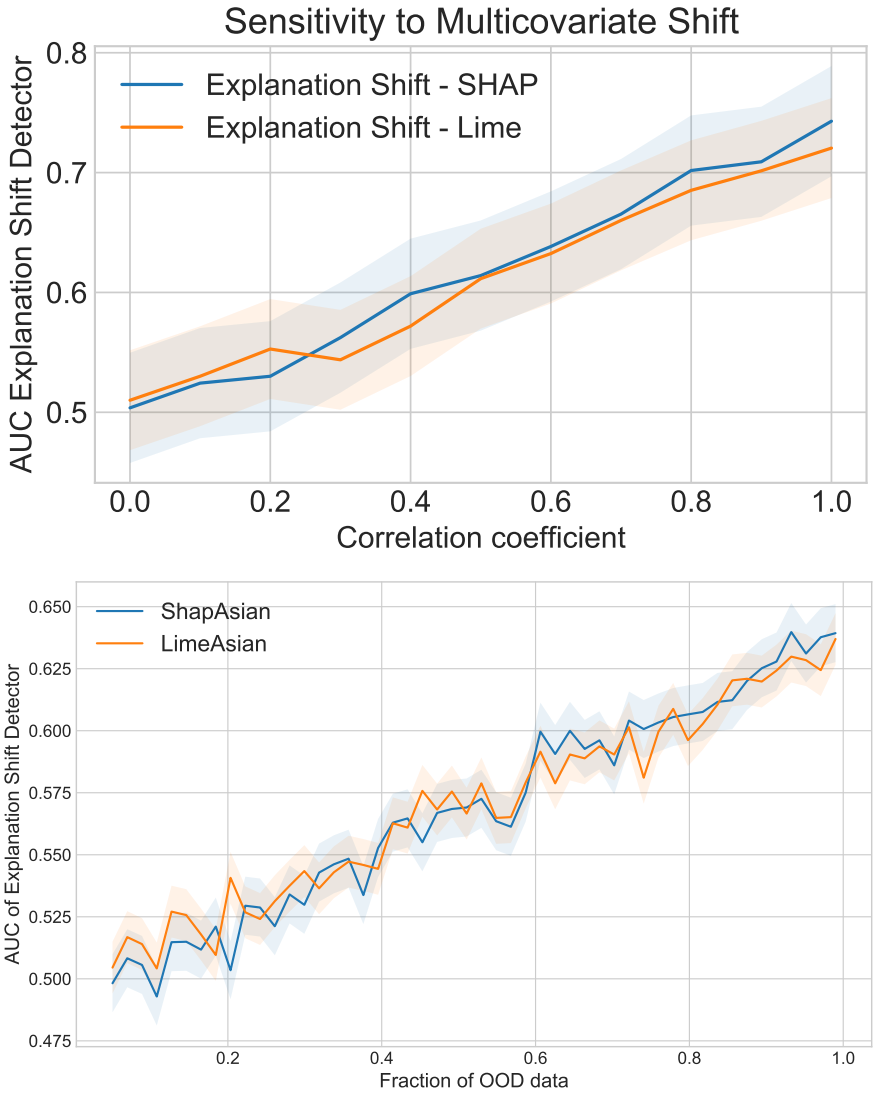

FIGURE 4.12: Comparison of the explanation distribution generated by LIME and SHAP. The above figure shows the sensitivity of the predicted probabilities to multi covariate changes using the synthetic data experimental setup of 4.2 on the main body of the paper. The figure below shows the distribution of explanation shifts for a New Covariate Category shift (Asian) in the ACS Income dataset.

Figure 4.12 compares the explanation distributions generated by LIME and SHAP. The left plot shows the sensitivity of the predicted probabilities to multicovariate changes using the synthetic data experimental setup from Figure 4.2 in the main body of the paper. The below plot shows the distribution of explanation shifts for a New Covariate

Category shift (Asian) in the ASC Income dataset. The performance of OOD explanations detection is similar between the two methods, but LIME suffers from two drawbacks: its theoretical properties rely on the definition of a local neighborhood, which can lead to unstable explanations (false positives or false negatives on explanation shift detection), and its computational runtime required is much higher than that of SHAP (see experiments below).

4.5.1 Runtime

We conducted an analysis of the runtimes of generating the explanation distributions using the two proposed methods. The experiments were run on a server with 4 vCPUs and 32 GB of RAM. We used `shap` version 0.41.0 and `lime` version 0.2.0.1 as software packages. In order to define the local neighborhood for both methods in this example we use all the data provided as background data. As an f_θ model, we use an `xgboost` and compare the results of TreeShap against LIME. When varying the number of samples we use 5 features and while varying the number of features we use 1000 samples.

Figure 4.13, shows the wall time required for generating explanation distributions using SHAP and LIME with varying numbers of samples and columns. The runtime required of generating an explanation distributions using LIME is much higher than using SHAP, especially when producing explanations for distributions. This is due to the fact that LIME requires training a local model for each instance of the input data to be explained, which can be computationally expensive. In contrast, SHAP relies on heuristic approximations to estimate the feature attribution with no need to train a model for each instance. The results illustrate that this difference in computational runtime becomes more pronounced as the number of samples and columns increases.

We note that the computational burden of generating the explanation distributions can be further reduced by limiting the number of features to be explained, as this reduces the dimensionality of the explanation distributions, but this will inhibit the quality of the explanation shift detection as it won't be able to detect changes on the distribution shift that impact model on those features.

Given the current state-of-the-art of software packages we have used SHAP values due to lower runtime required and that theoretical guarantees hold with the implementations. In the experiments performed in this paper, we are dealing with a medium-scaled dataset with around $\sim 1,000,000$ samples and $20 - 25$ features. Further work can be envisioned on developing novel mathematical analysis and software that study under which conditions which method is more suitable.

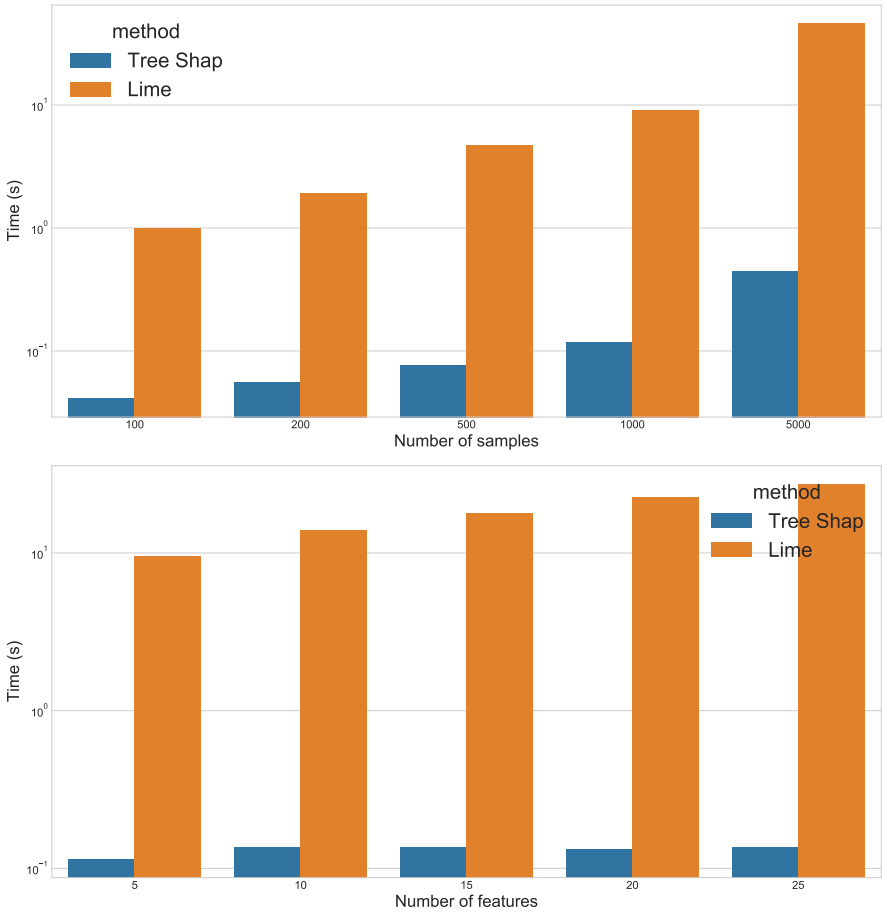

FIGURE 4.13: Wall time for generating explanation distributions using SHAP and LIME with different numbers of samples (above) and different numbers of columns (below). Note that the y-scale is logarithmic. The experiments were run on a server with 4 vCPUs and 32 GB of RAM. The runtime required to create an explanation distributions with LIME is far greater than SHAP for a gradient-boosting decision tree

4.6 Related Work on Tabular Data

4.6.1 Classifier two-sample test

Evaluating how two distributions differ has been a widely studied topic in the statistics and statistical learning literature (Hastie et al., 2001; Quiñonero-Candela et al., 2009; Liu et al., 2020a) and has advanced in recent years (Park et al., 2021a; Lee et al., 2018; Zhang et al., 2013). The use of supervised learning classifiers to measure statistical tests has been explored by Lopez-Paz and Oquab (2017) proposing a classifier-based approach that returns test statistics to interpret differences between two distributions. We adopt their power-test analysis and interpretability approach but apply it to the explanation distributions instead of input data distributions.

4.6.2 Out-Of-Distribution Detection

Evaluating how two distributions differ has been a widely studied topic in the statistics and statistical learning literature Hastie et al. (2001); Quiñonero-Candela et al. (2009); Liu et al. (2020a), that have advanced recently in last years Park et al. (2021a); Lee et al. (2018); Zhang et al. (2013). Rabanser et al. (2019) provides a comprehensive empirical investigation, examining how dimensionality reduction and two-sample testing might be combined to produce a practical pipeline for detecting distribution shifts in real-life machine learning systems. Other methods to detect if new data is OOD have relied on neural networks based on the prediction distributions Fort et al. (2021); Garg et al. (2020). They use the maximum softmax probabilities/likelihood as a confidence score Hendrycks and Gimpel (2017), temperature or energy-based scores Ren et al. (2019); Liu et al. (2020b); Wang et al. (2021), they extract information from the gradient space Huang et al. (2021), relying on the latent space Crabbé et al. (2021), they fit a Gaussian distribution to the embedding, or they use the Mahalanobis distance for out-of-distribution detection Lee et al. (2018); Park et al. (2021b).

Many of these methods are explicitly developed for neural networks that operate on image and text data, and often, they can not be directly applied to traditional ML techniques. For image and text data, one may build on the assumption that the relationships between relevant predictor variables (X) and response variables (Y) remain unchanged, i.e., that no *concept shift* occurs. For instance, the essence of how a dog looks remains unchanged over different data sets, even if contexts may change. Thus, one can define invariances on the latent spaces of deep neural models, which do not apply to tabular data in a similar manner. For example, predicting buying behaviour before, during, and after the COVID-19 pandemic constitutes a conceptual shift that is not amenable to such methods. We focus on such tabular data where techniques such as gradient boosting decision trees achieve state-of-the-art model performance Grinsztajn et al. (2022b); Elsayed et al. (2021b); Borisov et al. (2021).

Other research lines, such as membership inference (or membership classifiers), aim to detect that, given a data record and a machine learning model, whether the record was in the model's training dataset or not Shokri et al. (2017). Our approach is slightly different, as we don't attempt to predict if a certain instance belongs to the training data or not, but if its distribution is as the training dataset.

4.6.3 Detecting distribution shift and its impact on model behaviour

Extensive related work has aimed at detecting that data is from out-of-distribution. To this end, they have created several benchmarks that measure whether data comes from in-distribution or not (Koh et al., 2021; Sagawa et al., 2021; Malinin et al., 2021a, 2022,

2021b). In contrast, our main aim is to evaluate the impact of the distribution shift on the use of model.

A typical example is two-sample testing on the latent space such as described by Rabanser et al. (2019). However, many of the methods developed for detecting out-of-distribution data are specific to neural networks processing image and text data and can not be applied to traditional machine learning techniques. These methods often assume that the relationships between predictor and response variables remain unchanged, i.e., no concept shift occurs. Our work is applied to tabular data where techniques such as gradient boosting decision trees achieve state-of-the-art model performance (Grinsztajn et al., 2022b; Elsayed et al., 2021b; Borisov et al., 2021).

4.6.4 Impossibility of model monitoring

Recent research findings have formalized the limitations of monitoring machine learning models in the absence of labelled data. Specifically (Garg et al., 2022; Chen et al., 2022b) prove the impossibility of predicting model degradation or detecting out-of-distribution data with certainty (Fang et al., 2022; Zhang et al., 2021; Guerin et al., 2022). Although our approach does not overcome these limitations, it provides valuable insights for machine learning engineers to better understand changes in interactions between learned models and shifting data distributions.

4.6.5 Model monitoring and distribution shift under specific assumptions:

Under specific types of assumptions, model monitoring and distribution shift become feasible tasks. One type of assumption often found in the literature is to leverage causal knowledge to identify the drivers of distribution changes (Budhathoki et al., 2021; Zhang et al., 2022; Schrouff et al., 2022). For example, Budhathoki et al. (2021) use graphical causal models and feature attributions based on Shapley values to detect changes in the distribution. Similarly, other works aim to detect specific distribution shifts, such as covariate or concept shifts. Our approach does not rely on additional information, such as a causal graph, labelled test data, or specific types of distribution shift. Still, by the nature of pure concept shifts, the model behaviour remains unaffected and new data need to come with labelled responses to be detected.

4.6.6 Explainability and distribution shift

Lundberg et al. (2020a) applied Shapley values to identify possible bugs in the pipeline by visualizing univariate SHAP contributions. Following this line of work, Nigenda et al. (2022) compare the order of the feature importance using the NDCG between

training and unseen data. We go beyond their work and formalize the multivariate explanation distributions on which we perform a two-sample classifier test to detect how distribution shift impacts interaction with the model. Furthermore, we provide a mathematical analysis of how the SHAP values contribute to detecting distribution shift. In Appendix 4.4 we provide a formal comparison against Nigenda et al. (2022).

An approach using Shapley values by Balestra et al. (2022) allows for tracking distributional shifts and their impact among for categorical time series using slidSHAP, a novel method for unlabelled data streams. This approach is particularly useful for unlabelled data streams, offering insights into the changing data distribution dynamics. In contrast, our work focuses on defining explanation distributions and leveraging their theoretical properties in the context of distribution shift detection, employing a two-sample classifier test for detection.

Another perspective in the field of explainability is explored by Adebayo et al. (2020, 2018b), who investigate the effectiveness of post-hoc model explanations for diagnosing model errors. They categorize these errors based on their source. While their work is geared towards model debugging, our research takes a distinct path by aiming to quantify the influence of distribution shifts on the model.

