UNIVERSITY OF CALIFORNIA
SANTA CRUZ

TOWARDS SOCIALLY ACCEPTABLE ALGORITHMIC MODELS: A STUDY WITH ACTIONABLE RECOURSE

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Jayanth Yetukuri

June 2024

The Dissertation of Jayanth Yetukuri
is approved:

Dr. Yang Liu, Chair

Dr. Leilani Gilpin

Dr. Zhe Wu

Peter F. Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Towards socially acceptable algorithmic models:

A study with Actionable Recourse

by

Jayanth Yetukuri

The integration of machine learning (ML) models into our daily lives has become ubiquitous, influencing almost every aspect of our interaction with technology. However, as these models become more prevalent, particularly in sensitive areas such as healthcare, banking, and criminal justice, they must undergo rigorous scrutiny. This scrutiny addresses several critical social challenges, including data accessibility and integrity, privacy, safety, algorithmic bias, the explainability of outcomes, and transparency.

To foster trust and transparency in ML models, tools like **Actionable Recourse** (AR) have been developed. AR empowers negatively impacted users by providing recommendations for cost-efficient changes to their *actionable* features, thereby helping them achieve favorable outcomes. Traditional approaches to providing recourse focus on optimizing properties such as proximity, sparsity, validity, and distance-based costs. However, our work recognizes the importance of incorporating *User Preference* into the recourse generation process. By capturing user preferences through soft constraints—such as scoring continuous features, bounding feature values, and ranking categorical features—we propose a gradient-based approach to identify *User Preferred Actionable Recourse (UP-AR)*. Our extensive experiments validate the effectiveness of this approach.

Moreover, as ML models automate decisions in various applications, it is crucial to provide recourse that considers latent characteristics not captured in the model, such as age, sex, and marital status. We explore how the cost and feasibility of recourse vary across *latent groups*. We introduce a notion of group-level plausibility and develop a clustering procedure to identify groups with shared latent characteristics. By employing a constrained optimization approach, which we call *Fair Feasible Training (FFT)* procedure, we aim to equalize the cost of recourse over these groups. Our empirical study on simulated and real-world datasets demonstrates that our approach can produce models with improved performance in terms of cost and feasibility at the group level.

In addition to addressing group-level disparities, our study suggests a model-agnostic set of actions from a presupposed catalog called *Conformal Recourse AcTions Framework (CRAFT)*, ensuring the high probability of including the *Desired* action. This framework is adaptable to a black-box model setup and can be generalized across different models. It is intuitive, requiring only a set of calibration data points, and its effectiveness is corroborated by extensive experiments with real-world datasets.

The challenge of integrating ML models into practical applications extends to search engines, which play a pivotal role in retrieving relevant items based on user-specified queries. A significant challenge arises when there is a mismatch between the buyer's and seller's vocabularies, leading to insufficient recall or unsatisfactory results. This issue is exemplified by "Null and Low" (N&L) queries, which can significantly degrade the user experience. Our analysis of user search behavior data from a major e-commerce company revealed that approximately 29% of search queries have multiple

category interpretations, a phenomenon we term "multi-faceted query interpretations." Drawing a conceptual parallel between N&L query reformulation and counterfactual explanation literature, we propose a novel method that utilizes a neural translation model to provide diverse and multiple reformulations, thereby enhancing the user experience for N&L queries.

In conclusion, the advancement of machine learning models necessitates a multifaceted approach to ensure their ethical and equitable application. This thesis represents a necessary step in the pursuit of Trustworthy ML. By addressing the challenges of transparency, user preference, latent group disparities, and practical search engine limitations, we can move towards more responsible and user-centric machine learning systems. Additionally, it sheds light on potential avenues for future studies, underscoring the importance of continuous innovation and advancement within this vital field.

Acknowledgments

ఓం నమః శివాయ

ఓం నమో వెంకటేశాయ

I want to start by thanking GOD for every good thing in my life. I found peace and solace that were critical during this journey.

Firstly, I am honored and humbled to have worked with Professor Yang Liu, whose constant support and timely interventions proved to be extremely essential for my Ph.D. journey. Your expertise and insightful feedback have been instrumental in shaping my research and academic growth. Thank you for your patience, encouragement, and belief in my potential even when I doubted myself. I am thankful to many wonderful professors at UCSC who played a very critical role in mentoring during my initial days, including Dr. David Helmbold and Dr. Suresh Lodha. My collaboration with Dr. Berk Ustun gave me a deep understanding of the field and his hands on approach played an essential role which got one my research works accepted. I also thank Dr. Zhe Wu for being a part of my dissertation committee and always being supportive during my internship at eBay. He is hands down the best manager/person to work with.

I am thankful to Dr. Madhusudan Therani for being my mentor since 2016, whose words have always helped me find the right direction whenever I am confused about making a decision. I attribute the start and success of my Ph.D. journey to my dear friend Ganesh Sundar Kumar Govindaraju. Dr. Subramany Bharadwaj is another dear friend and mentor who has always been there for me.

Finally, I would like to express my deep love and respect for my mother Thulasi Yetukuri, who is the only reason why I am what I am today. My wife Swetha Daddala has provided me with all the essential support, trust, and confidence during my journey. My son Kartikeya Yetukuri has brought enormous luck, love and levity to our lives and I attribute all my milestones to him. I also want to thank my sister Padmaja Yetukuri for being a part of my life and for providing me with constant support.

Chapter 1

Introduction

This chapter gives a brief and sufficient introduction to all the necessary background required for the subsequent chapters of this thesis. We start with a general introduction to Artificial Intelligence and Machine learning followed by a quick look at adversarial machine learning. We then proceed to introduce Trustworthy Machine Learning and discuss the role of counterfactual and actionable recourse towards achieving the end goal of socially acceptable models.

1.1 Artificial Intelligence and Machine Learning (AIML)

Artificial Intelligence (AI) represents a frontier in computational technology. Here machines exhibit capabilities similar to human intelligence. These systems can perform several tasks, such as recognizing speech, making decisions, and solving critical problems. Machine Learning (ML), a subset of AI, enables computers to learn from historical data. Instead of having an explicitly programmed logic, ML algorithms use foundational statistical techniques to infer existing patterns and improve their functions over time. This self-improvement aspect of ML allows machines to adapt and perform tasks with increasing accuracy, revolutionizing industries from healthcare to finance by providing insights and automation that were previously neither unattainable nor scalable. Together, AI and ML are shaping a future in which intelligent machines augment human capabilities.

Building on its foundational principles, Machine Learning continues to evolve, using vast amounts of data to refine algorithms. Deep learning, a complex ML technique, mimics the neural networks of the human brain, allowing machines to process data with layers of abstraction. This has led to breakthroughs in image and speech recognition, natural language processing, and autonomous systems. As computational power grows and data become more accessible, the potential of ML expands, promising personalized medicine, predictive maintenance, and smarter cities. Ethical considerations also emerge, as ML's transformative power necessitates discussions on privacy, bias, and the future of work in an increasingly automated world.

1.1.1 Supervised machine learning

Within the last couple of decades, we have witnessed exponential growth in the field of Artificial Intelligence (AI). Specifically, some systems have achieved better than human-level performance in a variety of tasks such as speech recognition [29], image classification [46], face recognition [58] and self-driving cars [12]. These achievements were made possible by the exponential increase in machine learning techniques. Depending on the available data type, machine learning tasks can be classified into supervised learning, unsupervised learning, and semi-supervised learning. If the training data contains features explaining the problem along with labels for these features, this is called labeled data, and learning from such data is called supervised learning. Unsupervised learning handles unlabeled data that do not contain labels for the features. In this study, we do not consider unsupervised learning tasks. Based on the type of label available, supervised learning tasks can be further classified into classification problems or regression problems. We refer to classification models if the dataset contains countable discrete labels, which often have a symbolic meaning. For example, if the dataset contains images of cats and dogs, the labels would be 0 and 1 or vice versa. If the dataset contains continuous labels, we refer to regression models.

1.1.2 Adversarial machine learning

Adversarial machine learning deals with various adversarial attack techniques and defense techniques [95]. Although adversarial examples improve Deep Neural Network's (DNN) image recognition performance, [88], adversarial examples during post-training inference have been shown to affect the model's performance severely. There is a plethora of research beginning with [77] to understand the process of generating adversarial examples and defending a DNN against them. Its pros and cons accompany every strategy. For example, the most effective method to improve robustness is through *adversarial training* [25] and its variants [**?**]. Adversarial machine learning research has seen a surge in recent years. Since its dawn, *adversary* has been considered to have malicious intent. Such adversaries can be seen in spam detection [44], fraud detection [96] and in image classification [25]. In [25], authors proposed adversarial training to improve the performance of the model under attacks. Here, an adversary generates perturbed samples and augments them into the training data for training. Such an adversary within the system is working towards improving the system. An inherent assumption in this technique is the availability of an adversarial module that benefits the system. To distinguish between an external adversary trying to harm the system and an internal adversary working to improve its performance, we call the latter an *Ethical Adversary.*

1.2 Trustworthy Machine Learning

Recent years have seen significant growth in real-world applications of Machine learning models directly impacting society. Several AI regulations and policies discuss the crucial components of trustworthy machine learning [80, 42] or responsible machine learning. These challenges can be summarized as *data accessibility and integrity, privacy, safety, algorithmic bias, the explainability of outcomes, and transparency* [64]. Active research is being conducted to address each of these challenges independently. We intend to resolve these limitations by leveraging the techniques of an ethical adversary.

Trustworthy Machine Learning has become paramount as AI systems become integral to our daily lives. It emphasizes the creation of algorithms that are not only effective in terms of performance but also fair, transparent, and accountable. Ensuring trust involves addressing critical biases in the data, providing explainable AI decisions, and maintaining robustness against manipulation of inputs. Researchers and practitioners continuously strive to develop models that respect privacy and ethical standards, and improve user confidence. As regulatory frameworks evolve, they aim to guarantee that the ML applications adhere to societal norms and legal requirements. Trustworthy ML thus represents a commitment to advancing technology responsibly, prioritizing human values alongside innovation for sustainable progress in the AI domain.

In the pursuit of Trustworthy Machine Learning, interdisciplinary collaboration becomes highly essential. Experts in the fields of ethics, law, and social sciences should join forces with data scientists to design systems that respect human rights and democratic values. This holistic approach ensures that the ML tools are not only technically sound but also socially beneficial. Continuous monitoring and evaluation are key, with mechanisms for feedback and redress to address any adverse impacts swiftly. As we embed AI more deeply into the societal fabric, education and awareness are crucial, empowering users to understand and engage with ML technology. Ultimately, Trustworthy ML seeks to cultivate an ecosystem where innovation thrives without compromising the trust and well-being of its users.

1.3 Counterfactuals and Actionable recourse

1.3.1 Counterfactuals

The notion of Counterfactuals plays a critical role in the field of Trustworthy Machine Learning, particularly in the realm of explainability. They provide the hypothetical scenarios that answers the "what-if" questions, allowing users to understand how different model inputs can alter an AI model's decision. For example, in a loan approval domain, a counterfactual explanation can illustrate how a higher income or a lower debt-to-income ratio could have led to the model outcome. This not only aids model transparency but also empowers the individuals with actionable insights which can potentially help them change future model outcomes.

Counterfactuals also contribute to model debugging and fairness analysis. By analyzing how modifications in the data instances affect model predictions, developers can identify and address biases within the algorithms. They can also serve as a critical tool for regulatory compliance, helping to meet requirements that mandate explanations of the algorithmic decisions.

1.3.2 Actionable Recourse

Actionable recourse is a vital extension of counterfactual explanations for building a Trustworthy Machine Learning system, which focuses on providing users with actionable steps to alter any unfavorable decisions by the ML system. It goes beyond explaining what factors has led to a decision, offering a definitive steps for individuals to change the model outcome. For example, if a credit scoring model denies a loan application, actionable recourse will suggest specific actionable updates to an individual's features to improve the applicant's score for a future decision.

Incorporating actionable recourse requires careful consideration of the user's circumstances and the feasibility of any suggested actions. It's not enough to just propose theoretical changes; the recommendations must be realistic and achievable from user's perspective. This approach enhances user satisfaction, as it aligns with the principles individual's ability to influence model decisions.