Hinder et al. (2022) proposes to explain concept drift by contrasting explanations describing characteristic changes of spatial features. Haug and Kasneci (2021) track changes in the distribution of model parameter values that are directly related to the input features to identify concept drift early on in data streams. In a more recent paper, Haug et al. (2022) also exploits the idea that local changes to feature attributions and distribution shifts are strongly intertwined and uses this idea to update the local feature attributions efficiently. Their work focuses on model retraining and concept shift, in our work, the original model f_θ remains unaltered, and since we are in an unsupervised monitoring scenario, we can't detect concept shifts see discussion in Section 4.7

4.7 Discussion

In this study, we conducted a comprehensive evaluation of explanation shift by systematically varying models (f), model parametrizations (θ), feature attribution explanations ($\mathcal{S}$), and input data distributions ($\mathcal{D}_X$). Our objective was to investigate the impact of distribution shift on the model by explanation shift and gain insights into its characteristics and implications.

The Shapley value, a key component in our method, describes how a model's prediction for a specific data point deviates from the mean. These theoretical considerations, which we laid out in Section 4.2, have been confirmed by our experimental sections.

In Section 4.3.3, we have studied input distribution shift. Our experiment shows that explanation shift detects input distribution shifts better than the best baseline methods. Table 4.6 showcases the two top-performing methods—comparing input distributions with Kolmogorov-Smirnoff *(B1)* and our method — with statistically insignificant differences.

In Section 4.3.2, we have studied co-variate shift. Considering, Table 4.5, the best method for detecting input distribution shifts, *(B1)*, fails completely on this task. The second best method is *(B2)* comparison of prediction distributions using the Wasserstein distance, which also did quite well wrt. predicting input distribution shifts and came rather close behind our approach in both experiments.

Moreover, in our geopolitical and temporal shift experiment (Section 4.3.5), we demonstrate the ability to account for the drivers of model changes under such input data shifts. Cross-task comparisons in experiments (Figure 3.9 or Figure 3.8) highlight how explanation shift feature importance varies even when input distribution shifts remain constant during cross-task. These capabilities are not offered by any of the competing baselines. These observations are further supported by additional experiments in Appendix 3.6.4, where we solely vary model complexity, showcasing the adaptability of explanation shifts to changes in model characteristics.

Our approach cannot detect concept shifts, as concept shift requires understanding the interaction between prediction and response variables. By the nature of pure concept shifts, such changes do not affect the model. To be understood, new data need to come with labelled responses. We work under the assumption that such labels are not available for new data, nor do we make other assumptions; therefore, our method is not able to predict the degradation of prediction performance under distribution shifts. All papers such as (Garg et al., 2022; Baek et al., 2022; Chen et al., 2022b; Fang et al., 2022; Miller et al., 2021; Lu et al., 2023) that address the monitoring of prediction performance have the same limitation. Only under specific assumptions, e.g., no occurrence of concept shift or causal graph availability, can performance degradation be predicted with reasonable reliability.

The potential utility of explanation shifts as distribution shift indicators that affect the model in computer vision or natural language processing tasks remains an open question. We have used feature attribution explanations to derive indications of explanation shifts, but other AI explanation techniques may be applicable and come with their advantages.

4.8 Conclusions

Commonly, the problem of detecting the impact of the distribution shift on the model has relied on measurements for detecting shifts in the input or output data distributions or relied on assumptions either on the type of distribution shift or causal graphs availability. In this paper, we proposed explanation shifts as an indicator for detecting and identifying the impact of distribution shifts on machine learning models. We provide software, mathematical analysis examples, synthetic data, and real-data experimental evaluation. We found that measures of explanation shift can provide more insights than input distribution and prediction shift measures when monitoring machine learning models.

Limitations

The potential utility of explanation shifts as distribution shift indicators that affect the model in computer vision or natural language processing tasks remains an open question. We have used feature attribution explanations to derive indications of explanation shifts, but other AI explanation techniques may be applicable and come with their advantages. Also, our approach cannot detect concept shifts, as concept shift requires understanding the interaction between input data and response variables. By the nature of pure concept shifts, such changes do not affect the model. We work under the assumption that such labels are not available for new data, nor do we make other assumptions; therefore, our method is not able to predict the degradation of prediction performance under distribution shifts.

Furthermore, our use of the `shap` Python package for Shapley values approximation can introduce known drawbacks, as highlighted in recent literature (Bilodeau et al., 2022; Slack et al., 2020b). Additionally, our current implementation relies on linear Shapley value interaction approximations, which can be extended following the work of Fumagalli et al. (2023); Bordt and von Luxburg (2023).

Chapter 5

Building Indicators of Model Performance Deterioration

Monitoring machine learning models in production is not an easy task. There are situations when the true label of the deployment data is available, and performance metrics can be monitored. But there are cases where it is not, and performance metrics are not so trivial to calculate once the model has been deployed. Model monitoring aims to ensure that a machine learning application in a production environment displays consistent behavior over time.

Being able to explain or remain accountable for the performance or the deterioration of a deployed model is crucial, as a drop in model performance can affect the whole business process, potentially having catastrophic consequences[1]. Once a deployed model has deteriorated, models are retrained using previous and new input data in order to maintain high performance. This process is called continual learning (Diethe et al., 2019) and it can be computationally expensive and put high demands on the software engineering system. Deciding when to retrain machine learning models is paramount in many situations.

Traditional machine learning systems assume that training data has been generated from a stationary source, but *data is not static, it evolves*. This problem can be seen as a distribution shift, where the data distributions of the training set and the test set differ. Detecting distribution shifts has been a longstanding problem in the machine learning (ML) research community (Shimodaira, 2000; Sugiyama et al., 2007; Masashi and Klaus-Robert, 2005; Tasche, 2017; Zadrozny, 2004; Stolzenberg and Relles, 1997; Heckman, 1990; Cortes et al., 2008; Huang et al., 2007; He et al., 2014), as it is one of the main

[1]The Zillow case is an example of consequences of model performance degradation in an unsupervised monitoring scenario, see https://edition.cnn.com/2021/11/09/tech/zillow-ibuying-home-zestimate/index.html (Online accessed January 26, 2022).

sources of model performance deterioration (Quiñonero-Candela et al., 2009). Furthermore, data scientists in machine learning competitions claim that finding the train/validation split that better resembles the test (evaluation) distribution is paramount to winning a Kaggle competition (Guschin et al., 2018).

However, despite the fact that a shift in data distribution can be a source of model deterioration, the two are not identical. Indeed, if we shift a random noise feature we have caused a change in the data distribution, but we should not expect the performance of a model to decline when evaluated on this shifted dataset. Thus, we emphasize here that our focus is on *model deterioration* and not distribution shift, despite the correlation between the two.

Established ways of monitoring distribution shift when the real target distribution is not available are based on statistical changes either the input data (Diethe et al., 2019; Rabanser et al., 2019) or on the model output Garg et al. (2022). These statistical tests correctly detect univariate changes in the distribution but are completely independent of the model performance and can therefore be too sensitive, indicating a change in the covariates but without any degradation in the model performance. This can result in false positives, leading to unnecessary model retraining. It is worth noting that several authors have stated the clear need to identify how non-stationary environments affect the behavior of models (Diethe et al., 2019).

Aside from merely indicating that a model has deteriorated, it can in some circumstances be beneficial to identify the *cause* of the model deterioration by detecting and explaining the lack of knowledge in the prediction of a model. Such explainability techniques can provide algorithmic transparency to stakeholders and to the ML engineering team (Mougan et al., 2021a; Bhatt et al., 2021; Koh and Liang, 2020; Ribeiro et al., 2016b; Sundararajan et al., 2017).

This chapter's primary focus is on non-deep learning models and small to medium-sized tabular datasets, a size of data that is very common in the average industry, where, non-deep learning-based models achieve state-of-the-art results Grinsztajn et al. (2022c); Borisov et al. (2022); Elsayed et al. (2021a).

Our contributions are the following:

1. We use this non-parametric uncertainty estimation method to develop a machine learning monitoring system for regression models, which outperforms previous monitoring methods in terms of detecting deterioration of model performance.

2. We use explainable AI techniques to identify the source of model deterioration for both entire distributions as a whole as well as for individual samples, where classical statistical indicators can only determine distribution differences.

3. We release an open source Python package, `doubt`, which implements our uncertainty estimation method and is compatible with all `scikit-learn` models (Pedregosa et al., 2011).

5.1 Related Work

5.1.1 Model Monitoring

From a software perspective, a machine learning model is another software component whose functionalities need to be tested before deployment. This approach has one structural weakness: *data is not static, it evolves* (Diethe et al., 2019). The conditions in which a system is developed can differ from deployment time for a range of reasons, from an evolutionary environment, a sampling bias at training time, or the inability to reproduce test conditions at training time (Quiñonero-Candela et al., 2009).

Model monitoring techniques help to detect unwanted changes in the behavior of a machine learning application in a production environment. One of the biggest challenges in model monitoring is distribution shift, which is also one of the main sources of model degradation (Quiñonero-Candela et al., 2009; Diethe et al., 2019).

Diverse types of model monitoring scenarios require different supervision techniques. We can distinguish two main groups: Supervised learning and unsupervised learning. Supervised learning is the appealing one from a monitoring perspective, where performance metrics can easily be tracked. Whilst attractive, these techniques are often unfeasible as they rely either on having ground truth labeled data available or maintaining a hold-out set, which leaves the challenge of how to monitor ML models to the realm of unsupervised learning Diethe et al. (2019). Popular unsupervised methods that are used in this respect are the Population Stability Index (PSI) and the Kolmogorov-Smirnov test (K-S), all of which measure how much the distribution of the covariates in the new samples differs from the covariate distribution within the training samples. These methods are often limited to real-valued data, low dimensions, and require certain probabilistic assumptions (Diethe et al., 2019; Malinin et al., 2021c).

Another approach suggested by Lundberg et al. (2019) is to monitor the SHAP value contribution of input features over time together with decomposing the loss function across input features in order to identify possible bugs in the pipeline as well as distribution shift. This technique can account for previously unaccounted bugs in the machine learning production pipeline but fails to monitor the model degradation.

Is worth noting that prior work (Garg et al., 2022; Jiang et al., 2021) has focused on monitoring models either on out-of-distribution data or in-distribution data (Neyshabur et al., 2017, 2019). Such a task, even if challenging, does not accurately represent the

different types of data a model encounters in the wild. In a production environment, a model can encounter previously seen data (training data), unseen data with the same distribution (test data), and statistically new and unseen data (out-of-distribution data). That is why we focus our work on finding an unsupervised estimator that replicates the behavior of the model performance.

The idea of mixing uncertainty with dataset shift was introduced by Ovadia et al. (2019). Our work differs from theirs, in that they evaluate uncertainty by shifting the distributions of their dataset, where we aim to detect model deterioration under dataset shift using uncertainty estimation. Their work is also focused on deep learning classification problems, while we estimate uncertainty using model agnostic regression techniques. Further, our contribution allows us to pinpoint the features/dimensions that are main causes of the model degradation.

Garg et al. (2022) introduces a monitoring system for classification models, based on imposing thresholds on the softmax values of the model. Our method differs from theirs in that we work with regression models and not classification models, and that our method utilizes external uncertainty estimation methods, rather than relying on the model's own "confidence" (i.e., the outputted logits and associated softmax values).

Rabanser et al. (2019), presents a comprehensive empirical investigation of dataset shift, examining how dimensionality reduction and two-sample testing might be combined to produce a practical pipeline for detecting distribution shift in a real-life machine learning system. They show that the two-sample-testing-based approach performs best. This serves as a baseline comparison within our models, even if their idea is more focused on binary classification, whereas our works focus on building a regression indicator.

5.1.2 Uncertainty

Uncertainty estimation is being developed at a fast pace. Model averaging (Kumar and Srivastava, 2012; Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017; Arnez et al., 2020) has emerged as the most common approach to uncertainty estimation. Ensemble and sampling-based uncertainty estimates have been successfully applied to many use cases such as detecting misclassifications (Lakshminarayanan, 2021), out-of-distribution inputs (D'Angelo and Henning, 2021), adversarial attacks (Carlini and Wagner, 2017; Smith and Gal, 2018), automatic language assessments (Malinin, 2019) and active learning (Kirsch et al., 2019). In our work, we apply uncertainty to detect and explain model performance for seen data (train), unseen and identically distributed data (test), and statistically new and unseen data (out-of-distribution). Barber et al.

(2021) recently introduced a new non-parametric method of creating prediction intervals using the Jackknife+. Our method differs from theirs in that we are using general bootstrapped samples for our estimates rather than leave-one-out estimates.

In this thesis, we use the method introduced by Mougan and Nielsen (2023), which is an extension of the method proposed by Kumar and Srivastava (2012), that takes into account the model's variance in the construction of the prediction intervals. Kumar and Srivastava (2012), introduced a non-parametric method to compute prediction intervals for any ML model using a bootstrap estimate with theoretical guarantees.

5.2 Methodology

5.2.1 Evaluation of Deterioration Detection Systems

The problem we are tackling in this paper is evaluating and accounting for model predictive performance deterioration, which in general is an impossible task. In order to achieve possible results, we simulate a gradual covariate distribution shift scenario in which we *have* access to the true labels, which we can use to measure the model deterioration and thus evaluate the monitoring system. A naive simulation in which we simply manually shift a chosen feature of a dataset would not be representative, as the associated true labels could have changed if such a shift happened "in the wild".

Therefore, we propose the following alternative approach. Starting from a real-life dataset $\mathcal{D}$ and a numerical feature F of $\mathcal{D}$, we sort the data samples of $\mathcal{D}$ by the value of F, and split the sorted $\mathcal{D}$ in three equally sized sections: $\{\mathcal{D}^{below}, \mathcal{D}^{tr}, \mathcal{D}^{upper}\} \subseteq \mathcal{D}$. The model is then fitted to the middle section ($\mathcal{D}^{tr}$) and evaluated on all of $\mathcal{D}$. The goal of the monitoring system is to input the model, the labelled data segment $\mathcal{D}^{tr}$ and a sample of unlabelled data $\mathcal{D}_X^{te} \subseteq \mathcal{D}$, and output a "monitoring value" which behaves like the model's performance on $\mathcal{D}^{te}$. Such a prediction will thus have to take into account the training performance, generalization performance, and the out-of-distribution performance of the model. This setup aims to evaluate a gradual covariate shift, in which we assume that in the vicinity of the training data, concept shift has not occurred.

In the experimental section, we compare our monitoring technique to several other such systems. To enable comparison between the different monitoring systems, we standardize all monitoring values as well as the performance metrics of the model. From these standardized values, we can now directly measure the goodness-of-fit of the model monitoring system by computing the absolute difference between its (standardized) monitoring values and the (standardized) ground truth model performance metrics. Our chosen evaluation method is very similar to the one used by Garg et al. (2022). They focus on classification models and their systems output estimates of the

model's accuracy on the dataset. They evaluate these systems by computing the absolute difference between the system's accuracy estimate and the actual accuracy that the model achieves on the dataset.

As we are working with regression models in this paper, we will only operate with a single model performance metric: mean squared error. We will introduce our monitoring system, which is based on an uncertainty measure, and will compare our monitoring system against statistical tests based on input data or prediction data. In that section, we will also compare our uncertainty estimation method to current state-of-the-art uncertainty estimation methods.

5.2.2 Uncertainty Estimation

In order to estimate uncertainty in a general way for all machine learning models, we use a non-parametric regression technique. The method aims to determine prediction intervals for outputs of general non-parametric regression models using bootstrap methods.

We estimate the uncertainty of a model by computing a bootstrap estimate of the variance of the predictions. This is part of the method used in Kumar and Srivastava (2012) and is computed as follows. From a dataset X and a new sample $x_0 \notin X$, we first bootstrap X $B > 0$ times, yielding bootstrapped subsamples X_b for each $b < B$. We next fit our model on each of the X_b's, call the resulting fitted models M_b, and then computing the predictions of the fitted models of x_0, $M_b(x_0)$. We then use the length of the 95% confidence interval of these bootstrapped predictions as our uncertainty measure. Concretely, this is computed as

$$\texttt{uncertainty}_M(x_0) := (q_{0.025}(\{M_b(x_0) \mid b < B\}), q_{0.975}(\{M_b(x_0) \mid b < B\})), \quad (5.1)$$

with $q_\alpha(A)$ being the α'th quantile of the set A.

5.2.3 Detecting the Source of Uncertainty/Model Deterioration

Using uncertainty as a method to monitor the performance of an ML model does not provide any information on *what* features are the cause of the model degradation, only a goodness-of-fit to the model performance. We propose to solve this issue with the use of Shapley values.

We start by fitting a model f_θ to the training data, X^{train}. We next shift the test data by five standard deviations (call the shifted data X^{ood}) and compute uncertainty estimates Z of f_θ on X^{ood}. We next fit a second model g_ψ on (X^{ood}) to predict the uncertainty estimate Z, and compute the associated Shapley values Lundberg et al. (2019) of g_ψ. These

Shapley values thus signify which features are the ones contributing the most to the uncertainty values. With the correlation between uncertainty values and model deterioration that we hope to conclude from the experiment described in the experimental section, this thus also provides us with a plausible cause of the model deterioration, if deterioration has taken place. Particularly, this methodology can be extended to large-scale datasets and deep learning-based models.

5.3 Experiments

Our experiments have been organized into two main groups: Firstly, we assess the performance of our proposed uncertainty method for monitoring the performance of a machine learning model. Then, we evaluate the usability of the explainable uncertainty for identifying the features that are driving model degradation in local and global scenarios. We present the results over several real-world datasets, Table 5.1; we provide experiments on synthetic datasets that exhibit non-linear and linear behaviour.

TABLE 5.1: Statistics of the regression datasets used in this paper.

Dataset	# Samples	# Features
Airfoil Self-Noise	1,503	5
Bike Sharing	17,379	16
Concrete Strength	1,030	8
QSAR Fish Toxicity	908	6
Forest Fires	517	12
Parkinsons	5,875	22
Power Plant	9,568	4
Protein	45,730	9

5.3.1 Evaluating Model Deterioration

The scenario we are addressing is characterized by regression data sets with statistically seen data (train data), IID statistically unseen data (test data), and out-of-distribution data. Following the open data for reproducible research guidelines described in Community et al. (2019) and for measuring the performance of the proposed methods, we have used eight open-source datasets (cf. Table 5.1) for an empirical comparison coming from the UCI repository (Dua and Graff, 2017b). As described in the methodology, to benchmark our algorithm we, for each feature F in each dataset $\mathcal{D}$, sort $\mathcal{D}$ according to F and split $\mathcal{D}$ into three equally sized sections $\{\mathcal{D}^{below}, \mathcal{D}^{tr}, \mathcal{D}^{upper}\} \subseteq \mathcal{D}$. We then train the model on $\mathcal{D}^{tr}$ and test the performance of all of $\mathcal{D}$. In this way we obtain a mixture of train, test, and out-of-distribution data, allowing us to evaluate our monitoring techniques in all three scenarios.

In evaluating a monitoring system we need to make a concrete choice of the sampling method to get the unlabelled data $\mathcal{S} \subseteq \mathcal{D}$. We are here using a rolling window of fifty samples, which has the added benefit of giving insight into the performance of the monitoring system on each of the three sections $\mathcal{D}^{lower}$, $\mathcal{D}^{tr}$ and $\mathcal{D}^{upper}$ (cf. Figure 5.1).

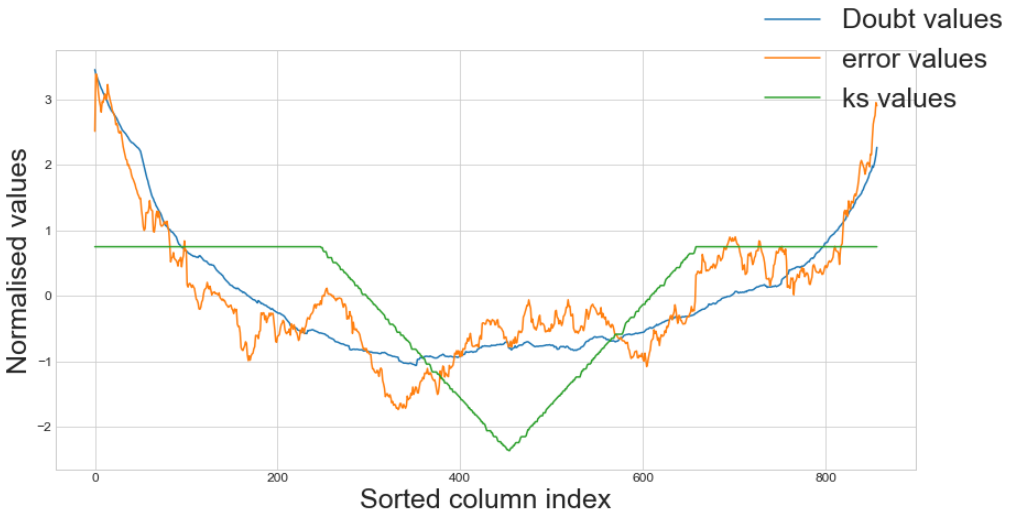

FIGURE 5.1: Comparison of different model degradation detection methods for the Fish Toxicity dataset. Each of the plots represents an independent experiment where each of the six features has been shifted, using the method described in the methodology section. Doubt achieves a better goodness-of-fit than previous statistical methods. A larger version of this figure can be found in Appendix.

We compare our monitoring system using the uncertainty estimation method against: (i) two classical statistical methods on input data: the Kolmogorov-Smirnov test statistic (K-S) and the Population-Stability Index (PSI) (Diethe et al., 2019), (ii) a Kolmogorov-Smirnov statistical test on the predictions between train and test Garg et al. (2022) that we denominate prediction shift and (iii) the previous state-of-the-art uncertainty estimate MAPIE. We evaluate the monitoring systems on a variety of model architectures: generalized linear models, tree-based models as well as neural networks.

The average performance across all datasets can be found in Table 5.2.[2] From these we can see that our methods outperform K-S and PSI in all cases except for the Random Forest case, where our method is still on par with the best method, in that case, K-S. We have included a table with each dataset and all the estimators in the appendix, where it can be seen that both K-S and PSI easily identify a shift in the distribution but fail to detect when the model performance degrades, giving too many false positives.

[2]See the appendix for a more detailed table.

	PSI	K-S	Pred.Shift	MAPIE	Doubt	Exp.Shift
Linear Reg.	0.87 ± 0.08	0.81 ± 0.10	0.86 ± 0.13	0.77 ± 0.10	$\mathbf{0.71 \pm 0.14}$	0.85 ± 0.10
Poisson	0.93 ± 0.08	0.94 ± 0.20	1.00 ± 0.15	0.83 ± 0.18	$\mathbf{0.79 \pm 0.14}$	0.94 ± 0.14
Decision Tree	0.97 ± 0.10	0.52 ± 0.12	0.80 ± 0.14	0.60 ± 0.16	$\mathbf{0.49 \pm 0.10}$	0.77 ± 0.16
Random Forest	0.95 ± 0.08	$\mathbf{0.50 \pm 0.12}$	0.73 ± 0.18	0.86 ± 0.15	0.74 ± 0.18	0.89 ± 0.20
Gradient Boosting	0.95 ± 0.08	0.61 ± 0.19	0.75 ± 0.20	0.73 ± 0.18	$\mathbf{0.58 \pm 0.23}$	0.87 ± 0.18
MLP	0.84 ± 0.16	0.72 ± 0.22	0.74 ± 0.22	0.74 ± 0.38	$\mathbf{0.68 \pm 0.38}$	0.84 ± 0.10

TABLE 5.2: Performance of model monitoring systems for model deterioration for a variety of model architectures on eight regression datasets from the UCI repository Dua and Graff (2017b). The scores are the means and standard deviations of the absolute deviation from the true labels on $\mathcal{D}^{lower}$ and $\mathcal{D}^{upper}$ (lower is better). K-S and PSI are the monitoring systems obtained by computing the Kolmogorov-Smirnov test values and the Population Stability Index, respectively, Prediction Shift is the statistical comparison of the model prediction, and Doubt is our method. The best results for each model architecture are shown in bold. See the Appendix for all the raw scores.

5.3.2 Detecting the Source of Uncertainty

For this experiment, we make use of two datasets: a synthetic one and the popular House Prices regression dataset[3], where the goal is to predict the selling price of a given property. We select two of the features that are the most correlated with the target, `GrLivArea` and `TotalBsmtSF`, and also create a new feature of random noise, to have an example of a feature with minimum correlation with the target. A model deterioration system should therefore highlight the `GrLivArea` and `TotalBsmtSF` features, and *not* highlight the random features.

Concretely, we compute an estimation of the Shapley values using TreeSHAP Lundberg et al. (2019), which is an efficient estimation approach values for tree-based models, that allows for this second model to identify the features that are the source of the uncertainty, and thus also provide an indicator for what features might be causing the model deterioration.

We fitted an MLP on the training dataset, which achieved a R^2 value of 0.79 on the validation set. We then shifted all three features by five standard deviations and trained a gradient boosting model on the uncertainty values of the MLP on the validation set, which achieves a good fit (an R^2 value of 0.94 on the hold-out set of the validation). We then compare the SHAP values of the gradient boosting model with the PSI and K-S statistics for the individual features.

In Figure 5.2, classical statistics and SHAP values to detect the source of the model deterioration are compared. We see that the PSI and K-S value correctly capture the shift in each of the three features (including the random noise). On the other hand, our SHAP method highlights the two substantial features (`GrLivArea` and `TotalBsmtSF`) and correctly does not assign a large value to the random feature, despite the distribution shift.

[3] https://www.kaggle.com/c/house-prices-advanced-regression-techniques

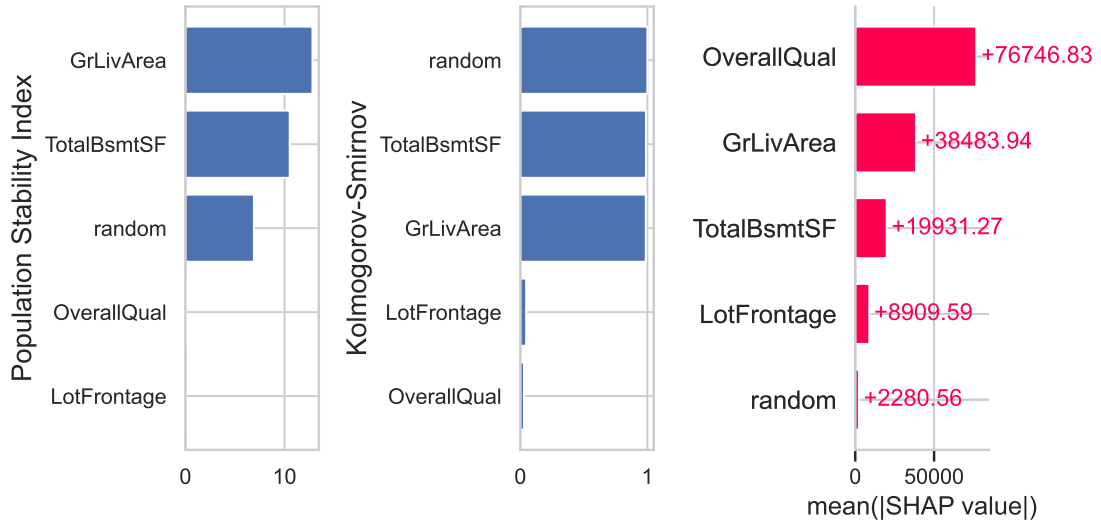

FIGURE 5.2: Global comparison of different distribution shift detection methods. Statistical methods correctly indicate that there exists a distribution shift in the shifted data. Shapley values indicate the contribution of each feature to the drop in predictive performance of the model.

Figure 5.3 shows features contributing to pushing the model output from the base value to the model output. Features pushing the uncertainty prediction higher are shown in red, and those pushing the uncertainty prediction lower are in blue (Bala et al., 2018; Lundberg et al., 2020a; Lundberg and Lee, 2017). From these values, we can, at a local level, also identify the two features (`GrLivArea` and `TotalBsmtSF`) causing the model deterioration in this case.

5.4 Synthetic Data Experiments

In this section we apply our model monitoring method to synthetic data to demonstrate main differences between our approach, methods from classical statistics and other model agnostic uncertainty approaches.

We sample data from a three-variate normal distribution $X = (X_1, X_2, X_3) \sim N(1, 0.1 \cdot I_3)$ with I_3 being an identity matrix of order 3. We construct the target variable as $Y := X_1^2 + X_2 + \varepsilon$, with $\varepsilon \sim N(0, 0.1)$ being random noise. We thus have a non-linear feature X_1^2, a linear feature X_2 and a feature X_3 which is not used, all of them being independent among each other. We draw $10,000$ random samples for both training and test data, yielding the training set (X^{tr}, Y^{tr}) and the test set (X^{te}, Y^{te}) and train a linear model f_θ.