Building models that provide actionable recourse encourages a more responsible AI design, as it prompts developers to consider the broader impact of their systems on people's lives and the overall health of the society. By ensuring that a model does not dictate outcomes without the possibility of a recourse, Trustworthy Machine Learning can help maintain social cohesion and trust in AI systems.

1.3.3 Fairness of algorithmic recourse

The concept of fairness in actionable recourse is pivotal in ensuring justice and equity within various systems, including legal, financial, and social institutions. Actionable recourse provides individuals with the means to seek redress or correction when they have been wronged or harmed. Fairness in this context implies that the mechanisms for recourse are accessible, unbiased, and transparent, allowing for consistent and impartial outcomes. It requires that all parties have the opportunity to present their case, be heard, and receive a fair evaluation based on established rules and ethical considerations. Ultimately, the fairness of actionable recourse upholds the integrity of institutions and fosters trust among the individuals they serve.

1.4 Societal aspects of Actionable Recourse

The various societal aspects of the subfield of actionable recourse in trustworthy machine learning highlight the importance of equitable access to opportunities for all individuals of a society. When an AI system provides steps for recourse, it must account for the diverse socio-economic backgrounds of the affected individuals. Such an inclusivity ensures that this procedure is not only available to the privileged few who have the resources to act on it, but can also be tailored to those who may face systemic barriers to take actions.

For example, actionable recourse in a job application screening tool should offer realistic advice to applicants from various educational and professional backgrounds. Similarly, in healthcare, personalized recommendations should consider patient's varying access to medical facilities or treatments.

Furthermore, the societal impact of actionable recourse is closely tied to the concept of justice. When AI systems are used in critical high-stakes decisions, such as criminal sentencing or welfare distribution, the ability to understand and challenge these decisions is a matter of civil rights. Actionable recourse in these contexts must be transparent, accessible, and sensitive to the complexities of human life. It should empower people to advocate for themselves and seek redress when necessary.

In a broader scope, the integration of actionable recourse into AI systems can catalyze positive social change. By motivating AI developers with this information and tools, society can push for more ethical and responsible ML systems. This accountability can lead to an uplifting of the overall health of the society. As such, actionable recourse is not just a feature of ethical AI but a cornerstone for building a more equitable society in the contemporary age of artificial intelligence.

Complex Machine learning models are making several crucial decisions with a direct impact on an average individual. For the group of individuals adversely affected by its decisions, it becomes crucial and often required to provide recourse actions that, when acted upon, can help achieve a desired outcome from the model. This recourse action is typically aimed at satisfying individual preferences without any guarantees of acceptance and is designed to work only with the model in context.

1.5 Research Questions and Contributions

In the previous subsections, we have motivated that ML models must consider individual centric approach while designing and deploying them in to the several domains of the society. However, for the sole purpose of improving the trustworthiness of the models, this thesis answers the following critical research questions.

1. How to capture individual preferences for generating recourse actions? The subfield of actionable recourse often considers human in the loop approach to capture the individual centric preferences. However, existing literature lacks an easily comprehensible techniques for gathering individual preferences.

2. Can we quantify recourse action plausibility unfairness at group level? Unfairness at group level is typically measured in terms of average recourse costs for each group.

However, we argue that the novel notion of recourse plausibility which considers the latent characteristics of the individual provides a nuanced understanding of recourse difficulty at ground level.

3. How can we enable an independent entity capable of auditing recourse actions? Existing studies are highly specific to the ML model in context and lack the provision of providing recourses with user acceptance guarantees. We explore the possibility of incorporating a framework of a universal independent ethical entity with the capability of providing recourses with data driven acceptance guarantees.
4. Can we identify other domains with critical recourse generation? The notion of actionable recourse has been identified to be applicable to several real-world domains. We also aim to extend this set and discuss its applicability to the domain of e-Commerce.

To address these critical questions, this thesis consolidates my research into multiple chapters with extensive details of the proposed solutions. Here is the list of publications contributed as part of writing this thesis:

1. **Jayanth Yetukuri** and Yang Liu. Conformal Recourse Actions framework (under review) *2024*
2. **Jayanth Yetukuri**, Yuyan Wang, Ishita Khan, Liyang Hao, Zhe Wu and Yang Liu. 2024. Multifaceted reformulations for Null & Low Queries and its Parallelism with Counterfactuals in 2024 IEEE 40th International Conference on Data Engineering (ICDE), Utrecht, Netherlands.

3. **Jayanth Yetukuri**, Ian Hardy, Vorobeychik, Y., Berk Ustun, and Yang Liu. 2024. Providing Recourse over Plausible Groups. Proceedings of the AAAI Conference on Artificial Intelligence. 38, 19 (Mar. 2024), 21753-21760.

   Doi: https://doi.org/10.1609/aaai.v38i19.30175.

4. **Jayanth Yetukuri**, Ian Hardy, and Yang Liu. 2023. Towards User Guided Actionable Recourse. In AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montréal, QC, Canada. ACM, New York, NY, USA 10 Pages.

   Doi: https://doi.org/10.1145/3600211.3604708.

5. **Jayanth Yetukuri**. 2023. Individual and Group-level considerations of Actionable Recourse. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23). Association for Computing Machinery, New York, NY, USA, 1008–1009.

   Doi: https://doi.org/10.1145/3600211.3604758.

6. Ian Hardy, **Jayanth Yetukuri**, and Yang Liu. 2023. Adaptive Adversarial Training Does Not Increase Recourse Costs. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23). Association for Computing Machinery, New York, NY, USA, 432–442.

   Doi: https://doi.org/10.1145/3600211.3604704

7. **Jayanth Yetukuri** and Yang Liu, Robust Stochastic Bandit algorithms to defend against Oracle attack using Sample Dropout, 2022 IEEE International Conference

on Big Data (Big Data), Osaka, Japan, 2022, pp. 5845-5854.

Doi: http://dx.doi.org/10.1109/BigData55660.2022.10020649.

My research integrates individual preferences into recourse generation proposes method to mitigate plausibility bias to promote socially responsible ML models.

Chapter 2

User Preferred Actionable Recourse

2.1 Introduction

Actionable Recourse (AR) [82] refers to a list of actions an individual can take to obtain a desired outcome from a fixed Machine Learning (ML) model. Several domains such as lending [75], insurance [71], resource allocation [14, 74] and hiring decisions [1] are required to suggest recourses to ensure the trust of a decision system; in such scenarios, it is critical to ensure the actionability (the viability of taking a suggested action) of recourse, otherwise the suggestions are pointless. Consider an individual named Alice who applies for a loan, and a bank, which uses an ML-based classifier, who denies it. Naturally, Alice asks - *What can I do to get the loan?* The inherent question is what action she must take to obtain the loan in the future. *Counterfactual explanation* introduced in *Wachter* [86] provides a *what-if* scenario to alter the model's decision, but it does not account for actionability. AR aims to provide Alice with a *feasible* action set which is both actionable by Alice and which suggests as low-cost modifications as possible.

While some features (such as age or sex) are inherently inactionable for all individuals, Alice's personalized constraints may also limit her ability to take action on certain suggested recourses (such as a strong reluctance to secure a co-applicant). We call these localized constraints *User Preferences*, synonymous to user-level constraints introduced as *local feasibility* by [52]. Figure 2.1 illustrates the motivation behind UP-AR. Note that how similar individuals can prefer contrasting recourse.

Actionability, as we consider it, is centered explicitly around individual preferences, and similar recourses provided to two individuals (Alice and Bob) with identical feature vectors may not necessarily be equally actionable. Most existing methods of finding actionable recourse are restricted to *omissions* of features from the *actionable feature set* and *box constraints* [56] that bound actions. In this chapter, we discuss three forms of user preferences and propose a user-provided score formulation for capturing these different idiosyncrasies. We believe that communicating in terms of preference scores (by say, providing a 1-10 rating on the actionability of specific features) improves the explainability of a recourse generation mechanism, which ultimately improves trust in the underlying model. Such a system could also be easily re-run with different preference scores, allowing for diversifiable recourse. We surveyed 40 individuals and found that an overwhelming 60% majority preferred to provide their preferences on individual features for influencing a recourse mechanism, as opposed to receiving multiple "stock" recourse options or simply receiving a single option. Additional details of our survey are included in the Appendix. We provide a hypothetical example of UP-AR's ability to adapt to preferences in Table 2.1.

Motivated by the above considerations, we capture soft user preferences along with hard constraints and identify recourse based on local desires without affecting the success rate of identifying recourse. For example, consider Alice prefers to have 80% of the recourse "cost" from loan duration and only 20% from the loan amount, meaning she prefers to have recourse with a minor reduction in the loan amount. Such recourse enables Alice to get the benefits of a loan on her terms, and can easily be calculated to Alice's desire. We study the problem of providing *user preferred recourse* by solving a custom optimization for individual user-based preferences. Our contributions include:

- We start by enabling Alice to provide three types of user preferences: i) *Scoring*, ii) *Ranking*, and iii) *Bounding*. We embed them into an optimization function to guide the recourse generation mechanism.
- We then present *User Preferred Actionable Recourse (UP-AR)* to identify a recourse tailored to her liking. Our approach highlights a cost correction step to address the *redundancy* induced by our method.
- We consolidate performance metrics with empirical results of UP-AR across multiple datasets and compare them with state-of-art techniques.

2.1.1 Related Works

Several methods exist to identify counterfactual explanations, such as FACE [63], which uses the shortest path to identify counterfactual explanations from high-density

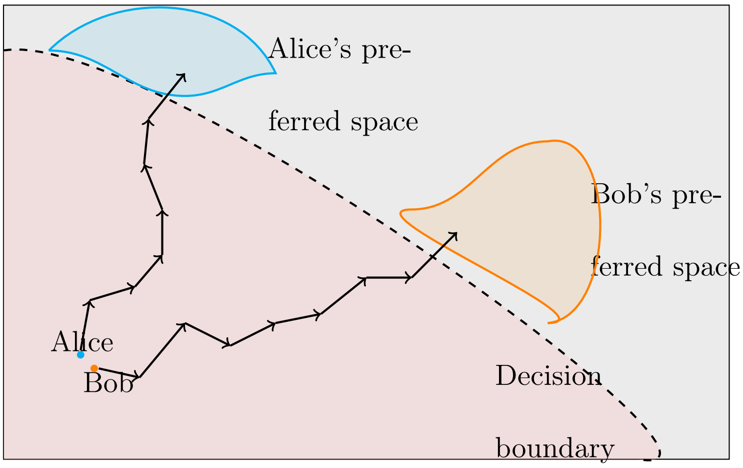


Figure 2.1: Illustration of UP-AR. Similar individuals Alice and Bob with contrasting preferences can have different regions of desired feature space for a recourse.

Actionable Features	**Curr. val.**	**UP-AR values**	
		Alice	**Bob**
`LoanDuration`	18	8	17
`LoanAmount`	$1940	$1840	$1200
`HasGuarantor`	0	0	1
`HasCoapplicant`	0	1	0

Table 2.1: A hypothetical actionable feature set of adversely affected individuals sharing similar features and corresponding suggested actions by AR and UP-AR. UP-AR provides personalized recourses.

regions, and Growing Spheres (GS) [49] which employs random sampling within increasing hyperspheres for finding counterfactuals. CLUE [7] identifies counterfactuals with low uncertainty in terms of the classifier's entropy within the data distribution. Similarly, manifold-based CCHVAE [61] generates high-density counterfactuals through the use of a latent space model. However, there is often no guarantee that the *what-if* scenarios identified by these methods are attainable.

Existing research focuses on providing feasible recourses, yet comprehensive literature on understanding and incorporating user preferences within the recourse generation mechanism is lacking. It is worth mentioning that instead of understanding user preferences, [56] provides a user with diverse recourse options and hopes that the user will benefit from at least one. The importance of diverse recourse recommendations has also been explored in recent works [86, 56, 69], which can be summarized as increasing the chances of actionability as intuitively observed in the domain of unknown user preferences [38]. [41] and [13] also resolve uncertainty in a user's cost function by inducing *diversity* in the suggested recourses. Interestingly, only 16 out of the 60 recourse methods explored in the survey by [38] include diversity as a constraint where diversity is measured in terms of distance metrics. Alternatively, studies like [82, 66, 15] optimize on a universal cost function. This does not capture individual idiosyncrasies and preferences crucial for actionability.

Efforts of eliciting user preferences include recent work by [16]. The authors provide interactive human-in-the-loop approach, where a user continuously interacts with the system. However, learning user preferences by asking them to select from one of the *partial interventions* provided is a derivative of providing a diverse set of recourse candidates. In this chapter, we consider fractional cost as a means to communicate with Alice, where fractional cost of a feature refers to *fraction of cost incurred from a feature i out of the total cost of the required intervention.*

The notion of user preference or user-level constraints was previously studied as *local feasibility* [52]. Since users can not precisely quantify the cost function [66], [90] diverged from the assumption of a universal cost function and optimizes over the distribution of cost functions. We argue that the inherent problem of feasibility can be solved more accurately by capturing and understanding Alice's recourse preference and adhering to her constraints which can vary between *Hard Rules* such as unable to bring a co-applicant and *Soft Rules* such as hesitation to reduce the amount, which should not be interpreted as unwillingness. This is the first study to capture individual idiosyncrasies in the recourse generation optimization to improve feasibility.

2.2 Problem Formulation

Consider a binary classification problem where each instance represents an individual's feature vector $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \cdot, \mathbf{x}_D]$ and associated binary label $y \in \{-1, +1\}$. We are given a model $f(\mathbf{x})$ to classify $\mathbf{x}$ into either -1 or $+1$. Let $f(\mathbf{x}) = +1$ be the desirable output of $f(\mathbf{x})$ for Alice. However, Alice was assigned an undesirable label of -1 by f. We consider the problem of suggesting action $\mathbf{r} = [\mathbf{r}_1, \mathbf{r}_2, \cdot, \mathbf{r}_D]$ such that $f(\mathbf{x} + \mathbf{r}) = +1$. Since suggested recourse only requires actions to be taken on *actionable* *features* denoted by F_A, we have $\mathbf{r}_i \equiv 0 : \forall i \notin F_A$. We further split F_A into *continuous actionable features* F_{con} and *categorical actionable features* F_{cat} based on feature domain. Action $\mathbf{r}$ is obtained by solving the following optimization, where $userCost(\mathbf{r}, \mathbf{x})$ is any predefined cost function of taking an action $\mathbf{r}$ such that:

$$\min_{\mathbf{r}} \quad userCost(\mathbf{r}, \mathbf{x}) \tag{2.1}$$

$$s.t. \quad userCost(\mathbf{r}, \mathbf{x}) = \sum_{i \in F_A} userCost(\mathbf{r}_i, \mathbf{x}_i) \tag{2.2}$$

$$\text{and } f(\mathbf{x} + \mathbf{r}) = +1. \tag{2.3}$$

2.2.1 Capturing individual idiosyncrasies

A crucial step for generating recourse is identifying *local feasibility* constraints captured in terms of individual user preferences. In this chapter, we assume that every user provides their preferences in three forms. Every continuous actionable feature $i \in F_{con}$ is associated with a *preference score* Γ_i obtained from the affected individual. Additional preferences in the form of feature value bounds and ranking for preferential treatment of categorical features are also requested from the user.

User Preference Type I (Scoring continuous features): User preference for continuous features are captured in $\Gamma_i \in [0, 1] : \forall i \in F_{con}$ subject to $\sum_{i \in F_{con}} \Gamma_i = 1$. Such *soft constraints* capture the user's preference without omitting the feature from the actionable feature set. Γ_i refers to the fractional cost of action Alice prefers to incur from a continuous feature i. For example, consider $F_{con} = \{$*LoanDuration*, *LoanAmount*$\}$ with corresponding user-provided scores $\Gamma = \{0.8, 0.2\}$ implying that Alice prefers to incur 80%

of fractional feature cost from taking action on *LoanDuration*, while only 20% of fractional cost from taking action on *LoanAmount*. Here, Alice prefers reducing *LoanDuration* to *LoanAmount* and providing recourse in accordance improves actionability.

User Preference Type II (Bounding feature values): Users can also provide constraints on values for individual features in F_A. These constraints are in the form of lower and upper bounds for individual feature values represented by $\underline{\delta i}$ and $\overline{\delta i}$ for any feature i respectively. These constraints are used to discretize the steps. For a continuous feature i, action steps can be discretized into pre-specified step sizes of $\Delta_i = \{s : s \in [\underline{\delta i}, \overline{\delta i}]\}$. For categorical features, steps are defined as the feasible values a feature can take. For all categorical features we define, $\Delta_i = \{\underline{\delta i}, \ldots, \overline{\delta i}\} : \forall i \in F_{cat}$ representing the possible values for categorical feature i.

User Preference Type III (Ranking categorical features): Users are also asked to provide a ranking function $\mathcal{R} : F_{cat} \to \mathbb{Z}^{+1}$ on F_{cat}. Let $\mathcal{R}_i$ refers to the corresponding rank for a categorical feature i. Our framework identifies recourse by updating the candidate action based on the ranking provided. For example, consider $F_{cat} = \{$*HasCoapplicant*, *HasGuarantor*, *CriticalAccountOrLoansElsewhere*$\}$ for which Alice ranks them by $\{3, 2, 1\}$. The recourse generation system considers suggesting an action on *HasGuarantor* before *HasCoapplicant*. Ranking preferences can be easily guaranteed by a simple override in case of discrepancies while finding a recourse.

2.2.1.1 Cognitive simplicity of preference scores

The user preferences proposed are highly beneficial for guiding the recourse generation process. Please note that in the absence of these preferences, the recourse procedure falls back to the default values set by a domain expert. Additionally, the users can be first presented with the default preferences, and asked to adjust as per their individual preferences. A simple user interface can help them interact with the system intuitively. For example, adjusting a feature score automatically adjusts the corresponding preference type scores.

2.2.2 Proposed optimization

We depart from capturing a user's cost of feature action and instead obtain their preferences for each feature. We elicit three forms of preferences detailed in the previous section and iteratively take steps in the action space. We propose the following optimization over the basic predefined steps based on the *user preferences.* Let us denote the inherent hardness of feature action $\mathbf{r}_i$ for feature value $\mathbf{x}_i$ using $\text{cost}(\mathbf{r} \mid \mathbf{x})$ which can be any cost function easily communicable to Alice. Here, $\text{cost}(\mathbf{r}_i^{(t)} \mid \mathbf{x}_i)$ refers to a "universal" cost of taking an action $\mathbf{r}_i^{(t)}$ for feature value $\mathbf{x}_i$ at step t. Note that this cost function or quantity differs from the $\textit{userCost}\,(\cdot, \cdot)$ function specified earlier. This

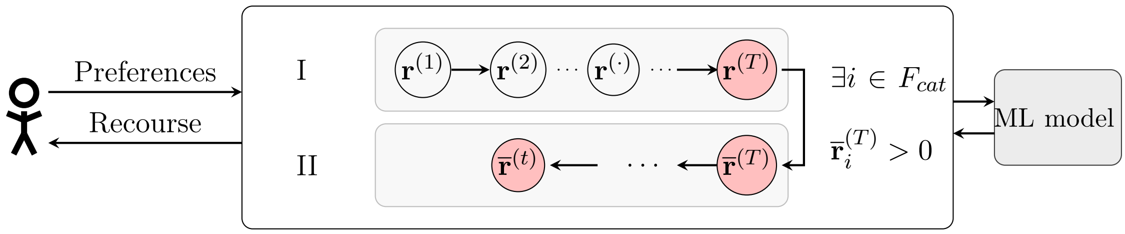


Figure 2.2: Framework of UP-AR. Successful recourse candidates; $\mathbf{r}^{(\cdot)}$, $\bar{\mathbf{r}}^{(\cdot)}$ are colored in pink.

quantity is capturing the inherent difficulty of taking an action.

$$\max_{\mathbf{r}} \sum_{i \in F_A} \frac{\Gamma_i}{\text{cost}(\mathbf{r}_i \mid \mathbf{x}_i)} \tag{Type I}$$

$$s.t.\ f(\mathbf{x}+\mathbf{r}) = +1$$

$$\Gamma_i = 0 : \ \forall i \notin F_A \tag{actionability}$$

$$\Gamma_j = 1 : \ \forall j \in F_{cat}$$

$$\mathbf{r}_i \in \Delta_i : \ i \in F_A \tag{Type II}$$

$$\mathbf{1}\{\mathbf{r}_i > 0\} \geq \mathbf{1}\{\mathbf{r}_j > 0\} : \mathcal{R}_i \geq \mathcal{R}_j \ \forall i, j \in F_{cat} \tag{Type III}$$

The proposed method minimizes the cost of a recourse weighted by Γ_i for all actionable features. We discuss the details of our considerations of cost function in Section 2.3.1. The order preference of categorical feature actions can be constrained by restrictions while finding a recourse. The next section introduces UP-AR as a stochastic solution to the proposed optimization.

2.3 User Preferred Actionable Recourse (UP-AR)

Our proposed solution, User Preferred Actionable Recourse (UP-AR), consists of two stages. The first stage generates a candidate recourse by following a connected gradient-based iterative approach. The second stage then improves upon the *redundancy* metric of the generated recourse for better actionability. The details of UP-AR are consolidated in Algorithm 1 and visualized in Figure 2.2.

2.3.1 Stage 1: Stochastic gradient-based approach

[63] identifies a counterfactual by following a high-density connected path from the feature vector $\mathbf{x}$. With a similar idea, we follow a connected path guided by the user's preference to identify a feasible recourse. We propose incrementally updating the candidate action with a predefined step size to solve the optimization. At each step t, a candidate intervention is generated, where any feature i is updated based on a Bernoulli trial with probability $I_i^{(t)}$ derived from user preference scores and the cost of taking a predefined step $\delta_i^{(t)}$ using the following procedure:

$$I_i^{(t)} \sim \mathit{Bernoulli}\left(\sigma\left(z_i^{(t)}\right)\right) \tag{2.4}$$

$$\text{where} \quad \sigma\left(z_i^{(t)}\right) = \frac{\mathrm{e}^{z_i^{(t)}/\tau}}{\sum_{j \in F_A} \mathrm{e}^{z^{(t)}/\tau}}, \; z_i^{(t)} = \frac{\Gamma_i}{\mathrm{cost}(\mathbf{r}_i^{(t)} \mid \mathbf{x}_i)} \tag{2.5}$$

With precomputed costs for each step, *weighted inverse cost* is computed for each feature, and these values are mapped to a probability distribution using a function like softmax. *Softmax* gives a probabilistic interpretation $P\left(I_i^{(t)} = 1 | z_i^{(t)}\right) = \sigma\left(z_i^{(t)}\right)$ by converting $z_i^{(t)}$ scores into probabilities.

We leverage the idea of *log percentile shift* from AR to determine the cost of action since it is easier to communicate with the users in terms of percentile shifts. Specifically, we follow the idea and formulation in [82] to define the cost:

$$\operatorname{cost}(\mathbf{r}_i \mid \mathbf{x}_i) = log \left(\frac{1 - Q_i\left(\mathbf{x}_i + \mathbf{r}_i\right)}{1 - Q_i\left(\mathbf{x}_i\right)} \right) \tag{2.6}$$

were $Q_i\left(\mathbf{x}_i\right)$ representing the *percentile* of feature i with value $\mathbf{x}_i$ is a score below which $Q_i\left(\mathbf{x}_i\right)$ percentage of scores fall in the frequency distribution of feature values in the target population.