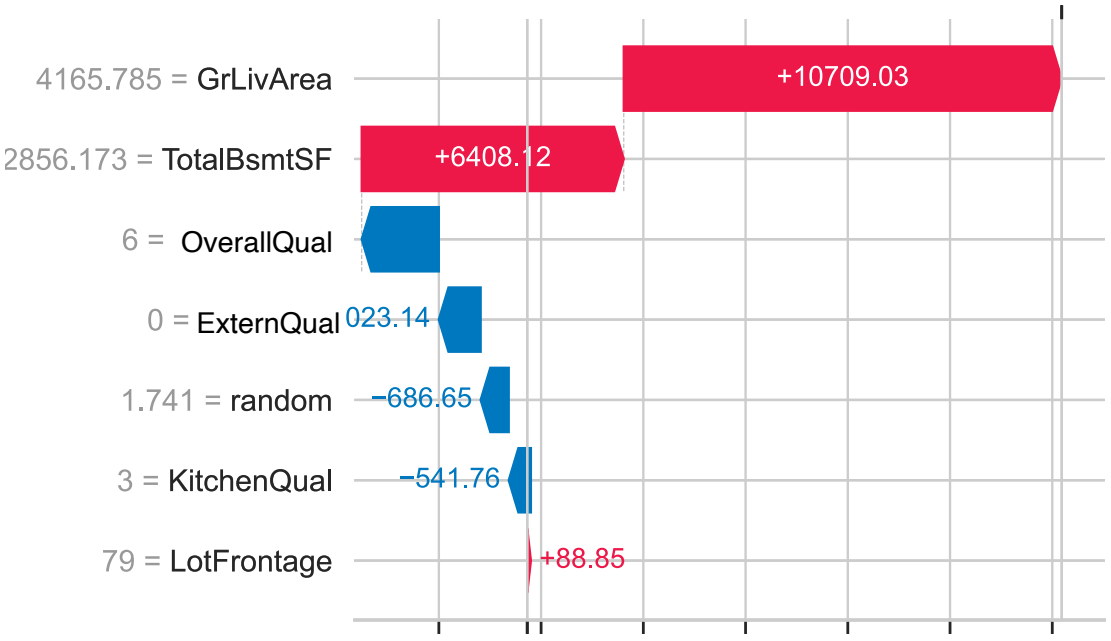

FIGURE 5.3: Individual explanation that displays the source of uncertainty for one instance. The previous method allowed only for comparison between distributions, now with explainable uncertainty, we are able to account for individual instances. In red, features pushing the uncertainty prediction higher are shown; in blue, those pushing the uncertainty prediction lower.

5.4.1 Evaluating Model Deterioration

In this experiment we aim to compare indicators of model deterioration by replacing the value of each feature, $j = 1, 2, 3$, with a continuous vector with evenly spaced numbers over a specified range $(-3, 4)$

	Doubt	NASA	MAPIE	KS	PSI	Pred. Shift	Exp. Shift
Quadratic feature	**0.09**	0.11	**0.09**	0.96	0.95	0.94	0.98
Linear feature	**0.11**	0.12	0.70	0.35	1.39	1.46	1.20
Random feature	**0.34**	0.35	0.42	1.39	1.46	5.58	1.21
Mean	**0.18**	0.20	0.40	1.24	1.29	2.69	1.13

TABLE 5.3: Performance of model monitoring systems for model deterioration for a linear regression model on a synthetic dataset (cf. Figure 5.5). K-S and PSI are the monitoring systems obtained by computing the Kolmogorov-Smirnov test values and the Population Stability Index, respectively, and Doubt is our method. Explanation Shift is the method presented in the previous chapter of the thesis. Best results are shown in bold.

In Figure 5.5 and Table 5.3 we can see what is the impact of the synthetic distribution shift is in terms of model mean squared error. We see that our uncertainty method follows a better goodness-of-fit than both the Kolmogorov-Smirnov test and the Population Stability Index on the input data and and target data. The latter two methods

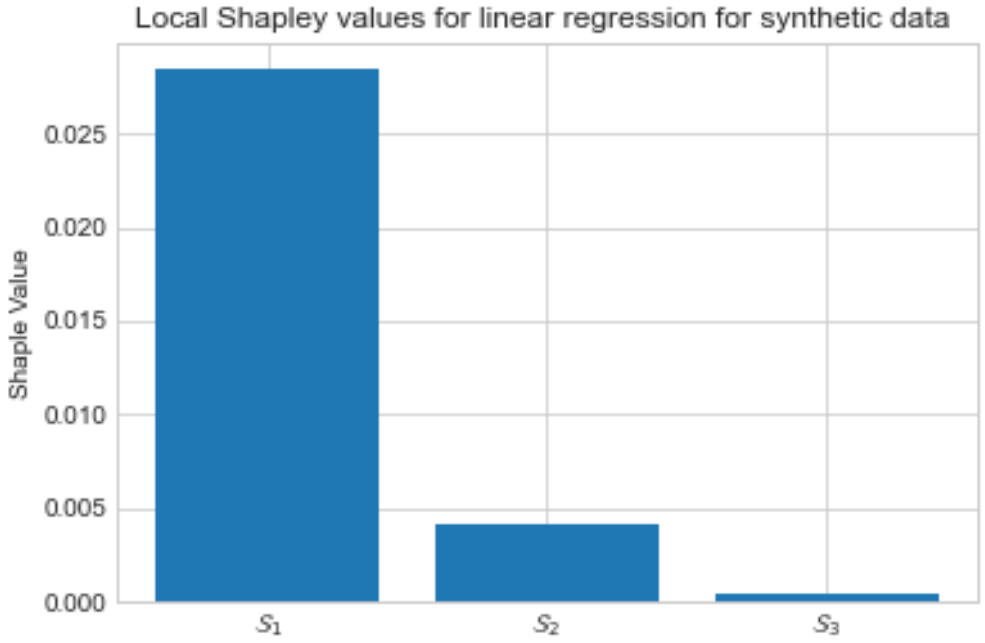

FIGURE 5.4: Shapley values for the data point $x := (10, 10, 10)$ and the model g_θ. Explainable uncertainty estimation allows to account for the drivers of uncertainty that serves as a proxy for model predictive performance degradation.

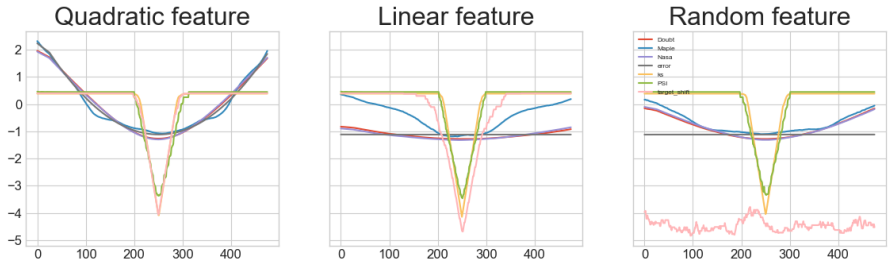

FIGURE 5.5: Comparison of different monitoring methods over the synthetic dataset. Each of the plots represents an independent experiment where each of the features has been replaced by a continuous vector with evenly spaced numbers in the range $(-3, 4)$. The x-axis represents the index of the sorted column. Doubt achieves a better goodness-of-fit than previous statistical methods (cf. Table 5.3)

have identical values for the linear and random features since they are independent of the target values.

5.4.2 Detecting the Source of Uncertainty/Model Deterioration

For this example, we simulate out-of-distribution data by shifting X by 10; i.e., $X_j^{ood} := X_j^{tr} + 10$ for $j = 1, 2, 3$. We train a linear regression model f_θ on (X^{tr}, Y^{tr}) and use Doubt to get uncertainty estimates Z on both X^{te} and X^{ood}. We then train another linear model g_ψ on $((X^{te}, X^{ood}))$ to predict Z. The coefficients of g_ψ are $\beta_1 = 0.03478, \beta_2 =$

0.005023, $\beta_3 = 0.000522$. Since the features are independent and we are dealing with linear regression, the interventional conditional expectation Shapley values can be computed as $\beta_i(x_i - \mu_i)$ Chen et al. (2020), where μ is the mean of the training data . So for the data point $x := (10, 10, 10)$, the Shapley values are $(0.0291, 0.0042, 0.0004)$, where the most relevant shifted feature in the model is the one that receives the highest Shapley value. In this experiment with synthetic data, statistical testing on the input data would have flagged the three feature distributions as equally shifted. With our proposed method, we can identify the more meaningful features towards identifying the source of model predictive performance deterioration. It is worth noting that this explainable AI uncertainty approach can be used with other uncertainty estimation techniques.

5.5 Computational Performance Comparison

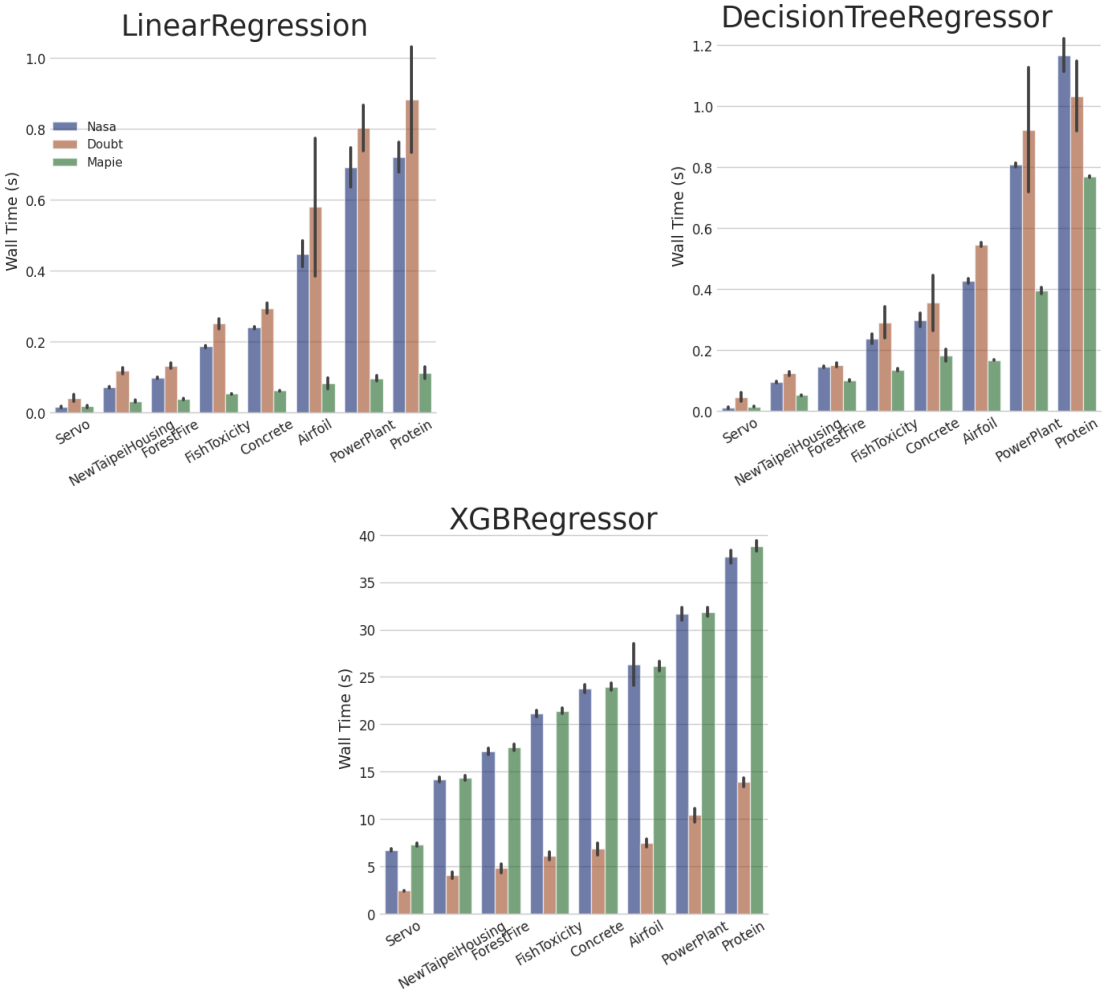

FIGURE 5.6: Wall time computational performance of the three uncertainty estimation methods described through the paper. Datasets are sorted by the number of rows and the y-axis is not shared accross plots.

In this section we compare the wall clock time of the NASA method Kumar and Srivastava (2012), MAPIE Barber et al. (2021) and Doubt with three estimators: linear regression, decision tree and gradient boosting decision tree, over the range of dataset previously used. We set the number of model bootstraps in the Doubt method to be equal to the number of cross-validation splits in the MAPIE method; namely, the square root of the total number of samples in the dataset. The experiment are performed on a virtual machine with 8 vCPUs and 60 GB RAM.

See the results in Figure 5.6. We see here that the MAPIE approach is significantly faster than the NASA and Doubt methods in the cases where the model used is a linear regression model or a decision tree. However, interestingly, in the case where the model is an XGBoost model, the Doubt method is significantly faster than the other two.

TABLE 5.4: Non-aggregated version of Table 5.2. Model monitoring systems for model deterioration for a variety of model architectures on eight regression datasets from the UCI repository.

		Airfoil	Concrete	ForestFire	Parkinsons	PowerPlant	Protein	Bike	FishToxicity	Mean	Std
Linear Regression	Uncertainty	0.67	0.59	0.6	0.6	1.0	0.72	0.86	0.64	0.71	0.14
	MAPIE	0.72	0.79	0.71	0.62	0.85	0.74	0.79	0.94	0.77	0.09
	K-S	0.78	0.67	0.81	0.92	0.8	0.99	0.85	0.71	0.81	0.10
	PSI	1.02	0.9	0.78	0.83	0.93	0.79	0.91	0.81	0.87	0.08
	Target	0.94	0.68	0.7	0.84	0.8	0.97	0.97	1.0	0.86	0.12
Poisson Regressor	Uncertainty	0.81	0.49	0.93	0.82	0.97	0.83	0.79	0.71	0.79	0.14
	MAPIE	0.84	0.55	0.89	0.86	0.87	0.91	0.61	1.1	0.82	0.17
	K-S	0.83	0.68	1.08	1.08	0.8	1.16	1.16	0.79	0.94	0.19
	PSI	1.01	0.9	1.04	1.04	0.94	0.85	0.85	0.86	0.93	0.08
	Target	1.02	0.75	1.15	1.15	0.79	1.1	1.1	0.95	1.00	0.15
Decision Tree	Uncertainty	0.48	0.44	0.74	0.51	0.45	0.44	0.46	0.47	0.49	0.10
	MAPIE	0.59	0.52	0.96	0.61	0.61	0.51	0.49	0.45	0.59	0.15
	K-S	0.49	0.56	0.66	0.49	0.3	0.49	0.68	0.5	0.52	0.11
	PSI	1.14	1.08	1.02	0.9	0.92	0.9	0.86	0.95	0.97	0.09
	Target	0.96	0.77	0.81	0.9	0.64	0.54	0.94	0.79	0.79	0.14
Random Forest	Uncertainty	0.57	0.577	0.93	0.5	0.8	0.92	0.67	0.98	0.74	0.18
	MAPIE	0.76	0.91	0.95	0.58	0.99	1.0	0.75	0.98	0.86	0.15
	K-S	0.48	0.51	0.74	0.41	0.41	0.41	0.62	0.45	0.50	0.11
	PSI	1.11	1.02	1.0	0.89	0.92	0.92	0.85	0.93	0.95	0.08
	Target	0.95	0.7	0.7	0.87	0.61	0.38	0.89	0.8	0.73	0.18
Gradient Boosting	Uncertainty	0.48	0.49	1.0	0.45	0.45	0.32	0.57	0.88	0.58	0.23
	MAPIE	0.66	0.86	0.94	0.55	0.67	0.51	0.64	1.03	0.73	0.18
	K-S	0.45	0.47	0.65	0.9	0.9	0.42	0.64	0.48	0.61	0.19
	PSI	1.11	1.01	0.99	0.91	0.92	0.9	0.85	0.98	0.95	0.08
	Target	1.05	0.73	0.54	0.92	0.65	0.46	0.88	0.8	0.75	0.19
MultiLayer Perceptron	Uncertainty	1.5	0.93	0.81	0.43	0.43	0.42	0.52	0.41	0.68	0.38
	MAPIE	1.6	0.85	0.95	0.49	0.58	0.42	0.55	0.55	0.74	0.38
	K-S	1.03	0.72	1.18	0.61	0.62	0.64	0.62	0.62	0.75	0.22
	PSI	1.1	0.75	1.08	0.76	0.88	0.79	0.74	0.66	0.84	0.16
	Target	0.63	0.8	1.24	0.6	0.59	0.54	0.73	0.78	0.73	0.22

5.6 Conclusion

In this work, we have provided methods and experiments to monitor and identify machine learning model deterioration via non-parametric bootstrapped uncertainty estimation methods and use explainability techniques to explain the source of the model deterioration under gradual covariate shift.