We adapt and extend the idea that counterfactual explanations and adversarial examples [78] have a similar goal but with contrasting intention [59]. A popular approach to generating adversarial examples [25] is by using a gradient-based method. We employ the learning of adversarial example generation to determine the direction of feature modification in UP-AR: the Jacobian matrix is used to measure the local sensitivity of outputs with respect to each input feature. Consider that $f : \mathbb{R}^D \rightarrow \mathbb{R}^C$ maps a D-dimensional feature vector to a C-dimensional vector, such that each of the partial derivatives exists. For a given $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_D]$ and $f(\mathbf{x}) = [f_{[1]}(\mathbf{x}), \ldots, f_{[j]}(\mathbf{x}), \ldots, f_{[K]}(\mathbf{x})]$, the Jacobian matrix of f is defined to be a $D \times C$ matrix denoted by $\mathbf{J}$, where each (j, i) entry is $\mathbf{J}_{j,i} = \frac{\partial f_{[j]}(\mathbf{x})}{\partial \mathbf{x}_i}$. For a neural network (NN) with at least one hidden layer, $\mathbf{J}_{j,i}$ is obtained using the chain rule during backpropagation. For an NN with one hidden layer represented by *weights* $\{w\}$, we have:

$$\mathbf{J}_{j,i} = \frac{\partial f_{[j]}(\mathbf{x})}{\partial \mathbf{x}_i} = \sum_l \frac{\partial f_{[l]}(\mathbf{x})}{\partial a_l} \frac{\partial a_l}{\partial \mathbf{x}_i} \quad \text{where} \quad a_l = \sum_i w_{li} \mathbf{x}_i \tag{2.7}$$

Where in Equation 2.7, a_l is the output (with possible activation) of the hidden layer and

w_l is the weight of the node l. Notice line 4 in Algorithm 1 which *updates the candidate* action for a feature i at step t as:

$$\mathbf{r}_i^{(t)} = \mathbf{r}_i^{(t-1)} + sign\left(\mathbf{J}_{+1,i}^{(t)}\right) \cdot I_i^{(t)} \cdot \delta_i^{(t)} \tag{2.8}$$

Following the traditional notation of a binary classification problem and with a bit of abuse of notation $-1 \rightarrow 1, +1 \rightarrow +1$, $sign\left(\mathbf{J}_{+1,i}^{(t)}\right)$ captures the direction of the feature change at step t. This direction is iteratively calculated, and additional constraints such as non-increasing or non-decreasing features can be placed at this stage.

2.3.1.1 Calibrating frequency of categorical actions

We employ *temperature scaling* [26] parameter τ observed in Equation 2.5 to calibrate UP-AR's recourse generation cost. Updates on categorical features with fixed step sizes are expensive, especially for binary categorical values. Hence, tuning the frequency of categorical suggestions can significantly impact the overall cost of a recourse. τ controls the frequency with which categorical actions are suggested. Additionally, if a user prefers updates on categorical features over continuous features, UP-AR has the flexibility to address this with a smaller τ.

To study the effect of τ on overall cost, we train a Logistic Regression (LR) model on a processed version of *German* [9] dataset and generate recourses for the 155 individuals who were denied credit. The cost gradually decreases with decreasing τ since the marginal probability of suggesting a categorical feature change is diminished and the corresponding experiment is deferred to the Appendix. Hence, without affecting the success rate of recourse generation, the overall cost of generating recourses can be brought

Algorithm 1 User Preferred Actionable Recourse (UP-AR)

Input: Model f, user feature vector $\mathbf{x}$, cost function $\mathrm{cost}(\cdot \mid \cdot)$, step size $\Delta_i : \forall i \in F_A$, maximum steps T, action $\mathbf{r}$ initialized to $\mathbf{r}^{(0)}$, fixed τ, $t = 1$.

```
**while** $t \leq T$ or $f\left(\mathbf{x} + \mathbf{r}^{(t)}\right) \neq +1$ **do**
    $z_i^{(t)} = \frac{\Gamma_i}{\mathrm{cost}(\mathbf{r}_i^{(t)} \mid \mathbf{x}_i)} : \quad \forall i$
    $I_i^{(t)} \sim Bern(\sigma(z_i^{(t)})) : \quad \forall i, \text{where } \sigma(z_i^{(t)}) = \frac{\mathrm{e}^{z_i^{(t)}/\tau}}{\sum_{j \in F_A} \mathrm{e}^{z^{(t)}/\tau}}$
    $\mathbf{r}_i^{(t)} = \mathbf{r}_i^{(t-1)} + sign\left(\mathbf{J}_{+1,i}^{(t)}\right) \cdot I_i^{(t)} \cdot \delta_i^{(t)} : \ \forall i \in F_A$
    $t = t + 1$
Let $\hat{t}$ be the smallest step such that $f(\mathbf{x} + \mathbf{r}^{(\hat{t})}) = +1$ and initialize $t = \hat{t}$
**if** $\exists i \in F_{cat} : \mathbf{r}_i^{(t)} > 0$ **then**
    **while** $f\left(\mathbf{x} + \overline{\mathbf{r}}^{(t)}\right) = +1$ **do**
        $\overline{\mathbf{r}}^{(t)} = \mathbf{r}^{(t)}$
        $\overline{\mathbf{r}}_i^{(t)} = \mathbf{r}_i^{(\hat{t})} : \quad \forall i \in F_{cat}$
        $t = t - 1$
**return** $\overline{\mathbf{r}}^{(t)}$ as action $\mathbf{r}$ $=0$
```

down by decreasing τ. In simple terms, with a higher τ, UP-AR frequently suggests recourses with expensive categorical actions. We note that τ can also be informed by a user upon seeing an initial recourse. After the strategic generation of an intervention, we implement a cost correction to improve upon the potential redundancy of actions in a recourse option.

2.3.2 Stage 2: Redundancy & Cost Correction (CC)

In our experiments, we observe that once an expensive action is recommended for a categorical feature, some of the previous action steps might become redundant. Consider an LR model trained on the processed *german* dataset. Let $F_A = \{\textit{LoanDuration}, \textit{LoanAmount}, \textit{HasGuarantor}\}$ out of all the 26 features, where *HasGuarantor* is a binary feature which represents the user's ability to get a guarantor for the loan. Stage 1 takes several steps over *LoanAmount* and *LoanDuration* before recommending to update *HasGuarantor*. These steps are based on the feature action probability from Equation 2.5. Since categorical feature updates are expensive and occur with relatively low probability, Stage 1 finds a low-cost recourse by suggesting low-cost steps more frequently in comparison with high-cost steps.

Once an update to a categorical feature is recommended, some of the previous low-cost steps may be redundant, which can be rectified by tracing back previous continuous steps. Consider a scenario such that $\exists i \in F_{cat} : \mathbf{r}_i^{(T)} > 0$ for a recourse obtained after T steps in Stage 1. The CC procedure updates all the intermediary recourse candidates to reflect the categorical changes i.e., $\forall i \in F_{cat} : \mathbf{r}_i^{(T)} > 0$, we update

Features to change	Current values	Stage 1 values	Stage 2 values
LoanDuration	18	8	12
LoanAmount	$1940	$1040	$1540
HasGuarantor	0	1	1

Table 2.2: Redundancy corrected recourse for a hypothetical individual.

$\mathbf{r}_i^{(t)} = \mathbf{r}_i^{(T)} : \forall t \in \{1, 2, \ldots, T-1\}$ to obtain $\bar{\mathbf{r}}^{(t)}$. We then perform a linear retracing procedure to return $\bar{\mathbf{r}}^{(t)}$ such that $f\left(\mathbf{x} + \bar{\mathbf{r}}^{(t)}\right) = +1$ for the smallest t.

2.4 Discussion and analysis

In this section, we analyze the user preference performance of UP-AR. For simplicity, a user understands cost in terms of log percentile shift from her initial feature vector described in Section 2.3. Let $\hat{\Gamma}_i$ be the observed fractional cost for feature i formally defined in Equation 2.11. Any cost function can be plugged into UP-AR with no restrictions. A user prefers to have Γ_i fraction of the total desired percentile shift from feature i. Consider $F_A = \{\mathit{LoanDuration}, \mathit{LoanAmount}\}$ and let the corresponding user scores provided by all the adversely affected individuals be: $\Gamma = \{0.8, 0.2\}$. Here, "Denied loan applicants prefers reducing *LoanDuration* to *LoanAmount* by $8:2$." Figure 2.3 shows the frequency plot of feature cost ratio for feature *LoanDuration* out of total incurred cost

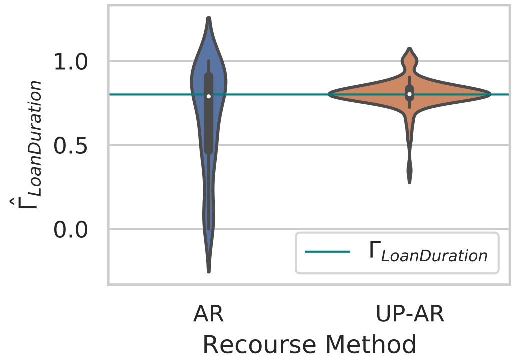


Figure 2.3: AR and UP-AR's distribution of $\hat{\Gamma}_{LoanDuration}$ for a *Logistic Regression* model trained on *German*.

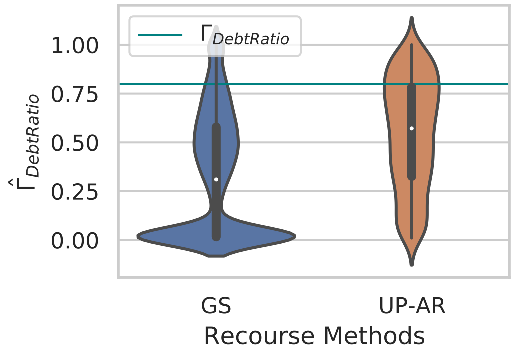


Figure 2.4: GS and UP-AR's distribution of $\hat{\Gamma}_{DebtRatio}$ for a *Neural Network* model trained on *GMSC*.

from *LoanDuration* and *LoanAmount*. i.e., $y-$axis represents $\hat{\Gamma}_i$. Also, Figure 2.4 further shows the fractional cost of feature *DebtRatio* for recourses obtained for a NN based model trained on *Give Me Some Credit (GMSC)* dataset. These experiments signify the adaptability of UP-AR to user preferences and provides evidence that distribution of $\hat{\Gamma}_i$ is centered around Γ_i.

Lemma 1. Consider UP-AR identified recourse $\mathbf{r}$ for an individual $\mathbf{x}$. If $C_{i,min}^{(T^*)}$ and $C_{i,max}^{(T^*)}$ represent the minimum and maximum cost of any step for feature i until T^*, then:

$$\mathbb{E}\left[\text{cost}(\mathbf{r}_i \mid \mathbf{x}_i)\right] \leq T^* \sigma \left(\frac{\Gamma_i}{C_{i,min}^{(T^*)}} \right) C_{i,max}^{(T^*)}. \tag{2.9}$$

Lemma 1 implies that the expected cost $\mathbb{E}\left[\text{cost}(\mathbf{r}_i \mid \mathbf{x}_i)\right]$, specifically for a continuous feature action is positively correlated to the probabilistic interpretation of user preference scores. Hence $\mathbf{r}$ satisfies users critical Type I constraints in expectation. Recall that Type II and III constraints are also applied at each step t. Lemma 1 signifies

			Alice		Bob		Chris	
Features to change	**Current values**	**AR values**	**User Pref**	**UP-AR values**	**User Pref**	**UP-AR values**	**User Pref**	**UP-AR values**
LoanDuration	30	25	0.8	20	0.8	10	0.2	27
LoanAmount	\$8072	\$5669	0.2	\$7372	0.2	\$6472	0.8	\$5272
HasGuarantor	0	1	1	1	0	0	1	1

Table 2.3: Recourses generated by UP-AR for similar individuals with a variety of preferences.

that UP-AR adheres to user preferences and thereby increases the actionability of a suggested recourse.