Our monitoring system is based on a model agnostic uncertainty estimation method, which produces prediction intervals with theoretical guarantees and which achieves

better coverage than the current state-of-the-art. The resulting monitoring system more accurately detects model deterioration than methods using classical statistics. Finally, we used SHAP values in conjunction with these uncertainty estimates to identify the features that are driving the model deterioration at both a global and local level, and qualitatively showed that these more accurately detect the source of the model deterioration compared to classical statistical methods.

5.6.1 Limitations:

Detecting performance degradation may be impossible in situations where no labelled data is available and no characterizations of the shift can be made. In this work, we have engineered an experiment of a gradual covariate shift, where uncertainty estimation by model averaging achieves the best result.

We emphasize here that due to computationally limitations, we have only benchmarked on datasets of relatively small to medium size (cf. Table 5.1), and further work needs to be done to see if these results are also valid for datasets of significantly larger size. This work also focused on tabular data and non-deep learning models.

Chapter 6

Conclusion and Remarks

6.1 Summary of Findings

This thesis studies the problem of model monitoring in the absence of labelled deployment data, aiming to improve the understanding of supervised machine learning systems in production. It begins by exploring the AI system life cycle, emphasizing the critical role of model monitoring in ensuring the functionality and reliability of deployed models and the unsupervised monitoring stage.

In Chapter 2, we navigated through the foundations needed for this thesis. Starting from feature attribution distribution using both Shapley values and LIME in Section2.2. In section 2.3 we provide the foundations of AI alignment, drawing basic common understanding from political and moral philosophy. In section 2.4, we reviewed the current state-of-the-art of technical fairness metrics. Finally in section 2.5 we reviewed the problem of model monitoring under distribution shift with its inherent limitations.

Chapter 3 proposed a metric and a model for measuring *equal treatment*, aiming to align AI systems with diverse human values encompassing political philosophy and moral principles. We have examined the principle of equality in machine learning, specifically on liberalism-oriented philosophical notions, emphasizing the importance of equal treatment for all individuals, regardless of their protected characteristics. By proposing a new formalization for equal treatment incorporating feature values' influence through Shapley values, we have sought to provide a better understanding of how can we align machine learning systems with equal treatment notions from the social sciences. The theoretical properties of our formalization have been analyzed, and a classifier two-sample test based on the AUC of an equal treatment inspector has been devised. Our experiments on synthetic and natural datasets have demonstrated the efficacy of our formalization, shedding light on the complexities of achieving equality in machine learning.

Chapter 4 proposed to usage of explanation distributions shift to better detect changes in the model due to distribution shifts.From the predictive performance perspective, we have investigated the challenges posed by distribution shifts, which can significantly impact the performance of machine learning models on new data. By introducing the concept of explanation shift, we have developed a sensitive and explainable indicator to monitor changes in model behavior resulting from distribution shifts. Through theoretical analysis and empirical evaluations, we have demonstrated the effectiveness of explanation shifts and facilitated understanding of the interaction between distribution shifts and learned models.

Chapter 5 distinguishes between understanding changes in model behaviour and building indicators of performance degradation which remains an impossible task in many scenarios. We engineered an experiment where there is a gradual covariate shift and proposed model agnostic uncertainty estimation techniques to build an indicator of model performance degradation. Finally, we account for the drivers of model deterioration, training another model to predict the uncertainty, and explaining this model.

Finally, in Chapter 6, we summarise the findings, make the contributions explicit and propose future work directions.

6.2 Contributions to the Field

The contributions made in this thesis help to tackle some of the challenges of model monitoring in machine learning, particularly in the context of deployment without labelled data. By defining and making explicit some the intricacies of post deployment monitoring, it flags the importance of unsupervised model monitoring to maintain the functionality and reliability of machine learning systems in production. The exploration of feature attribution distributions for AI alignment and technical fairness metrics, along with the challenges of model monitoring under distribution shift, provided techniques and methods to monitor models in production before the predictions are served in the real world.

We now summarize the key findings for each of the research questions posed.

6.2.1 (RQ-1) How can we use feature attributions distributions for post-deployment monitor in the absence of labeled truth data?

We addressed this issue by segmenting it into performance monitoring and AI alignment monitoring, and divide into two different research question aiming to explore in depth.

This thesis's technical solution leverages feature attribution distributions for model monitoring in unlabeled data scenarios. These distributions represent a projection of the model and data into a distribution space that retains distinct theoretical properties, depending on the feature attribution method utilized. We explored LIME and Shapley feature attribution methods.

6.2.2 (RQ-2) Can feature attribution distributions be used to measure alignment of post-deployed models in a unsupervised monitoring set up??

We initiated this by gathering research requirements from two specific philosophical perspectives—liberal-oriented political philosophy and deontological ethics—to align with the notion of equal treatment.

We introduced a new metric that improved existing equal treatment measures based on the collected requirements. By integrating feature attribution methods and philosophical principles, our proposed metric provides a more precise approach to measuring AI systems alignment with human values of equality and fairness.

6.2.3 (RQ-3) How did Distribution Shift Impact the Model?

We studied the interactions between deployed models and shifting data distributions, introducing the concept of explanation shift to detect and understand changes in model behaviour due to distribution shifts. This method offers a novel perspective on maintaining consistent model behaviour over time.

6.2.4 (RQ-4) How can we identify and explain changes in model performance using feature attribution distribution?

We proposed the use of Shapley values in conjunction with a second model that predicts uncertainty estimates to identify the features driving model deterioration at both a global and local level.

6.2.5 Software Contributions

The developed software, `skshift` and `explanationspace`, stands as a practical implementation of these contributions, offering researchers and practitioners a valuable tool for model monitoring. This work not only advances our understanding of model monitoring in unlabelled environments but also sets a new standard for the alignment of AI systems with ethical principles and human values.

To ensure reproducibility, we have made the data, code repositories, and experiments publicly available. The data, code, and tutorials can be accessed at the following URLs:

- **Explanation Shift**, Dataset and code repositories: `https://github.com/cmougan/ExplanationShift`
- Open-source Python package with documentation and tutorials `skshift`: `https://skshift.readthedocs.io`
- **Equal Treatment**, Data, data preparation routines, and code repositories: `https://github.com/cmougan/xAIAuditing`
- Open-source Python package with documentation and tutorials `explanationspace`: `https://explanationspace.readthedocs.io`
- Data, data preparation routines, and code repositories for model monitoring with uncertainty estimation `https://github.com/cmougan/MonitoringUncertainty`

For our experiments, we used default parameters from the `scikit-learn` library Pedregosa et al. (2011), unless stated otherwise. The system requirements and software dependencies for reproducing our experiments are provided. All experiments were conducted on a 4 vCPU server with 32 GB RAM.

6.2.6 Limitations and Reflections

As this thesis draws to a close, we now reflect on its scope, limitations, and areas requiring further exploration. While, I believe, the proposed methodologies and contributions are significant, they come with challenges and constraints that require reflection to provide a holistic view of the work's impact and applicability.

6.2.6.1 Reliability of Explanations

This thesis introduces an integrated framework for model monitoring that combines value alignment and performance maintenance in deployment scenarios. A central component of this framework is the reliance on algorithmic explanations, particularly feature attribution methods, to assess model behavior and detect shifts in explanation distributions. However, it relies on explanations.

Feature attribution explanation methods, such as those produced by LIME or Shapley values, are approximations and can introduce new biases. As previously stated in the thesis (cf. Section 2), previous research has focused on studying the reliability of

local explanations; for example, LIME's reliance on local perturbations may inadvertently favour majority groups due to its neighbourhood sampling strategy. However, when studying distributions, SHAP and LIME achieved very similar results, seeming to overcompensate those biases. Further unexplored biases could compromise the fairness assessments derived from distributions of explanation-based methods. Therefore, a deeper technical exploration of these explanations is necessary. Can distributions of algorithmic explanations be considered objective or reliable enough to guide fairness and performance decisions? This question remains partially unanswered and warrants further investigation.

6.2.6.2 Alternative Explanation Methods

Feature attribution methods, such as those used in this thesis to monitor ML models after deployment, can be useful tools. Although this thesis predominantly focuses on Shapley values and LIME, other explanation techniques, such as counterfactual explanations or gradient-based methods, might offer complementary or superior insights in certain scenarios. Each method has its own set of assumptions, strengths, and weaknesses, which should be considered when selecting a monitoring strategy.

6.2.6.3 Beyond Tabular Data

Moreover, while the two explored feature attribution methods proposed in this thesis, have demonstrated efficacy in both theoretical and experimental settings, it might benefit from further validation across a wider range of models, data types, or deployment environments. The adaptability of these methods to different architectures, such as transformer models or non-tabular data like images and text, remains an open question. Future research could aim to generalize these techniques to broader contexts and evaluate their performance comprehensively.

6.2.6.4 Complexity and Practicality of Feature Attribution-Based Monitoring

Incorporating feature attribution methods into model monitoring introduces an additional layer of complexity. While this complexity may be justified in some scenarios where complexity is needed, it may not always be necessary or practical. Mathematically simpler or less computationally intensive methods might suffice in certain cases, particularly where the risk of model failure is lower or where interpretability is not a primary concern.

Furthermore, this work has been developed under specific assumptions, including controlled experimental setups and the availability of computational resources. These assumptions are explicitly stated to avoid overgeneralization. For instance, the impact of

population imbalances or other real-world data challenges has not been thoroughly investigated. The findings should, therefore, be framed as promising but context-dependent, highlighting the need for cautious interpretation and application.

6.3 Future Directions

As we conclude this thesis, we highlight potential avenues for future research and development in model monitoring in general and in each of the chapter presented on this thesis.

6.3.1 Software Improvement

One avenue for future work involves expanding and refining the *Equal Treatment Inspector* and the *Explanation Shift Detector* software, allocated in the Python package `explanationspace` and `skshift` respectively. Currently, it provides insights into the features with different distributions of explanations, helping identify unequal treatment and distribution shift that impacts the model. Improvementts could include the development of a more user-friendly interface, incorporating statistical significance tests to validate differences in explanation distributions, extending to non-tabular data and exploring additional visualization techniques to facilitate a deeper understanding of the root sources of unequal treatment.

6.3.2 Integration with Real-world Applications

To validate the practical utility of the proposed approach, future work can potentially involve integrating the *Equal Treatment Inspector* and *Explanation Shift Detector* into real-world machine learning applications. This would entail applying the method to diverse datasets in various domains, such as finance, healthcare, or criminal justice, to assess its effectiveness in uncovering and addressing disparate treatment issues. Additionally, feedback from practitioners and stakeholders in these domains could be incorporated to improve the tool's applicability and relevance.

6.3.3 Ethical and Societal Implications

Future research should also delve into the ethical and societal implications of incorporating political liberal and Kantian deontological notions into machine learning fairness. This involves exploring how aligning AI systems with the principles of treating individuals as ends in themselves and equally might impact various stakeholders, including marginalized communities, policymakers, and the general public. Addressing

these implications is crucial, as their implications within a machine learning context might be different than the ones from the social sciences.

6.3.4 Longitudinal Studies

Longitudinal studies could be conducted to assess the long-term impact and effectiveness of the proposed approach. These studies would involve monitoring the deployment of machine learning models incorporating the *Equal Treatment Inspector* and *Explanation Shift Detector* over an extended period, evaluating its ability to mitigate disparate treatment, and adapting the tool based on real-world feedback and evolving ethical standards.

By addressing these future directions, researchers can contribute to the ongoing discourse on fairness in machine learning and foster the development of ethically sound and socially responsible AI systems.

List of Figures

1.1 Model monitoring occurs within the AI system life cycle. During model development, task requirements and data are collected, models are trained and evaluated, and parameters are refined. Finally, the model is assessed against benchmarks and metrics using a hold-out dataset. 1
1.2 Life cycle of monitoring a supervised learning binary classifier. Once a model is deployed, it encounters new, unlabelled data, prompting the need to observe and analyse its behaviour to detect potential issues before serving the predictions in the real world. In some cases, after serving the model to users, the acquisition of ground truth data allows for further evaluation. Leveraging these results and the available ground truth data, the original model can undergo retraining. 2

3.1 Equal Treatment Inspector workflow. The model f_θ is learned based on training data, $\mathcal{D} = \{(x_i, y_i)\}$, and outputs the explanations $\mathcal{S}(f_\theta, \mathcal{D}_X)$. The Classifier Two-Sample Test receives the explanations to predict the protected attribute, Z. The AUC of the two-sample test classifier g_ψ decides for or against *equal treatment*. We can interpret the driver for unequal treatment on g_ψ with explainable AI techniques 39
3.2 Comparing the power (the higher, the better) of C2ST based on AUC with Brunner-Munzel test (AUC Test BM) vs Accuracy vs AUC with an asymptotic normal approximation of the Wilcoxon–Mann–Whitney statistics (AUC Test A). Upper: balanced groups ($P(Z = 1) = 0.5$). Lower: unbalanced groups ($P(Z = 1) = 0.2$). 46
3.3 Coefficient of g_ψ over γ for synthetic datasets in two experimental scenarios. 48
3.4 In the figure above, "Indirect case": All distributions capture this fairness violation; only the prediction distributions are less sensitive due to their low dimensionality. In the figure below, "Uninformative case", only the explanation distributions and prediction distributions detect that the model is non-discriminant, input data flags a false positive detection. . . 51
3.5 AUC of the ET inspect using SHAP vs using LIME. 54
3.6 In the figure abobe, a comparison of Equal Treatment and Demographic Parity measures on the US Income data. The AUC range for Equal Treatment is notably wider, and aligning with the theoretical section, there are indeed instances where Demographic Parity fails to identify discrimination that Equal Treatment successfully detects. For a detailed statistical analysis, please refer to Appendix 3.6.2. The figure below provides insight into the influential features contributing to unequal treatment. Higher feature values correspond to a greater likelihood of these features being the underlying causes of unequal treatment. 56