Corollary 1. For UP-AR with a linear $\sigma(\cdot)$, predefined steps with equal costs and $\mathrm{cost}(\mathbf{r} \mid \mathbf{x}) = \sum_{i \in F_A} \mathrm{cost}(\mathbf{r}_i \mid \mathbf{x}_i)$, total expected cost after T^* steps is:

$$\mathbb{E}\left[\mathrm{cost}(\mathbf{r} \mid \mathbf{x})\right] \leq T^* \sum_{i \in F_A} \sigma\left(\Gamma_i\right). \tag{2.10}$$

Corollary 1 states that with strategic selection of $\sigma(\cdot)$, $\delta\cdot(\cdot)$ and $\mathrm{cost}(\cdot \mid \cdot)$, UP-AR can also tune the total cost of suggested actions. In the next section, we will compare multiple recourses based on individual user preferences for a randomly selected adversely affected individual.

2.4.1 Case study of individuals with similar features

Given an LR model trained on *german* dataset and Alice, Bob and Chris be three adversely affected individuals. $F_A = \{LoanDuration, LoanAmount, HasGuarantor\}$ and corresponding user preferences are provided by the users. In Table 2.3, we consolidate the corresponding recourses generated for the specified disparate sets of preferences.

From Table 2.3 we emphasize the ability of UP-AR to generate a variety of user-preferred recourses based on their preferences, whereas AR always provides the same low-cost recourse for all the individuals. The customizability of feature actions for individual users can be found in the table. When the Type I score for *LoanAmount* is 0.8, UP-AR prefers decreasing loan amount to loan duration. Hence, the loan amount is much lesser for Chris than for Alice and Bob.

2.5 Empirical Evaluation

In this section, we demonstrate empirically: 1) that UP-AR respects Γ_i-fractional user preferences at the population level, and 2) that UP-AR also performs favorably on traditional evaluate metrics drawn from CARLA [60]. We used the native CARLA catalog for the `Give Me Some Credit (GMSC)` [36], `Adult Income (Adult)` [20] and `Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)` [6] data sets as well as pre-trained models (both the **Neural Network** (NN) and **Logistic Regression** (LR)). NN has three hidden layers of size [18, 9, 3], and the LR is a single input layer leading to a Softmax function. Although AR is proposed for *linear models*, it can be extended to *nonlinear models* by the local linear decision boundary approximation method LIME [67] (referred as AR-LIME).

PERFORMANCE METRICS: For UP-AR, we evaluate:

1. *Success Rate (Succ. Rate)*: The percentage of adversely affected individuals for whom recourse was found.

2. *Average Time Taken (Avg.Tim.)*: The average time (in seconds) to generate recourse for a single individual.

3. *Constraint Violations (Con. Vio.)*: The average number of non-actionable features modified.

4. *Redundancy (Red.)*: A metric that tracks superfluous feature changes. For each successful recourse calculated on a univariate basis, features are flipped to their original value. The redundancy for recourse is the number of flips that do not change the model's classification decision.

5. *Proximity (Pro.)*: The normalized l_2 distance of recourse to its original point.

6. *Sparsity (Spa.)*: The average number of features modified.

We provide comparative results for UP-AR against state-of-the-art counterfactual/recourse generation techniques such as GS, Wachter, AR(-LIME), CCHAVE and FACE. These methods were selected based on their popularity and their representation of both independence and dependence based methods, as defined in CARLA. In addition to the

		Neural Network							**Logistic Regression**						
Data.	**Recourse**	**Succ.**	***pRMSE***	**Avg**	**Con.**	**Red.**	**Pro.**	**Spa.**	**Succ.**	***pRMSE***	**Avg**	**Con.**	**Red.**	**Pro.**	**Spa.**
	Method	**Rate**		**Tim.**	**Vio.**				**Rate**		**Tim.**	**Vio.**			
GMSC	GS	0.75	0.16	0.02	0.00	6.95	1.01	8.89	0.62	0.18	0.03	0.00	4.08	1.39	8.99
	Wachter	1.00	0.18	0.02	1.49	6.84	1.08	8.46	1.00	0.17	0.03	1.23	3.51	1.42	7.18
	AR$_{\text{(-LIME)}}$	0.03	0.17	0.45	0.00	0.00	0.17	1.72	0.17	0.17	0.73	0.00	0.00	0.93	1.91
	CCHVAE	1.00	0.18	1.05	2.0	9.99	1.15	10.1	1.00	0.18	1.37	2.00	8.64	2.05	11.0
	FACE	1.00	0.17	8.05	1.57	6.65	1.20	6.69	1.00	0.16	11.9	1.65	7.47	2.30	8.45
	UP-AR	0.94	0.07	0.08	0.00	1.30	0.49	3.22	1.00	0.07	0.12	0.00	1.47	0.68	3.92
Adult	GS	0.84	0.10	0.03	0.00	2.86	1.30	5.09	0.84	0.10	0.04	0.00	1.76	2.05	5.85
	Wachter	0.55	0.10	0.04	1.44	3.05	0.74	4.90	1.00	0.11	0.10	1.68	0.90	1.44	5.81
	AR$_{\text{(-LIME)}}$	0.42	0.10	9.20	0.00	0.00	2.10	2.54	0.76	0.10	7.37	0.00	0.03	2.10	2.31
	CCHVAE	0.84	0.11	0.77	4.47	5.83	3.95	9.40	0.84	0.10	1.08	4.22	6.85	3.96	9.45
	FACE	1.00	0.10	6.78	4.58	7.54	4.11	7.91	1.00	0.10	8.37	4.53	5.91	4.28	7.81
	UP-AR	0.82	0.10	0.76	0.00	0.78	1.77	2.78	0.82	0.05	0.67	0.00	0.55	1.78	2.88
COMPAS	GS	1.00	0.15	0.03	0.00	1.09	0.47	3.35	1.00	0.14	0.04	0.00	0.34	1.12	3.98
	Wachter	1.00	0.14	0.05	1.00	1.61	0.56	4.35	1.00	0.14	0.04	1.00	0.85	1.06	4.83
	AR$_{\text{(-LIME)}}$	0.65	0.13	0.20	0.00	0.00	0.78	0.90	0.52	0.15	0.24	0.00	0.00	1.45	1.57
	CCHVAE	1.00	0.14	5.09	2.27	4.31	1.70	4.91	1.00	0.14	0.02	1.62	2.70	1.74	4.92
	FACE	1.00	0.15	0.37	2.39	3.96	2.35	4.72	1.00	0.15	0.40	2.47	4.38	2.46	4.81
	UP-AR	0.92	0.08	0.04	0.00	0.60	0.63	1.82	1.00	0.10	0.05	0.00	0.81	0.82	2.74

Table 2.4: Summary of performance evaluation of UP-AR. Top performers are highlighted in green.

traditional performance metrics, we also measure *Preference-Root mean squared error (pRMSE)* between the user preference score and the fractional cost of the suggested recourses. We calculate $pRMSE_i$ for a randomly selected continuous valued feature i using:

$$pRMSE_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\hat{\Gamma}_i^{(j)} - \Gamma_i^{(j)}\right)^2} \tag{2.11}$$

$$\text{where} \quad \hat{\Gamma}_i^{(j)} = \frac{\text{cost}(\mathbf{r}_i \mid \mathbf{x}_i)}{\sum_{k\in F_{con}} \text{cost}(\mathbf{r}_k \mid \mathbf{x}_k)} \tag{2.12}$$

Here $\Gamma_i^{(j)}$ and $\hat{\Gamma}_i^{(j)}$ are user provided and observed preference scores of feature i for an individual j. In Table 2.4, we summarize $pRMSE$, which is the average error across continuous features such that:

$$pRMSE = \frac{1}{|F_{con}|}\sum_{i\in F_{con}} pRMSE_i. \tag{2.13}$$

DATASETS: We train an LR model on the processed version of `german` [9] credit dataset from *sklearn's linear_model* module. We replicate [82]'s model training and recourse generation on `german`. The dataset contains 1000 data points with 26 features for a loan application. The model decides if an applicant's credit request should be approved or not. Consider $F_{con} = \{\textit{LoanDuration}, \textit{LoanAmount}\}$, and $F_{cat} = \{\textit{CriticalAccountOrLoansElsewhere}, \textit{HasGuarantor}, \textit{HasCoapplicant}\}$. Let the user scores for F_{con} be $\Gamma = \{0.8, 0.2\}$ and ranking for F_{cat} be $\{3, 1, 2\}$ for all the denied individuals. For this experiment, we set $\tau^{-1} = 4$. Out of 155 individuals with denied credit, AR and UP-AR provided recourses to 135 individuals.

Cost Correction: Out of all the denied individuals for whom categorical actions were suggested, an average of $\sim$ \$400 in *LoanAmount* was recovered by cost correction. For the following datasets, for traditional metrics, user preferences were set to be uniform for all actionable features to not bias the results to one feature preference over another:

1. **GMSC:** The data set from the 2011 Kaggle competition is a credit underwriting dataset with 11 features where the target is the presence of delinquency. Here, we measure what feature changes would lower the likelihood of delinquency. We again used the default protected features (*age* and *number of dependents*). The baseline accuracy for the NN model is 81%, while the baseline accuracy for the LR is 76%.
2. **Adult Income:** This dataset originates from 1994 census database with 14 attributes. The model decides whether an individual's income is higher than 50,000 USD/year. The baseline accuracy for the NN model is 85%, while the baseline accuracy for the LR is 83%. Our experiment is conducted on a sample of 1000 data points.
3. **COMPAS:** The data set consists of 7 features describing offenders and a target representing predictions. Here, we measure what feature changes would change an automated recidivism prediction.

The baseline accuracy for NN is 78%, while baseline accuracy for LR is 71%.

Performance analysis of UP-AR. We find UP-AR holistically performs favorably to its counterparts. Critically, it respects feature constraints (which we believe is fundamental to actionable recourse) while maintaining a significantly low redundancy and sparsity. This indicates that it tends to change fewer necessary features. Its speed makes it tractable for real-world use, while its proximity values show that it recovers relatively low-cost recourse. These results highlight the promise of UP-AR as a performative, low-cost option for calculating recourse when user preferences are paramount. UP-AR shows consistent improvements over all the performance metrics. The occasional lower success rate for a NN model is attributed to 0 constraint violations.

$pRMSE$**:** We analyze user preference performance in terms of $pRMSE$. From Table 2.4, we observe that UP-AR's $pRMSE$ is consistently better than the state of art recourse methods. The corresponding experimental details and visual representation of the distribution of $pRMSE$ is deferred to Appendix 2.5.1.

2.5.1 Random user preference study

We performed an experiment with increasing step sizes on *German* dataset. We observed that, with increasing step sizes, $pRMSE_i$ increased from 0.09 to 0.13, whereas it was consistent for AR. In the next experiment, we randomly choose user preference for *LoanDuration* from $[0.4, 0.5, 0.6, 0.7, 0.8]$. The rest of the experimental setup is identical to the setup discussed in Section 2.4. In this experiment, we observe $pRMSE$ with non-universal user preference for adversely affected individuals. Here the average $pRMSE$ of both *LoanDuration* and *LoadAmount* for UP-AR is 0.19, whereas for AR it is 0.34.

Further, using the CARLA package, we generated recourses for a set of 1000

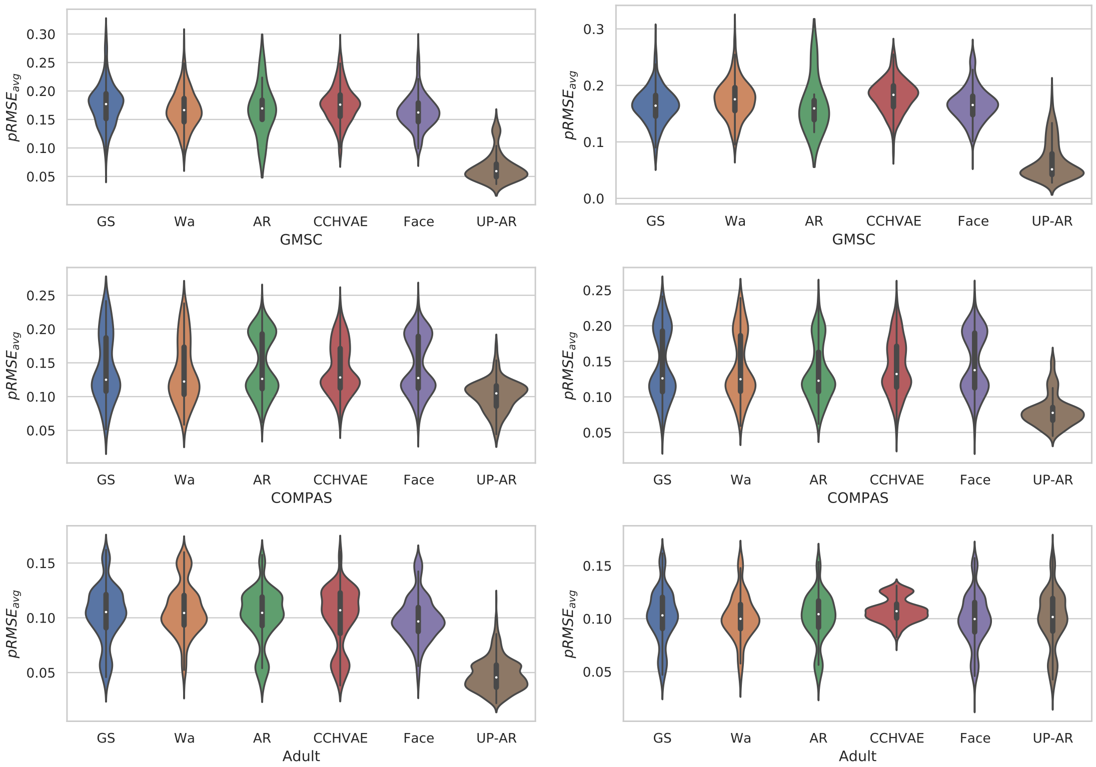


Figure 2.5: Logistic Regression model

Figure 2.6: Neural Network model

Figure 2.7: Distribution of the average $pRMSE$ of UP-AR and other recourse methodologies.

individuals and Γ for two continuous features was randomly selected from $[0.3, 0.6, 0.9]$. Figure 2.7 provides a visual analysis of the distribution of average $pRMSE$ using violin plots. The experiments were performed on the 3 datasets discussed in Section 3.4 for both the LR and NN models. For *GMSC* dataset, F_{con} = {*DebtRatio, MonthlyIncome*} and F_A = {*RevolvingUtilizationOf UnsecuredLines, NumberOfTime30-59DaysPastDueNotWorse, DebtRatio, MonthlyIncome, NumberOfOpenCreditLinesAndLoans, NumberOfTimes90DaysLate, NumberRealEstateLoansOrLines, NumberOfTime60-89DaysPastDueNotWorse*}. For *COMPAS* dataset, F_{con} = {*priors-count, length-of-stay*} and F_A = {*two-year-recid, priors-count' length-of-stay*}. For *Adult* dataset, F_{con} = {*education-num, capital-gain*} and F_A = {*education-num, capital-gain, capital-loss, hours-per-week, workclass-Non-Private, workclass-Private, marital-status-Married, marital-status-Non-Married, occupation-Managerial-Specialist, occu- pation-Other*}.

With these experiments we conclude that UP-AR's $\hat{\Gamma}$ deviation from the user's Γ is consistently lower than the existing recourse generation methodologies. We observe that AR is unaffected by the varying user preference due to the fact that AR and other state-of-the-art recourse methodologies lack the capability of capturing such idiosyncrasies. On the other hand, UP-AR is driven by those preferences and has significantly better $pRMSE$ in comparison to AR.

2.5.2 Cost Correction analysis

In Table 2.5 we explore the effect of UP-AR's cost correction procedure on the Adult and COMPAS datasets. We do not include the GMSC dataset as it does not

Metrics	**Adult**	**COMPAS**
Number of Factuals	1000	568
Success Rate	79.3%	99.6%
Percent of Recourse with a Binary Action	71.9%	82.6%
Percent of Recourse with Cost Correction	38.4%	25.5%
Average Percentage of Steps Saved	67.9%	63.5%
Average Percentage of Continuous Cost Saved	83.1%	76.0%

Table 2.5: The Frequency and Effect of Cost Correction

include binary features, and therefore does not utilize the cost correction procedure. In Table 2.5 we show the number of factuals, the percentage of factuals for which recourse was found, the percentage of recourse found which contained at least one binary action, the percent of recourse found which underwent cost correction, the average percentage of steps saved by the cost correction procedure, and the average percent of cost savings, measured as the percent reduction in continuous cost (l_2 distance) between a factual and its recourse before and after the cost-correction procedure.

2.5.3 Analysis

Interpretable and Incremental steps: In this chapter, each step $\delta_i^{(t)}$ is a predefined minimal feature modification inherently derived from the feature vector $\mathbf{x}$. A recourse suggested by UP-AR can be broken down into interpretable actions. Alice was denied a loan application, and her suggested recourse is to decrease the loan amount from \$8072 to \$6472 and decrease the loan duration from 30 years to 10 years. Here the recourse is broken down into reducing the loan amount by 16 steps of \$100 each, implying that the loan amount is 16 steps connected with the original feature value. Such steps increase the comprehensibility of recourse.

2.6 Ethics Statement

We proposed a recourse generation method for machine learning models that directly impact human lives. For practical purposes, we considered publicly available datasets for our experiments. Due care was taken not to induce any bias in this research. We further evaluated the primary performance metric for two groups (males and females) for *german* dataset.

This study reflects our efforts to bring human subjects within the framework of recourse generation. Comprehensible discussion with the users about the process improves trust and explainability of the steps taken during the entire mechanism. With machine learning models being deployed in high-impact societal applications, considering human inputs (in the form of preferences) for decision-making is a highly significant factor for improved trustworthiness. Additionally, comprehensible discussion with human subjects is another crucial component of our study. Our study motivates further research for capturing individual idiosyncrasies.

Gathering preferences from an individual could be another potential source of bias for UP-AR recourses, which needs to be evaluated with further research with human subjects. Preferential recourses will have a significant positive impact on humans conditioned on truthful reporting of various preferences. Preference scores are subject to various background factors affecting an individual, some of which can be sensitive. Additional care must be taken to provide confidentiality to these background factors while collecting individual preference scores, which have the potential to be exploited.

2.7 Ablation studies

In this section, we perform multiple experiments to understand several properties of UP-AR. First, run an experiment to measure the disparities in $pRMSE$ between the two gender groups. Secondly, we run experiments to understand the effects of the temperature parameter τ on UP-AR. Thirdly, we try to understand the relation between T^* and $\hat{\Gamma}$, if any.

2.7.1 UP-AR user preference disparities

UP-AR satisfies user Type I user preferences as observed in Section 2.4. For the following experiment, we consider a similar setup as in Section 2.4. We now evaluate similar performance among *males* and *females* separately in terms of $pRMSE$. With a similar setup as Section 2.4, Figure 2.8, shows a distribution of cost between the two gender groups. Observed $pRMSE_{LoanDuration}$ for males is 0.09, whereas for females it is 0.11. With this simple experiment, we conclude that UP-AR does not show any significant disparities in terms of adhering to user preferences.

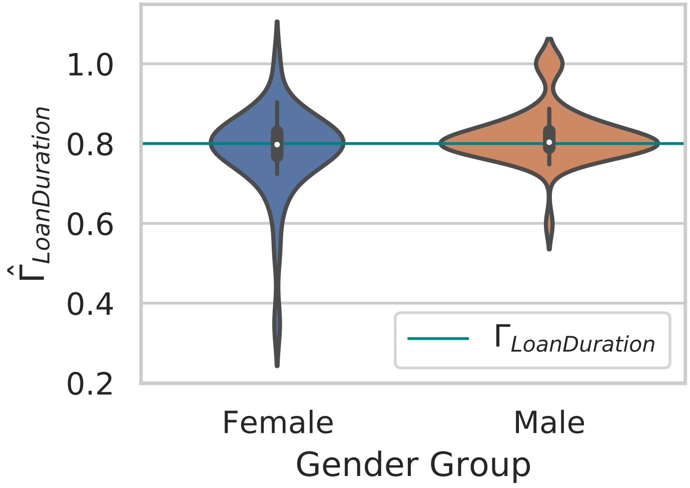


Figure 2.8: Comparison of UP-AR's distribution of $\hat{\Gamma}_{LoanDuration}$ between males and females for a *Logistic Regression* model trained on *German*.

2.7.2 Ablation study on τ

For the following experiment, we again consider a similar setup as in Section 2.4. Each data point in the plot represents the mean total cost of recourses for the target population for 20 independent runs of UP-AR, and the shaded region represents the ± 1 standard deviation of the 20 runs. We observe:

1. Effect of calibrating the overall cost of target population using τ. τ controls the frequency of categorical actions detailed in Section 2.3.1.1.

2. $\hat{\Gamma}_{LoanDuration}$ is not affected by any setting of τ as observed in Figure 2.10.

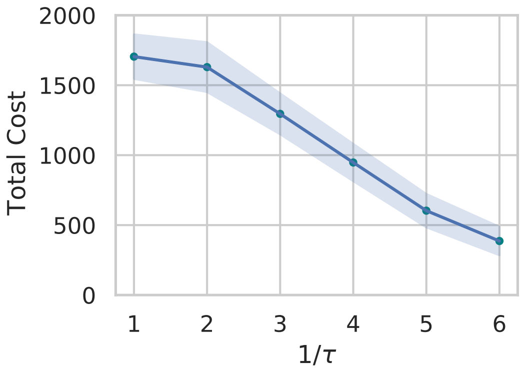


Figure 2.9: Total cost of the recourses generated for target population for varying τ. The user preference scores are fixed for the individuals.

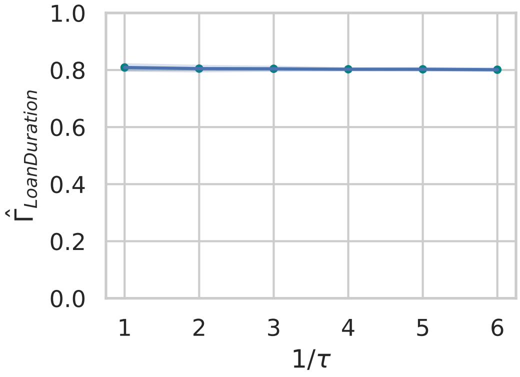


Figure 2.10: Mean fractional feature cost ratio of *LoanDuration* for varying τ. For this experiment, $\Gamma_{LoanDuration}$ is set to 0.8 for the target population.

2.7.3 Relation between $\hat{\Gamma}$ and T^*

Again considering a similar setup as in Section 2.4, Figure 2.11 visualizes the relation between the observed $\hat{\Gamma}_{LoanDuration}$ and the number of steps taken to identify a recourse T^*. We conclude that $\hat{\Gamma}_{LoanDuration}$ is not affected by the number of steps taken to identify a recourse by UP-AR.

2.7.4 Real cost vs Expected cost

In this experiment, we compare the expected cost and the actual observed cost of the recourses generated. Figure 2.12 visualizes the expected cost and observed cost for actionable features. We observe that with increasing τ, the total cost of recourses increases suggesting high categorical actions suggested in the generated recourses. Additionally, We also notice the consistency in $\hat{\Gamma}_{LoanDuration}$ for varying τ. Please note that careful calibration of τ can help individuals who prefer categorical feature actions over continuous

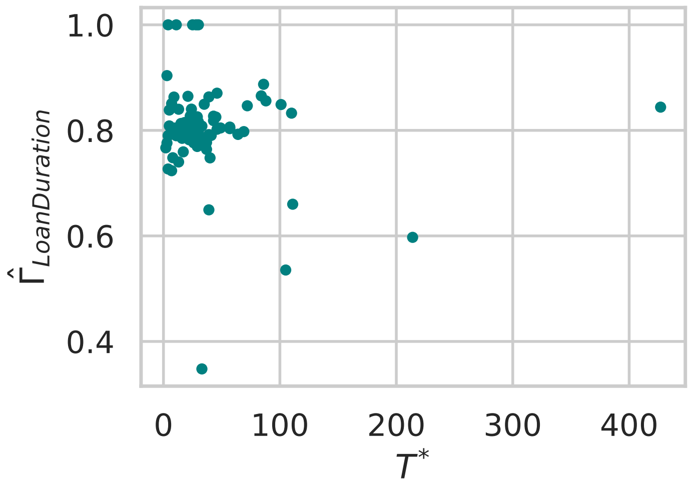


Figure 2.11: Scatter plot between $\hat{\Gamma}_{LoanDuration}$ and T^* on the recourses generated for adversely affected target population.

features.