3.7 Above: AUC of the inspector for Equal Treatment and DP, over the district of California 2014 for the ACS Employment dataset. Below: contribution of features to the Equal Treatment inspector performance. 57
3.8 Above AUC of the inspector for Equal Treatment and DP, over the district of California 2014 for the ACS Travel Time dataset. Below: contribution of features to the Equal Treatment inspector performance. 58
3.9 Above: AUC of the inspector for Equal Treatment and DP, over the district of California 2014 for the ACS Mobility dataset. Below: contribution of features to the Equal Treatment inspector performance. 59
3.10 AUC of the inspector for ET, over the district of CA14 for the US Income dataset. The model f_θ is a decision tree based algorithm, that we modify the maximum depth (x-axis), while the model g_ψ varies in each of the images. The image above shows the results using a logistic regression detector, while the image below uses an XGBoost detector. By changing the model, we can see that simpler models (decision tree with depth 1 or 2) are less discriminative, while when increasing the model complexity, the unequal treatment of the model starts taking higher values. 62

4.1 Our model for explanation shift detection. The model f_θ is trained on $\mathcal{D}^{tr}$ implying explanations for distributions $\mathcal{D}_X^{val}, \mathcal{D}_X^{new}$. The AUC of the two-sample test classifier g_ψ decides for or against explanation shift. If an explanation shift occurred, it could be explained which features of the $\mathcal{D}_X^{new}$ deviated in f_θ compared to $\mathcal{D}_X^{val}$. 75
4.2 In the figure above, we compare the Classifier Two-Sample Test on explanation distribution (ours) versus input distribution *(B6)* and prediction distribution *(B7)*. Explanation distribution shows the highest sensitivity. The figure below, related work comparison of distribution shift methods*(B1-B5)*, as the experimental setup has a gradual distribution shift, good indicators should follow a progressive steady positive slope, following the correlation coefficient, as our method does. In Table 4.5 we provide a quantitative evaluation. 83
4.3 Novel group shift experiment on the US Income dataset. Sensitivity (AUC) increases with the growing fraction of previously unseen social groups. *Left figure*: The explanation shift indicates that different social groups exhibit varying deviations from the distribution on which the model was trained (White). *Middle Figure*: We vary the model f_θ by training it using both `XGBoost` (solid lines) and Logistic Regression (dots). The novel ethnicity group is Black. We compare Explanation Shift against C2ST on input *(B6)* and output *(B7)*. *Right figure*: Comparison of Explanation Shift against Exp. NDCG *(B4)*. We see how monitoring method*(B4)* is more unstable with a linear model, and with an `xgboost` it erroneously finds a horizontal asymptote. We don't compare against methods relying purely on input data such as *(B1)* as we are changing the model, which they don't take into consideration. 85

4.4 Novel group shift experiment conducted on the 4 Datasets Comparison to NDCG. Sensitivity (AUC) increases as the proportion of previously unseen social groups grows. As the experimental setup has a gradual distribution shift, ideal indicators should exhibit a steadily increasing slope. However, in all figures, NDCG exhibits saturation and instability. These observations align with the analysis presented in the synthetic experiment section, as discussed in Section 4.3.2 of the main paper . . . 86
4.5 Both images represent the AUC of the *Explanation Shift Detector* for different countries on the StackOverflow survey dataset under novel group shift. In the left image, the estimator, f_θ, is a gradient boosting decision tree; in the right image, for both cases the detector, g_ψ, is a logistic regression. By changing the type of estimator model, we can see how different types of models are affected differently for the same distribution shift . . 87
4.6 In the above figure, a comparison of the performance of *Explanation Shift Detector* in different states. In the below figure, the strength analysis of features driving the change in the model, on the y-axis are the features, and on the x-axis are the different states. Explanation shifts allow us to identify how the distribution shift of different features impacted the model. 88
4.7 The above figure shows a comparison of the performance of the Explanation Shift Detector in different states for the ACS Employment dataset. The below figure shows the feature importance analysis for the same dataset. 90
4.8 In the left figure, comparison of the performance of *Explanation Shift Detector*, in different states for the ACS TravelTime prediction task. In the left figure, we can see how the state with the highest OOD AUC detection is KS18 and not PR18 as in other prediction tasks; this difference with respect to the other prediction task can be attributed to "Place of Birth", whose feature attributions the model finds to be more different than in CA14. 91
4.9 Left figure shows a comparison of the *Explanation Shift Detector*'s performance in different states for the ACS Mobility dataset. Except for PR18, all other states fall below an AUC of explanation shift detection of 0.70. The features driving this difference are Citizenship and Ancestry relationships. For the other states, protected social attributes, such as Race or Marital status, play an important role. 92
4.10 Images represent the AUC of the *Explanation Shift Detector*, on ACS Income and last two Stackoverflow under novel group shift. In the image above, the detector is a logistic regression, and in the images below, it is a gradient-boosting decision tree classifier. By changing the model, we can see that vanilla models (decision tree with depth 1 or 2) are unaffected by the distribution shift, while when increasing the model complexity, the out-of-distribution impact of the data in the model starts to be tangible 95
4.11 Images represent the AUC of the *Explanation Shift Detector*, on Stackoverflow under novel group shift. In the image above, the detector is a logistic regression, and in the image below, it is a gradient-boosting decision tree classifier. By changing the model, we can see that vanilla models (decision tree with depth 1 or 2) are unaffected by the distribution shift, while when increasing the model complexity, the out-of-distribution impact of the data in the model starts to be tangible 96

4.12 Comparison of the explanation distribution generated by LIME and SHAP. The above figure shows the sensitivity of the predicted probabilities to multi covariate changes using the synthetic data experimental setup of 4.2 on the main body of the paper. The figure below shows the distribution of explanation shifts for a New Covariate Category shift (Asian) in the ACS Income dataset. 97
4.13 Wall time for generating explanation distributions using SHAP and LIME with different numbers of samples (above) and different numbers of columns (below). Note that the y-scale is logarithmic. The experiments were run on a server with 4 vCPUs and 32 GB of RAM. The runtime required to create an explanation distributions with LIME is far greater than SHAP for a gradient-boosting decision tree 99

5.1 Comparison of different model degradation detection methods for the Fish Toxicity dataset. Each of the plots represents an independent experiment where each of the six features has been shifted, using the method described in the methodology section. Doubt achieves a better goodness-of-fit than previous statistical methods. A larger version of this figure can be found in Appendix. 112
5.2 Global comparison of different distribution shift detection methods. Statistical methods correctly indicate that there exists a distribution shift in the shifted data. Shapley values indicate the contribution of each feature to the drop in predictive performance of the model. 114
5.3 Individual explanation that displays the source of uncertainty for one instance. The previous method allowed only for comparison between distributions, now with explainable uncertainty, we are able to account for individual instances. In red, features pushing the uncertainty prediction higher are shown; in blue, those pushing the uncertainty prediction lower. 115
5.4 Shapley values for the data point $x := (10, 10, 10)$ and the model g_θ. Explainable uncertainty estimation allows to account for the drivers of uncertainty that serves as a proxy for model predictive performance degradation. 116
5.5 Comparison of different monitoring methods over the synthetic dataset. Each of the plots represents an independent experiment where each of the features has been replaced by a continuous vector with evenly spaced numbers in the range $(-3, 4)$. The x-axis represents the index of the sorted column. Doubt achieves a better goodness-of-fit than previous statistical methods (cf. Table 5.3) 116
5.6 Wall time computational performance of the three uncertainty estimation methods described through the paper. Datasets are sorted by the number of rows and the y-axis is not shared accross plots. 117

List of Tables

1.1 Requirements Elicitation Table for Equal Treatment 7

2.1 Notation used throughout the paper. 11
2.2 Summary of Technical Definitions of Fairness Metrics 28

3.1 Results of the C2ST on the "Equal Treatment Inspector". 45
3.2 AUC comparison of the "Explanation Shift Detector" between estimating the Shapley values between the interventional and the correlation-dependent approaches for the four prediction tasks based on the US census dataset Ding et al. (2021a). The % character represents the relative difference. The performance differences are negligible. 52
3.3 Linear regression coefficients comparison of the "Explanation Shift Detector" between estimating the Shapley values between the interventional and the correlation-dependent approaches for one of the US census based prediction tasks (ACS Income). The % character represents the relative difference. The coefficients show negligible differences between the calculation methods 52
3.4 Comparison ofEqual Treatment and Demographic Parity measured in different ways. The cases of equal Treatment violation but no demographic parity violation are highlighted in red. 60
3.5 AUC of theEqual Treatmentinspector for different combinations of models and inspectors. 61
3.6 Fairness metrics alignment with *Equal Treatment* criteria. Requirements are derived from a use case explained in Section 3.1.1, and metrics are further discussed and defined in Section 2.4 65

4.1 Displayed results are the one-tailed p-values of the Kolmogorov-Smirnov test comparison between two underlying distributions. Small p-values indicate that compared distributions would be very unlikely to be equally distributed. SHAP values correctly indicate the interaction changes that individual distribution comparisons cannot detect 79
4.2 Distribution comparison for synthetic concept shift. Displayed results are the one-tailed p-values of the Kolmogorov-Smirnov test comparison between two underlying distributions 80
4.3 Distribution comparison when modifying a random noise variable on test data. The input data shifts while explanations and predictions do not. 80
4.4 Distribution comparison over how the change on the contributions of each feature can cancel out to produce an equal prediction (cf. Section 4.2.2), while explanation shift will detect this behaviour changes on the predictions will not. 81

4.5 Pearson Correlation of the correlation coefficient ρ and baseline methods, extending Figure 4.2. Explanation Shift achieves better covariate shift detection on synthetic data. 84
4.6 Pearson correlation between baselines and the ratio of presence of the unseen group. The *Accountable* column relates to how well a method, like the Explanation Shift Detector, is underpinned by theoretical guarantees (e.g., Shapley value properties) and supported by empirical evidence (e.g., synthetic tests and real-world validations), as discussed in Section 4.2. 84
4.7 Comparison of explanation shift detection performance, measured by AUC, for different combinations of explanation shift detectors and estimators on the UCI Adult Income dataset using the Novel Covariate Group Shift for the "Asian" group with a fraction ratio of 0.5 (cf. Section 4.3). The table shows that the algorithmic choice for f_θ and g_ψ can impact the OOD explanation performance. We can see how, for the same detector, different f_θ models flag different OOD explanations performance. On the other side, for the same f_θ model, different detectors achieve different results. 93

5.1 Statistics of the regression datasets used in this paper. 111
5.2 Performance of model monitoring systems for model deterioration for a variety of model architectures on eight regression datasets from the UCI repository Dua and Graff (2017b). The scores are the means and standard deviations of the absolute deviation from the true labels on $\mathcal{D}^{lower}$ and $\mathcal{D}^{upper}$ (lower is better). K-S and PSI are the monitoring systems obtained by computing the Kolmogorov-Smirnov test values and the Population Stability Index, respectively, Prediction Shift is the statistical comparison of the model prediction, and Doubt is our method. The best results for each model architecture are shown in bold. See the Appendix for all the raw scores. 113
5.3 Performance of model monitoring systems for model deterioration for a linear regression model on a synthetic dataset (cf. Figure 5.5). K-S and PSI are the monitoring systems obtained by computing the Kolmogorov-Smirnov test values and the Population Stability Index, respectively, and Doubt is our method. Explanation Shift is the method presented in the previous chapter of the thesis. Best results are shown in bold. 115
5.4 Non-aggregated version of Table 5.2. Model monitoring systems for model deterioration for a variety of model architectures on eight regression datasets from the UCI repository. 118

References

Aas, K., Jullum, M., and Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intellience, Elsevier*, 298:103502.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., and Kim, B. (2018a). Sanity checks for saliency maps. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (Conference on Neural Information Processing Systems (NeurIPS))*, pages 9525–9536.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., and Kim, B. (2018b). Sanity checks for saliency maps. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 9525–9536.

Adebayo, J., Muelly, M., Liccardi, I., and Kim, B. (2020). Debugging tests for model explanations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (Neural Information Processing Systems'20)*, Neural Information Processing Systems'20, Red Hook, NY, USA. Curran Associates Inc.

Alvarez, J. M., Colmenarejo, A. B., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., Ghodsi, S., Mougan, C., Papageorgiou, I., Reyero, P., Russo, M., Scott, K. M., State, L., Zhao, X., and Ruggieri, S. (2024). Policy advice and best practices on bias and fairness in ai. *Ethics and Information Technology*.

Alvarez, J. M., Scott, K. M., Ruggieri, S., and Berendt, B. (2023). Domain adaptive decision trees: Implications for accuracy and fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery.

Alvarez-Melis, D. and Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (Conference on Neural Information Processing Systems (NeurIPS))*, pages 7786–7795.

Arnez, F., Espinoza, H., Radermacher, A., and Terrier, F. (2020). A comparison of uncertainty estimation approaches in deep learning components for autonomous vehicle applications.

Arrieta, A. B., Rodríguez, N. D., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58:82–115.

Ayvaz, D. S., Belenguer, L., He, H., Kanubala, D. D., Li, M., Low, S., Mougan, C., Onwuegbuche, F. C., Pi, Y., Sikora, N., Tran, D., Verma, S., Wang, H., Xie, S., and Pelletier, A. (2025). Measuring fairness in financial transaction machine learning models.

Baek, C., Jiang, Y., Raghunathan, A., and Kolter, J. Z. (2022). Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Bala, M., Monica S., V., Horibe, Mayumi, E., J., M., Trevor, A., David E., L., King-Wai, L. D., Shu-Fangand Kim Jerry, N., and Su-In, L. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(1):749–760.

Balestra, C., Li, B., and Müller, E. (2022). Enabling the visualization of distributional shift using shapley values. In *Conference on Neural Information Processing Systems (NeurIPS) 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Procedings ACM Human Computing Interaction*, 5(CSCW1):74:1–74:34.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507.

Barrabés, M., Montserrat, D. M., Geleta, M., i Nieto, X. G., and Ioannidis, A. G. (2023). Adversarial learning for feature shift detection and correction. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems*.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Ben-David, S., Lu, T., Luu, T., and Pál, D. (2010). Impossibility theorems for domain adaptation. In Teh, Y. W. and Titterington, D. M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pages 129–136. JMLR.org.

Bentham, J. (1996). *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation*. Clarendon Press.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G. G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., and Xiang, A. (2021). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty.

Bilodeau, B. L., Jaques, N., Koh, P. W., and Kim, B. (2022). Impossibility theorems for feature attribution. *CoRR*, abs/2212.11870.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *FAT*, volume 81 of *Proceedings of Machine Learning Research*, pages 149–159. PMLR.

Bordt, S. and von Luxburg, U. (2023). From shapley values to generalized additive models and back. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, pages 709–745. PMLR.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2022). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2021). Deep neural networks and tabular data: A survey.