2.7.5 Ablation study on Actionable Feature Set

We conducted an experiment on the average computational cost (modeled by execution time) of UP-AR and GS across a varying number of actionable features to explore how their performance changes as the actionable set size increases. Figures 2.13 and 2.14 show the performance trends for an *LR* model and *NN* model on the *Adult Income* dataset, while figures 2.15 and 2.16 show the performance trends for an *LR* model and *NN* model on the *German Credit* dataset. We observe that UP-AR's average time increases as the actionable feature dimension increases whereas gradient based GS remains relatively consistent. This can be attributed to the additional user scoring preference and ranking preference constraints while identifying a recourse, as well as the

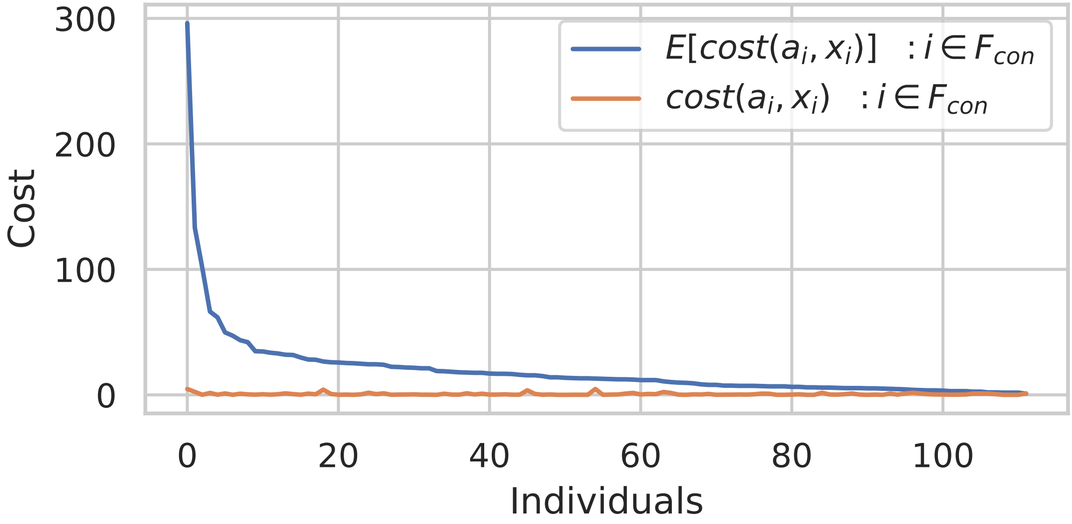
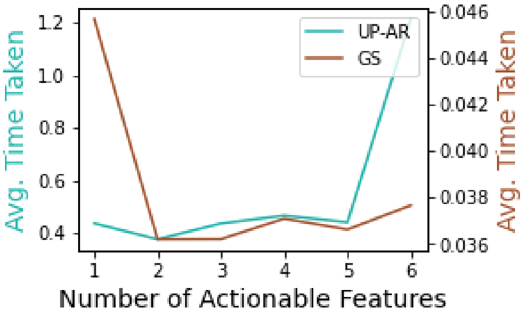


Figure 2.12: Expected and observed cost of modifications on F_{con} for all the recourses generated on the adversely affected target population.



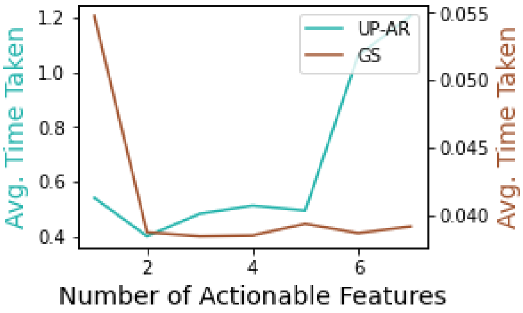


Figure 2.13: Average time to find recourse for *LR* model on the *Adult* dataset with a variable number of actionable features.

Figure 2.14: Average time to find recourse for *NN* on the *Adult* dataset with a variable number of actionable features.

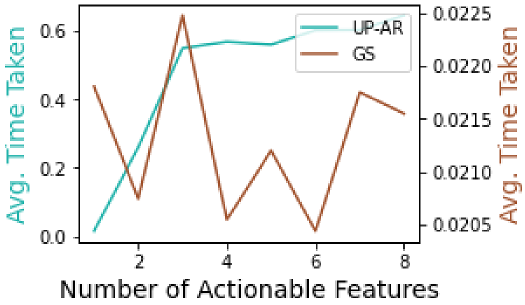


Figure 2.15: Average time to find recourse for *LR* model on the *Credit* dataset with a variable number of actionable features.

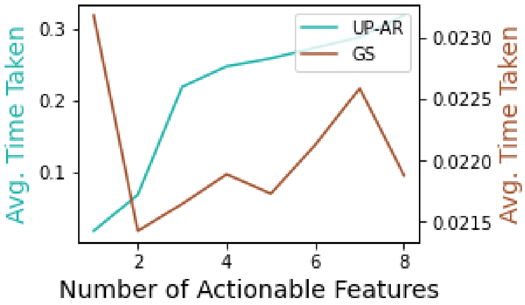


Figure 2.16: Average time to find recourse for *NN* on the *Credit* dataset with a variable number of actionable features.

cost correction procedure as the number of binary changes increases.

2.7.6 Additional proofs of results discussed in Section 2.4

2.7.6.1 Proof of Lemma 1

Consider that recourse $\mathbf{r}$ was suggested by UP-AR for Alice represented by a feature vector $\mathbf{x}$. Let $\mathbf{r}$ was obtained at time step T^*. Here $\text{cost}(\mathbf{r}_i^{(t)} \mid \mathbf{x}_i + \mathbf{r}_i^{(t-1)})$ measures the cost of taking an action $\mathbf{r}_i^{(t)}$ at time $t-1$ for feature i.

$$\begin{aligned}\mathbb{E}\left[\text{cost}(\mathbf{r}_i \mid \mathbf{x}_i)\right] &= \mathbb{E}\left[\sum_{t=1}^{T} I_i^{(t)} \cdot \text{cost}(\mathbf{r}_i^{(t)} \mid \mathbf{x}_i + \mathbf{r}_i^{(t-1)})\right] \\ &= \sum_{t=1}^{T} \mathbb{E}\left[I_i^{(t)} \cdot \text{cost}(\mathbf{r}_i^{(t)} \mid \mathbf{x}_i + \mathbf{r}_i^{(t-1)})\right] \\ &\leq \sum_{t=1}^{T} \mathbb{E}\left[I_i^{(t)} \ \cdot \ C_{i,max}^{(T^*)}\right]\end{aligned}$$

(where $C_{i,max}^{(T^*)}$ is the maximum cost of an individual feature change at any step)

$$\leq \sum_{t=1}^{T} \Pr\left(I_i^{(t)} = 1\right) C_{i,max}^{(T^*)}$$

Steps for each feature action at time t are decided by the inverse cost weighted by user preference score Γ_i. Let us call this *weighted inverse cost* which is then mapped to a probability distribution using usual choices such as normalization or a softmax function. Let $\sigma(\cdot)$ be a function which maps *weighted inverse cost* to a probability distribution. Let $C_{i,min}^{(T^*)}$ be the minimum cost of an individual feature change at any step. We have,

$$\mathbb{E}\left[\text{cost}(\mathbf{r}_i \mid \mathbf{x}_i)\right] \leq \sum_{t=1}^{T} \sigma\left(\frac{\Gamma_i}{C_{i,min}^{(T^*)}}\right) C_{i,max}^{(T^*)}$$

giving us Lemma 1.

2.7.6.2 Proof of Corollary 1

For simplicity, consider a cost function where the overall cost of recourse is the sum total of individual feature action costs, i.e., $\text{cost}(\mathbf{r} \mid \mathbf{x}) = \sum_{i \in F_A} \text{cost}(\mathbf{r}_i \mid \mathbf{x}_i)$. The total expected cost of a recourse $\mathbf{r}$ is:

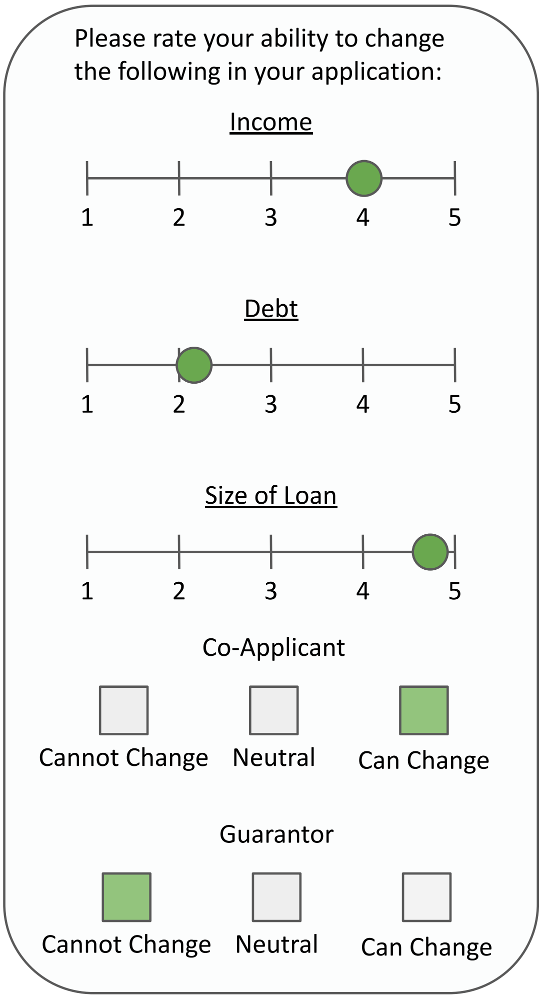
$$\mathbb{E}\left[\text{cost}(\mathbf{r} \mid \mathbf{x})\right] \leq \sum_{t \in T^*} \sum_{i \in F_A} \sigma\left(\frac{\Gamma_i}{C_{i,min}^{(T^*)}}\right) C_{i,max}^{(T^*)}$$

Considering that all the steps are of equal cost and a linear function $\sigma(\cdot)$, we get Corollary 1:

$$\mathbb{E}\left[\text{cost}(\mathbf{r} \mid \mathbf{x})\right] \leq T^* \sum_{i \in F_A} \sigma\left(\Gamma_i\right)$$

2.7.7 User interface example

Below we present an example user interface to capture various user preferences. A model could first provide the user with the default values and let the user adjust the preferences accordingly. Any user update will automatically readjust other feature default preferences. Such an interface will help reduce the cognitive burden on the end user while capturing necessary preferences.



Please rate your ability to change the following in your application:

Figure 2.17: An example interface to capture user preference.