Budhathoki, K., Janzing, D., Blöbaum, P., and Ng, H. (2021). Why did the distribution change? In Banerjee, A. and Fukumizu, K., editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 1666–1674. PMLR.

Burkov, A. (2020). *Machine Learning Engineering*. Kindle Direct Publishing, 1 edition.

Candelas, J. Q., Wu, Y., Hsu, B., Jain, S., Ramos, J., Adams, J., Hallman, R., and Basu, K. (2023). Disentangling and operationalizing AI fairness at linkedin. In *FAccT*, pages 1213–1228. ACM.

Carlini, N. and Wagner, D. (2017). *Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods*, page 3–14. Association for Computing Machinery, New York, NY, USA.

Castillo-Eslava, F., Mougan, C., Romero-Reche, A., and Staab, S. (2023). The role of large language models in the recognition of territorial sovereignty: An analysis of the construction of legitimacy. *CoRR*, abs/2304.06030.

Chakravarti, P., Kuusela, M., Lei, J., and Wasserman, L. (2023). Model-independent detection of new physics signals using interpretable semisupervised classifier tests. *The Annals of Applied Statistics*, 17(4):2759–2795.

Chang, P. W., Fishman, L., and Neel, S. (2023). Model explanation disparities as a fairness diagnostic. *CoRR*, abs/2303-01704.

Chen, H., Covert, I. C., Lundberg, S. M., and Lee, S. (2022a). Algorithms to estimate shapley value feature attributions. *CoRR*, abs/2207.07605.

Chen, H., Janizek, J. D., Lundberg, S. M., and Lee, S. (2020). True to the model or true to the data? *CoRR*, abs/2006.16234.

Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. (2019). L-shapley and c-shapley: Efficient model interpretation for structured data. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Chen, L., Zaharia, M., and Zou, J. Y. (2022b). Estimating and explaining model performance when both covariates and labels shift. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Chen, T. and Guestrin, C. (2016a). Xgboost: A scalable tree boosting system. In Krishnapuram, B., Shah, M., Smola, A. J., Aggarwal, C. C., Shen, D., and Rastogi, R., editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 785–794. ACM.

Chen, T. and Guestrin, C. (2016b). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

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

Commission, U. E. E. O. (1964). Title vii of the civil rights act of 1964. Public Law 88-352, 78 Stat. 241, 42 U.S.C. § 2000e et seq.

Community, T. T. W., Arnold, B., Bowler, L., Gibson, S., Herterich, P., Higman, R., Krystalli, A., Morley, A., O'Reilly, M., and Whitaker, K. (2019). The Turing Way: A Handbook for Reproducible Data Science.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 797–806. ACM.

Cortes, C. and Mohri, M. (2004). Confidence intervals for the area under the ROC curve. In *Neural Information Processing Systems*, pages 305–312.

Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. (2008). Sample selection bias correction theory.

Crabbé, J., Qian, Z., Imrie, F., and van der Schaar, M. (2021). Explaining latent representations with a corpus of examples. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 12154–12166.

D'Angelo, F. and Henning, C. (2021). Uncertainty-based out-of-distribution detection requires suitable function space priors.

Datta, A., Sen, S., and Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE Symposium on Security and Privacy*, pages 598–617. IEEE Computer Society.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

Diethe, T., Borchert, T., Thereska, E., Balle, B., and Lawrence, N. (2019). Continual learning in practice. ArXiv preprint, https://arxiv.org/abs/1903.05202.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021a). Retiring adult: New datasets for fair machine learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 6478–6490.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021b). Retiring adult: New datasets for fair machine learning. *CoRR*, abs/2108.04884.

Driver, J. (2022). The History of Utilitarianism. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

Dua, D. and Graff, C. (2017a). UCI machine learning repository.

Dua, D. and Graff, C. (2017b). UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, http://archive.ics.uci.edu/ml.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2012a). Fairness through awareness. In Goldwasser, S., editor, *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226. ACM.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2012b). Fairness through awareness. In *ITCS*, pages 214–226. ACM.

Edwards, H. and Storkey, A. J. (2016). Censoring representations with an adversary. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Ellis, C. M. (2023). Kaggle: What is adversarial validation? https://www.kaggle.com/code/carlmcbrideellis/what-is-adversarial-validation. Accessed 03/04/23.

Elsayed, S., Thyssens, D., Rashed, A., Jomaa, H. S., and Schmidt-Thieme, L. (2021a). Do we really need deep learning models for time series forecasting? *arXiv preprint arXiv:2101.02118*.

Elsayed, S., Thyssens, D., Rashed, A., Schmidt-Thieme, L., and Jomaa, H. S. (2021b). Do we really need deep learning models for time series forecasting? *CoRR*, abs/2101.02118.

Engels, F. (2023). Condition of the working class in England. In *British Politics and the Environment in the Long Nineteenth Century*, pages 31–38. Routledge.

Fang, Z., Li, Y., Lu, J., Dong, J., Han, B., and Liu, F. (2022). Is out-of-distribution detection learnable? In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Fazelpour, S. and Lipton, Z. C. (2020). Algorithmic fairness from a non-ideal perspective. In *AIES*, pages 57–63. ACM.

Feng, J., Subbaswamy, A., Gossmann, A., Singh, H., Sahiner, B., Kim, M.-O., Pennello, G., Petrick, N., Pirracchio, R., and Xia, F. (2023). Towards a post-market monitoring framework for machine learning-based medical devices: A case study. In *Conference on Neural Information Processing Systems (NeurIPS) 2023 Workshop on Regulatable ML*.

Fleisher, W. (2021). What's fair about individual fairness? In *AIES*, pages 480–490. ACM.

Fligner, M. A. and Policello, G. E. (1981). Robust rank procedures for the behrens-fisher problem. *Journal of the American Statistical Association*, .(373):162–168.

Fort, S., Ren, J., and Lakshminarayanan, B. (2021). Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34.

Friedman, M. (2022). Wikipedia,the free encyclopedia. [Online; accessed 12-May-2023].

Friedman, M., Friedman, R. D., and Friedman, R. D. (1990). *Free to choose*. Free to Choose Enterprise.

Frye, C., Rowat, C., and Feige, I. (2020). Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, Conference on Neural Information Processing Systems (NeurIPS) 2020, December 6-12, 2020, virtual*.

Fumagalli, F., Muschalik, M., Kolpaczki, P., Hüllermeier, E., and Hammer, B. E. (2023). SHAP-IQ: Unified approximation of any-order shapley interactions. In *Thirty-seventh Conference on Neural Information Processing Systems (Conference on Neural Information Processing Systems (NeurIPS) 2023)*.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds Mach.*, 30(3):411–437.

Gabriel, I. (2022). Toward a theory of justice for artificial intelligence. *Daedalus*, 151(2):218–231.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning.

Gao, W. and Zhou, Z. (2015). On the consistency of AUC pairwise optimization. In *IJCAI*, pages 939–945. AAAI Press.

Garg, S., Balakrishnan, S., Kolter, Z., and Lipton, Z. (2021). Ratt: Leveraging unlabeled data to guarantee generalization. In *International Conference on Machine Learning*, pages 3598–3609. PMLR.

Garg, S., Balakrishnan, S., Lipton, Z. C., Neyshabur, B., and Sedghi, H. (2022). Leveraging unlabeled data to predict out-of-distribution performance. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. (2020). A unified view of label shift estimation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3290–3300. Curran Associates, Inc.

Ghorbani, A. and Zou, J. Y. (2019). Data shapley: Equitable valuation of data for machine learning. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2242–2251. PMLR.

Gonçalves, L., Subtil, A., Oliveira, M. R., and de Zea Bermudez, P. (2014). Roc curve estimation: An overview. *REVSTAT-Statistical journal*, 12(1):1–20.

Good, P. I. and Hardin, J. W. (2012). *Common errors in statistics (and how to avoid them)*. John Wiley & Sons.

Grabowicz, P. A., Perello, N., and Mishra, A. (2022). Marrying fairness and explainability in supervised learning. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 1905–1916. ACM.

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *AAAI*, pages 51–60. AAAI Press.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022a). Why do tree-based models still outperform deep learning on typical tabular data? In *Conference on Neural Information Processing Systems (NeurIPS)*.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022b). Why do tree-based models still outperform deep learning on typical tabular data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022c). Why do tree-based models still outperform deep learning on typical tabular data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guerin, J., Delmas, K., Ferreira, R. S., and Guiochet, J. (2022). Out-of-distribution detection is not all you need. In *Conference on Neural Information Processing Systems (NeurIPS) ML Safety Workshop*.

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

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018b). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Guschin, A., Ulyanov, D., Trofimov, M., Altukhov, D., and Michaidilis, M. (2018). How to win a data science competition: Learn from top kagglers - national research university higher school of economics. https://www.coursera.org/lecture/competitive-data-science/categorical-and-ordinal-features-qu1TF. Accessed 02/11/20.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Haug, J., Braun, A., Zürn, S., and Kasneci, G. (2022). Change detection for local explainability in evolving data streams. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 706–716.

Haug, J. and Kasneci, G. (2021). Learning parameter distributions to detect concept drift in data streams. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9452–9459. IEEE.

He, X., Pan, J., Jin, O., Xu, T., Liu, B., Xu, T., Shi, Y., Atallah, A., Herbrich, R., Bowers, S., et al. (2014). Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, pages 1–9.

Heckman, J. (1990). Varieties of selection bias. *American Economic Review*, 80(2):313–18.

Heidari, H., Loi, M., Gummadi, K. P., and Krause, A. (2019). A moral framework for understanding fair ML through economic models of equality of opportunity. In danah boyd and Morgenstern, J. H., editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 181–190. ACM.

Hendrycks, D. and Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Hinder, F., Artelt, A., Vaquet, V., and Hammer, B. (2022). Contrasting explanation of concept drift. In *ESANN*.

Hsu, B., Mazumder, R., Nandy, P., and Basu, K. (2022). Pushing the limits of fairness impossibility: Who's the fairest of them all? In *Conference on Neural Information Processing Systems (NeurIPS)*.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. (2007). Correcting sample selection bias by unlabeled data. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Huang, R., Geng, A., and Li, Y. (2021). On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, Conference on Neural Information Processing Systems (NeurIPS) 2021*, abs/2110.00218.

Huyen, C. (2022). *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*. O'Reilly.

Jiang, Y., Nagarajan, V., Baek, C., and Kolter, J. Z. (2021). Assessing generalization of sgd via disagreement. In *International Conference on Learning Representations (ICLR)*.

Kamishima, T., Akaho, S., and Sakuma, J. (2011). Fairness-aware learning through regularization approach. In *ICDM Workshops*, pages 643–650. IEEE Computer Society.

Kant, I. (1785). *Groundwork of the Metaphysics of Morals*. Cambridge University Press.

Kearns, M. J., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Dy, J. G. and Krause, A., editors,

Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of *Proceedings of Machine Learning Research*, pages 2569–2577. PMLR.

Kenthapadi, K., Lakkaraju, H., Natarajan, P., and Sameki, M. (2022). Model monitoring in practice: Lessons learned and open challenges. In Zhang, A. and Rangwala, H., editors, *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 4800–4801. ACM.

Killenbeck, A. M. and Killenbeck, M. R. (2022). Brief amici curiae in support of neither party, students for fair admissions, inc. v. president and fellows of harvard college, no. 20-1199, and students for fair admissions, inc. v. university of north carolina. *President and Fellows of Harvard College*, No. 21-707(20-1199).

Kim, B., Xu, C., and Barber, R. F. (2020). Predictive inference is free with the jackknife+-after-bootstrap. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, Conference on Neural Information Processing Systems (NeurIPS) 2020, December 6-12, 2020, virtual*.

Kirsch, A., van Amersfoort, J., and Gal, Y. (2019). Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning.

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Papadimitriou, C. H., editor, *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Knoblauch, J., Husain, H., and Diethe, T. (2020). Optimal continual learning has perfect memory and is np-hard. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 5327–5337. PMLR.

Koh, P. W. and Liang, P. (2020). Understanding black-box predictions via influence functions.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. (2021). WILDS: A benchmark of in-the-wild distribution shifts. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR.

Korsgaard, C. M. (2018). *Fellow creatures: Our obligations to the other animals*. Oxford University Press.

Kumar, S. and Srivastava, A. (2012). Bootstrap prediction intervals in non-parametric regression with applications to anomaly detection. In *Proc. 18th ACM SIGKDD Conf. Knowl. Discovery Data Mining*.

Kuppler, M., Kern, C., Bach, R. L., and Kreuter, F. (2021). Distributive justice and fairness metrics in automated decision-making: How much overlap is there? *CoRR*, abs/2105.01441.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Neural Information Processing Systems*, pages 4066–4076.

Kwon, Y., Rivas, M. A., and Zou, J. (2021). Efficient computation and analysis of distributional shapley values. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 793–801. PMLR.

Kymlicka, W. (2002). *Contemporary political philosophy: An introduction*. Oxford University Press.

Labs, C. F. (2021). Inferring concept drift without labeled data. `https://concept-drift.fastforwardlabs.com/`.

Lakshminarayanan, B. (2021). Uncertainty and out-of-distribution robustness in deep learning.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles.

Lamont, J. and Favor, C. (2017). Distributive justice. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy (Winter 2017 Edition)*. Stanford. `https://plato.stanford.edu/archives/win2017/entries/justice-distributive/`.

Lee, K., Lee, K., Lee, H., and Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, Conference on Neural Information Processing Systems (NeurIPS) 2018, December 3-8, 2018, Montréal, Canada*, pages 7167–7177.

Leslie, D., Ashurst, C., González, N. M., Griffiths, F., Jayadeva, S., Jorgensen, M., Katell, M., Krishna, S., Kwiatkowski, D., Martins, C. I., Mahomed, S., Mougan, C., Pandit, S., Richey, M., Sakshaug, J. W., Vallor, S., and Vilain, L. (2024). Frontier ai, power, and the public interest: Who benefits, who decides? *Harvard Data Science Review*.

Lipton, Z. C., Wang, Y., and Smola, A. J. (2018). Detecting and correcting for label shift with black box predictors. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm,*

Sweden, July 10-15, 2018, volume 80 of *Proceedings of Machine Learning Research*, pages 3128–3136. PMLR.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. (2020a). Learning deep kernels for non-parametric two-sample tests. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6316–6326. PMLR.

Liu, J. et al. (2012). The enterprise risk management and the risk oriented internal audit. *International Business Journal*, 10(4):287–292.

Liu, W., Wang, X., Owens, J. D., and Li, Y. (2020b). Energy-based out-of-distribution detection. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33 (Conference on Neural Information Processing Systems (NeurIPS) 2020)*.

Lopez-Paz, D. and Oquab, M. (2017). Revisiting classifier two-sample tests. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Lu, Y., Wang, Z., Zhai, R., Kolouri, S., Campbell, J., and Sycara, K. P. (2023). Predicting out-of-distribution error with confidence optimal transport. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*.

Lundberg, S. M. (2020). Explaining quantitative measures of fairness. In *Fair & Responsible AI Workshop@ CHI2020*.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2019). Explainable ai for trees: From local explanations to global understanding.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020a). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839.

Lundberg, S. M., Erion, G. G., Chen, H., DeGrave, A. J., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S. (2020b). From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, 2(1):56–67.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, Long Beach, CA, USA. Curran Associates, Inc.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760.

Macleod, C. (2020). John Stuart Mill. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2020 edition.

Malinin, A. (2019). *Uncertainty estimation in deep learning with application to spoken language assessment*. PhD thesis, University of Cambridge.

Malinin, A., Athanasopoulos, A., Barakovic, M., Cuadra, M. B., Gales, M. J., Granziera, C., Graziani, M., Kartashev, N., Kyriakopoulos, K., Lu, P.-J., et al. (2022). Shifts 2.0: Extending the dataset of real distributional shifts. *arXiv preprint arXiv:2206.15407*.

Malinin, A., Band, N., Chesnokov, G., Gal, Y., Gales, M. J., Noskov, A., Ploskonosov, A., Prokhorenkova, L., Provilkov, I., Raina, V., et al. (2021a). Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*.

Malinin, A., Band, N., Gal, Y., Gales, M. J. F., Ganshin, A., Chesnokov, G., Noskov, A., Ploskonosov, A., Prokhorenkova, L., Provilkov, I., Raina, V., Raina, V., Roginskiy, D., Shmatova, M., Tigas, P., and Yangel, B. (2021b). Shifts: A dataset of real distributional shift across multiple large-scale tasks. In Vanschoren, J. and Yeung, S., editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks 2021, December 2021, virtual*.

Malinin, A., Band, N., Gal, Y., Gales, M. J. F., Ganshin, A., Chesnokov, G., Noskov, A., Ploskonosov, A., Prokhorenkova, L., Provilkov, I., Raina, V., Raina, V., Roginskiy, D., Shmatova, M., Tigas, P., and Yangel, B. (2021c). Shifts: A dataset of real distributional shift across multiple large-scale tasks. ,.

Manerba, M. M. and Guidotti, R. (2021). Fairshades: Fairness auditing via explainability in abusive language detection systems. In *CogMI*, pages 34–43. IEEE.

Marx, K. and Engels, F. (2019). The communist manifesto. In *Ideals and ideologies*, pages 243–255. Routledge.

Masashi, S. and Klaus-Robert, M. (2005). Input-dependent estimation of generalization error under covariate shift. *Statistics & Risk Modeling*, 23(4/2005):249–279.

Mase, M., Owen, A. B., and Seiler, B. (2019). Explaining black box decisions by shapley cohort refinement. *CoRR*, abs/1911.00467.

Max Roser, C. A. and Ritchie, H. (2013). Human height. *Our World in Data*. https://ourworldindata.org/human-height.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Mill, J. S. (1966). *Utilitarianism*. Springer.

Miller, A. S. and Howell, R. F. (1959). The myth of neutrality in constitutional adjudication. *University of Chicago Law Review*, 27:661.

Miller, J., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. (2021). Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7721–7735. PMLR.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163.

Mittelstadt, B. D., Russell, C., and Wachter, S. (2019). Explaining explanations in AI. In danah boyd and Morgenstern, J. H., editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 279–288. ACM.

Molnar, C. (2019). *Interpretable Machine Learning*. Online. https://christophm.github.io/interpretable-ml-book/.

Mougan, C., Alvarez, J., Ruggieri, S., and Staab, S. (2023a). Fairness implications of encoding protected categorical attributes. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 956–966, Montreal, Canada. Association for Computing Machinery.

Mougan, C. and Brand, J. (2023). Kantian deontology meets ai alignment: Towards morally robust fairness metrics.

Mougan, C., Broelemann, K., Kasneci, G., Tiropanis, T., and Staab, S. (2022). Explanation shift: Detecting distribution shifts on tabular data via the explanation space. In *Conference on Neural Information Processing Systems (NeurIPS) 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Mougan, C., Broelemann, K., Kasneci, G., Tiropanis, T., and Staab, S. (2025). Explanation shift: How did distribution shift impact the model. In *TMLR*.

Mougan, C., Gottron, T., Kanellos, G., Micheler, J., and Martínez, J. (2023b). Introducing explainable supervised machine learning into interactive feedback loops for statistical production systems. In Settlements, B. f. I., editor, *Data science in central banking: applications and tools*, volume 59. Bank for International Settlements.

Mougan, C., Kanellos, G., and Gottron, T. (2021a). Desiderata for explainable ai in statistical production systems of the european central bank. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 575–590, Cham. Springer International Publishing.

Mougan, C., Masip, D., Nin, J., and Pujol, O. (2021b). Quantile encoder: Tackling high cardinality categorical features in regression problems. In Torra, V. and Narukawa, Y., editors, *Modeling Decisions for Artificial Intelligence - 18th International Conference, MDAI 2021, Umeå, Sweden, September 27-30, 2021, Proceedings*, volume 12898 of *Lecture Notes in Computer Science*, pages 168–180. Springer.

Mougan, C. and Nielsen, D. S. (2023). Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap. In *AAAI Conference on Artificial Intelligence*.

Mougan, C., Plant, R., Teng, C., Bazzi, M., Ejea, A. C., Chan, R. S.-Y., Jasin, D. S., Stoffel, M., Whitaker, K. J., and Manser, J. (2023c). How to data in datathons. *Advances in Neural Information Processing Systems*.

Mougan, C., State, L., Ferrara, A., Ruggieri, S., and Staab, S. (2023d). Beyond demographic parity: Redefining equal treatment. In *First Workshop on AI meets Moral Philosophy and Moral Psychology. Neural Information Processing Systems*.

Munoz, C., Smith, M., and Patil, D. (2016). Big data: A report on algorithmic systems, opportunity, and civil right. *United States. Executive Office of the President*.

Mutlu, E. Ç., Yousefi, N., and Garibay, Ö. Ö. (2022). Contrastive counterfactual fairness in algorithmic decision-making. In *AIES*, pages 499–507. ACM.

Neubert, K. and Brunner, E. (2007). A studentized permutation test for the non-parametric behrens-fisher problem. *Comput. Stat. Data Anal.*, 51(10):5192–5204.

Neyshabur, B., Bhojanapalli, S., Mcallester, D., and Srebro, N. (2017). Exploring generalization in deep learning. *Advances in Neural Information Processing Systems*, 30:5947–5956.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. (2019). Towards understanding the role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations (ICLR)*.

Nigenda, D., Karnin, Z., Zafar, M. B., Ramesha, R., Tan, A., Donini, M., and Kenthapadi, K. (2022). Amazon sagemaker model monitor: A system for real-time insights into deployed machine learning models. In *KDD*, pages 3671–3681. ACM.

Nozick, R. (1974). *Anarchy, state, and utopia*, volume 5038. New York: Basic Books.

O'Neill, O. (2022). *A philosopher looks at digital communication*, volume 4. Cambridge University Press.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift.

Paleyes, A., Urma, R., and Lawrence, N. D. (2023). Challenges in deploying machine learning: A survey of case studies. *ACM Computing Surveys*, 55(6):114:1–114:29.

Papageorgiou, I. and Mougan, C. (2023). Processing sensitive personal data for bias monitoring under the ai act: Necessity principle. In *Beyond Data Protection Conference*.

Park, C., Awadalla, A., Kohno, T., and Patel, S. N. (2021a). Reliable and trustworthy machine learning for health using dataset shift detection. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, Conference on Neural Information Processing Systems (NeurIPS) 2021, December 6-14, 2021, virtual*, pages 3043–3056.

Park, C., Awadalla, A., Kohno, T., and Patel, S. N. (2021b). Reliable and trustworthy machine learning for health using dataset shift detection. *CoRR*, abs/2110.14019.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Pedreschi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In Li, Y., Liu, B., and Sarawagi, S., editors, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 560–568. ACM.

Pham, Q., Liu, C., and Hoi, S. C. H. (2021). Dualnet: Continual learning, fast and slow. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 16131–16144.

Podesta, J., Pritzker, P., Moniz, E. J., Holden, J., and Zients, J. (2014). Big data: Seizing opportunities and preserving values. *United States. Executive Office of the President*.

Prokhorenkova, L. O., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, Conference on Neural Information Processing Systems (NeurIPS) 2018, December 3-8, 2018, Montréal, Canada*, pages 6639–6649.

Quiñonero-Candela, J., Sugiyama, M., Lawrence, N. D., and Schwaighofer, A. (2009). *Dataset shift in machine learning*. Mit Press.

Rabanser, S., Günnemann, S., and Lipton, Z. C. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, Conference on Neural Information Processing Systems (NeurIPS) 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1394–1406.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Hildebrandt, M., Castillo, C., Celis, L. E., Ruggieri, S., Taylor, L., and Zanfir-Fortuna, G., editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 33–44. ACM.

Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. A. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In Bonet, B. and Koenig, S., editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 3571–3577. AAAI Press.

Rappoport, A. S. (1924). Dictionary of socialism. *(No Title)*.

Rawls, J. (1958). Justice as fairness. *The philosophical review*, 67(2):164–194.

Rawls, J. (1971). *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, Massachusetts.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., DePristo, M. A., Dillon, J. V., and Lakshminarayanan, B. (2019). Likelihood ratios for out-of-distribution detection. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32 (Conference on Neural Information Processing Systems (NeurIPS) 2019)*, pages 14680–14691.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016a). Model-agnostic interpretability of machine learning.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016b). "why should I trust you?": Explaining the predictions of any classifier.

Rosenblatt, L. and Witter, R. T. (2023). Counterfactual fairness is basically demographic parity. In *AAAI*, pages 14461–14469. AAAI Press.

Ruf, B. and Detyniecki, M. (2021). Towards the right kind of fairness in AI. *CoRR*, abs/2102.08453.

Ruggieri, S., Álvarez, J. M., Pugnana, A., State, L., and Turini, F. (2023). Can we trust fair-AI? In *AAAI*, pages 15421–15430. AAAI Press.

Sagawa, S., Koh, P. W., Lee, T., Gao, I., Xie, S. M., Shen, K., Kumar, A., Hu, W., Yasunaga, M., Marklund, H., Beery, S., David, E., Stavness, I., Guo, W., Leskovec, J., Saenko, K., Hashimoto, T., Levine, S., Finn, C., and Liang, P. (2021). Extending the WILDS benchmark for unsupervised adaptation. *CoRR*, abs/2112.05090.

Salimi, B., Rodriguez, L., Howe, B., and Suciu, D. (2019). Interventional fairness: Causal database repair for algorithmic fairness. In *SIGMOD Conference*, pages 793–810. ACM.

Schroeder, M. (2021). Value Theory. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition.

Schrouff, J., Harris, N., Koyejo, O. O., Alabdulmohsin, I., Schnider, E., Opsahl-Ong, K., Brown, A., Roy, S., Mincu, D., Chen, C., Dieng, A., Liu, Y., Natarajan, V., Karthikesalingam, A., Heller, K. A., Chiappa, S., and D'Amour, A. (2022). Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Selbst, A. D. and Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review 1085*, 87:2825–2830.

Shapley, L. S. (1953). *A Value for n-Person Games*, pages 307–318. Princeton University Press.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 3–18. IEEE Computer Society.

Simons, J., Bhatti, S. A., and Weller, A. (2021). Machine learning and the meaning of equal treatment. In Fourcade, M., Kuipers, B., Lazar, S., and Mulligan, D. K., editors, *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, pages 956–966. ACM.

Singer, P. (2011). *Practical Ethics*. Cambridge University Press, 3rd edition.

Sinnott-Armstrong, W. (2023). Consequentialism. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020a). Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In Markham, A. N., Powles, J., Walsh, T., and Washington, A. L., editors, *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, pages 180–186. ACM.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020b). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the*

AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), AIES '20, page 180–186, New York, NY, USA. Association for Computing Machinery.

Smith, L. and Gal, Y. (2018). Understanding measures of uncertainty for adversarial example detection.

Stackoverflow (2019). Developer survey results 2019.

Stevens, A., Deruyck, P., Veldhoven, Z. V., and Vanthienen, J. (2020). Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva. In *SSCI*, pages 1241–1248. IEEE.

Stolzenberg, R. M. and Relles, D. A. (1997). Tools for intuition about sample selection bias and its correction. *American Sociological Review*, 62(3):494–507.

Strumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*, 41(3):647–665.

Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005.

Sundararajan, M. and Najmi, A. (2020). The many shapley values for model explanation. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 9269–9278. PMLR.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks.

Sunstein, C. R. (1992). Neutrality in constitutional law (with special reference to pornography, abortion, and surrogacy). *Columbia Law Review*, 92:1.

Tasche, D. (2017). Fisher consistency for prior probability shift. *Journal of Machine Learning Research*, 18(95):1–32.

Vecchione, B., Levy, K., and Barocas, S. (2021). Algorithmic auditing and social justice: Lessons from the history of audit studies. In *Access in Algorithms, Mechanisms, and Optimization (EAAMO'21)*, pages 19:1–19:9. ACM.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wachter, S., Mittelstadt, B., and Russell, C. (2020). Bias preservation in machine learning: The legality of fairness metrics under eu non-discrimination law. *West Virginia Law Review*, 123:735–759.

Wachter, S., Mittelstadt, B., and Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41:105567.

Wang, H., Liu, W., Bocchieri, A., and Li, Y. (2021). Can multi-label classification networks know what they don't know? In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems 34 (Conference on Neural Information Processing Systems (NeurIPS) 2021)*, pages 29074–29087.

Watson, L. (2021). *The right to know: Epistemic rights and why we need them*. Routledge.

Yasar, A. G., Chong, A., Dong, E., Gilbert, T. K., Hladikova, S., Maio, R., Mougan, C., Shen, X., Singh, S., Stoica, A.-A., Thais, S., and Zilka, M. (2023). Ai and the eu digital markets act: Addressing the risks of bigness in generative ai. In *Proceedings of the 1st Workshop on Generative AI and Law*. International Conference on Machine Learning (ICML).

Yasar, A. G., Chong, A., Dong, E., Gilbert, T. K., Hladikova, S., Mougan, C., Shen, X., Singh, S., Stoica, A.-A., and Thais, S. (2024). Integration of generative ai in the digital markets act: Contestability and fairness from a cross-disciplinary perspective. In *Working Paper Series*. London School of Economics.

Yurochkin, M., Bower, A., and Sun, Y. (2020). Training individually fair ML models with sensitive subspace robustness. In *ICLR*. OpenReview.net.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, 2004.

Zern, A., Broelemann, K., and Kasneci, G. (2023). Interventional shap values and interaction values for piecewise linear regression trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, H., Singh, H., and Joshi, S. (2022). "why did the model fail?": Attributing model performance changes to distribution shifts. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013). Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 819–827. JMLR.org.

Zhang, L. H., Goldstein, M., and Ranganath, R. (2021). Understanding failures in out-of-distribution detection with deep generative models. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12427–12436. PMLR.