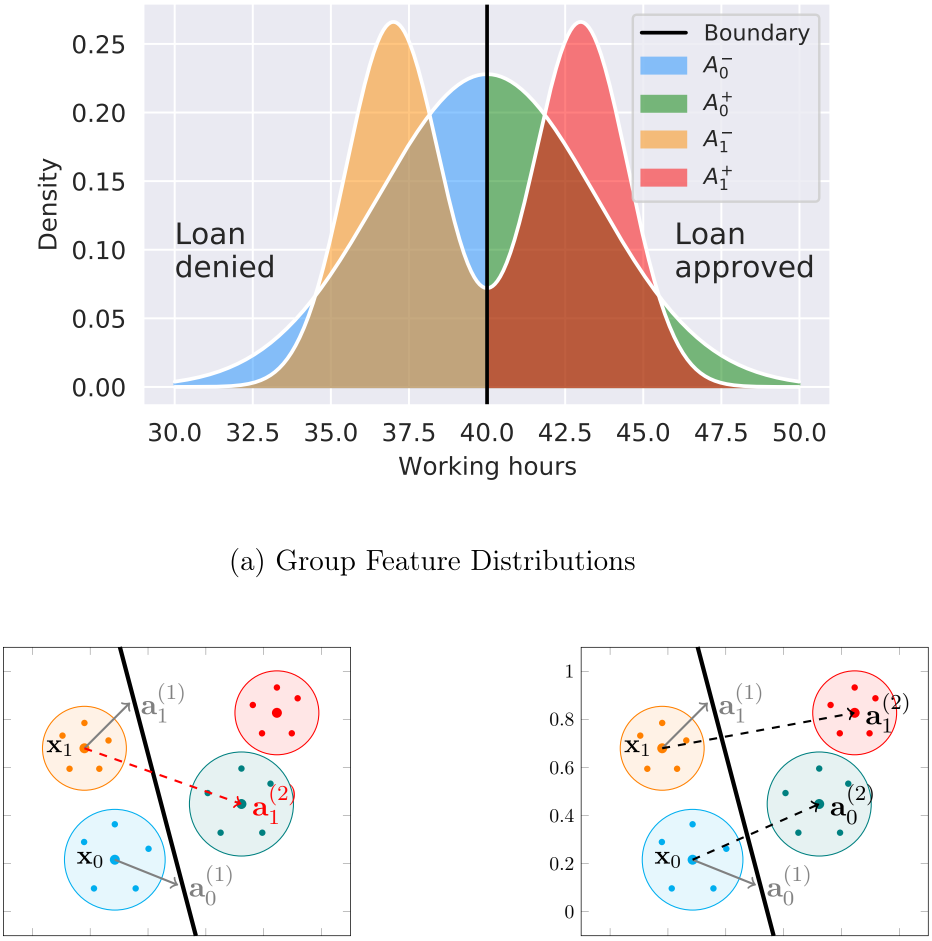
Chapter 3

Fair Recourse over Plausible Groups

3.1 Introduction

In this chapter, we explore the notion of recourse plausibility across groups and how a vanilla model training procedure can render recourses to be unfair. Existing methods for recourse provision may output actions that exhibit biases across groups in a target population. Such biases may affect the difficulty or feasibility of recourse. For example, research [22] suggests that race has a profound correlation with the level of education a person has access to. In the context of a lending model, this relationship would imply that actions that are identical may have diverging “actionability" across protected racial groups. In practice, they may arise due to historical biases within the training data [43] or due to the underlying model [17, 55].

Some existing literature seeks to address these issues through interventions at the group level. For example, [85] considers an individual's hidden feature(s) in recourse



(a) Group Feature Distributions

(b) Implausible $\mathbf{a}_1^{(2)}$

(c) Plausible $\mathbf{a}_0^{(2)}$ and $\mathbf{a}_1^{(2)}$

Figure 3.1: Toy example scenario demonstrating the existence of group-level actionability unfairness. In these figures, *orange* and *red* represent the negatively (A_1^-) and positively (A_1^+) affected sub-populations of the disadvantaged group (A_1), respectively. *Blue* and *Green* represents the negatively (A_0^-) and positively (A_0^+) affected sub-populations of the advantaged group (A_0). Consider a hypothetical situation where the average cost of recourse $\mathbf{a}_0^{(1)}$ and $\mathbf{a}_1^{(1)}$ for similar individuals $\mathbf{x}_0$ and $\mathbf{x}_1$ from A_0 and A_1, respectively, is identical. Such recourse can be commonly followed by A_0 but not necessarily by A_1.

generation, using group-level information to provide subsidies. Likewise, [51] identify *hidden confounders*, which are unobserved factors that alter the cost and feasibility of recourse at an individual level.

[27] argues that negatively impacted individuals from different groups should have equal chances of obtaining recourse, seeking to equalize the distance from the decision boundary across groups. In this chapter we consider the actionability at the group level instead of relying on a universal cost function. Consider an individual who applies for a loan and gets denied; we answer:

> "*What actions can I take to be part of the approved sub-group of people with my socioeconomic background?*"

The difference between the notion of *group-level fair actionability* and *fair recourse* is demonstrated using Figure 3.1 (a). Here, feature distribution for `working hours` follows a high variance unimodal distribution for group A_0, whereas we notice bimodal distribution for group A_1, implying that higher plausibility regime (of recourses) for group A_0 is closer to the decision boundary compared to A_1. Additionally, Figure 3.1 (b) shows the decision boundary using a scatter plot. Low density of individuals near the decision boundary for A_1, makes the recourse $\mathbf{a}_1^{(1)}$ predominantly undesirable in comparison with $\mathbf{a}_0^{(1)}$ for A_0. Alternatively, $\mathbf{a}_0^{(2)}$ and $\mathbf{a}_1^{(2)}$ from Figure 3.1 (c) shows post action features which fall within the corresponding high-density regions.

Group-level recourse plausibility of a post-action feature is defined as its believability or realizability with respect to the distribution of the group-specific approved *sub-population.* Given the spatial proximity nature [28] of plausibility, we observe that: "plausibility of post-action features is proportional to the *density* of the resulting region and *similarity* with the resulting region of approved profiles."

We leverage the *group-level approved sub-population* signals to understand actionability and thereby train a fair actionable model. Here, a group can be any immutable categorical feature in your dataset. We argue that a recourse $\mathbf{a}_0$ for an individual $\mathbf{x}_0 \in \mathcal{H}^-$ has higher chances of actionability if $\mathbf{x}_0 + \mathbf{a}_0 \in \mathcal{H}^+$, where $\mathcal{H}^+$ is the distribution of the approved group to which $\mathbf{x}_0$ belongs.

3.1.1 Motivating Scenarios

We describe two real-world scenarios for motivation for **loan approval**. Applicant A belongs to the *old* group, whereas Applicant B belongs to the *young* group, and both of them have approached a bank for a loan. Both the individuals' loan applications were denied by the bank and were suggested a similar recourse.

Applicant A: Single Parent. The recourse provided by the bank suggests increasing their working hours from 32 per week to 40 per week. Considering that they belong to the sub-population of *denied single parent*, the recourse may not be actionable, as they may not have the flexibility of increasing working hours per week. They are more likely to consider taking a second *remote job* instead. Hence, recourse actions that align with those of other single parents help improve the actionability and benefit such disadvantaged groups.

Applicant B: International Student. Applicant B is an undocumented employee with severe restrictions due to his immigration status, often limiting their flexibility in acting on the recourse provided. He may need more capabilities to act upon several features such as *income, working hours, job sector* etc. Such constraints are further exacerbated if Applicant B is a student. For the holistic benefit of society and improved trust in machine learning systems, the suggested recourses must be unbiased in terms of plausibility metrics. The main contributions of this chapter include:

1. We introduce a notion of group-level plausibility using latent characteristics related to immutable categorical features.
2. We introduce a fairness notion *group-level plausibility bias* and provide metrics for quantification using a general purpose clustering procedure.
3. We provide evidence of group-level plausibility bias using a real-world dataset dataset to show its detrimental effects on the trustworthiness of a model.
4. We consolidate the traditional performance metrics of recourse generation and compare the proposed fairness metric between naturally trained models and trained with our proposed optimization.

3.1.2 Broader Impacts

This chapter is primarily designed to mitigate specific failure modes of machine learning models used in consumer-facing applications such as lending, hiring, and the allocation of services. In particular, we seek to study how these models can assign predictions that are difficult or impossible to change across groups that are difficult to identify using features that are not used by the model. This chapter studies these biases in responsiveness through the lens of recourse and outlines a general-purpose approach to correct them. In particular, we (re)introduce plausible recourse as an alternative to a low-cost recourse.

3.2 Framework

We consider a classification task where a model $f : \mathcal{X} \to \mathcal{Y}$ assigns a binary label $\mathbf{y} \in \{\pm 1\}$ to an individual with features $\mathbf{x} = [x_1, \ldots, x_d] \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d \subseteq \mathbb{R}^d$. Let $\mathcal{D} = \left\{\left(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right)\right\}_{i=1}^{n}$ be the set of data samples observed from the true underlying distribution. Let g observing values $g \in \mathcal{G} = \{1, \ldots, K\}$ denote a categorical attribute encodes a protected characteristics.

We define the following subspaces based on the true label $\mathbf{y}$ and predicted label$f(\mathbf{x})$: $\mathcal{D}^- = \{\mathbf{x} \in \mathcal{X} : \mathbf{y} = -1\}$, $\mathcal{D}^+ = \{\mathbf{x} \in \mathcal{X} : \mathbf{y} = +1\}$, $\mathcal{H}^- = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) = -1\}$, and $\mathcal{H}^+ = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) = -1\}$. Let $v^{(i)} \in \mathcal{D}$ be a labeled example where each $v^{(i)}$ is associated with a group $g \in \mathcal{G}$. Given a *group membership function* $m : \mathbb{R}^d \to \{\pm 1\}$, we define $\mathcal{H}_g^+ = \{v^{(i)} \in \mathcal{D} |\ m\left(v^{(i)}\right) = g, f(\mathbf{x}) = +1\}$. and $\mathcal{H}_g^- = \{v^{(i)} \in \mathcal{D} |\ m\left(v^{(i)}\right) = g, f(\mathbf{x}) = -1\}$.

Recourse. Given an individual with features $\mathbf{x}_0$ such that $f\left(\mathbf{x}_0\right)=-1$, we return an action $\mathbf{a}_0$ that achieves recourse by solving an optimization problem of the form:

$$\begin{aligned} \min_{\mathbf{a}_0} \quad & \mathit{cost}\left(\mathbf{x}_0, \mathbf{a}_0\right) \\ \text{s.t.} \quad & f\left(\mathbf{x}_0+\mathbf{a}_0\right)=+1, \\ & \mathbf{a}_0 \in \mathcal{A}\left(\mathbf{x}_0\right). \end{aligned} \tag{3.1}$$

Here, $\mathit{cost}\left(\mathbf{x}_0, \mathbf{a}_0\right): \mathcal{A}\left(\mathbf{x}_0\right) \rightarrow \mathbb{R}^{+}$ is any cost function used to capture the difficulty of taking a set of actions $\mathbf{a}_0$ by an individual represented by $\mathbf{x}_0$ and let $\mathcal{A}\left(\mathbf{x}_0\right)$ be the set of feasible actions.

3.2.1 Measuring Plausibility

Recourse actions are traditionally specified by the cost of changes and actionability constraints, e.g., *feasibility sets* of [40]. Instead, we intend to maximize the overall feasibility in terms of a *proximity score* $\operatorname{prox}(\mathbf{x}_0+\mathbf{a}_0, \mathcal{S}_0)$ for an individual $\mathbf{x}_0$ with respect to a user-specified *Exemplar Set* $\mathcal{S}_0$.

An *exemplar set* contains all clusters of predefined individuals with certain robust properties, including prevalence and model agnostic adversarial robustness. Given a classifier f, a set of feasibility constraints $\mathcal{A}\left(\mathbf{x}_0\right)$, we recover an action by solving the optimization problem:

$$\begin{aligned}
\min_{\mathbf{a}_0} \quad & \textit{cost}\left(\mathbf{x}_0, \mathbf{a}_0\right) \\
\text{s.t.} \quad & \operatorname{prox}\left(\mathbf{x}_0 + \mathbf{a}_0, \mathcal{S}_0\right) \geq \rho, \\
& f\left(\mathbf{x}_0 + \mathbf{a}_0\right) = +1, \\
& \mathbf{a}_0 \in \mathcal{A}\left(\mathbf{x}_0\right).
\end{aligned} \tag{3.2}$$

Here:

- $\operatorname{prox}(\mathbf{x}_0 + \mathbf{a}_0, \mathcal{S}_0) : \mathcal{X} \to \mathbb{R}^+$ is a *proximity score* for the post-action features $\mathbf{x}_0 + \mathbf{a}_0$ to an *Exemplar Set* $\mathcal{S}_0$.
- ρ measures the minimum required proximity for $\mathbf{x}_0 + \mathbf{a}_0$ to be feasible and can vary for each group.

ρ can be specifically configured for every group based on the variance within $\mathcal{S}_0$. This ensures that $\mathbf{a}_0$ ensures underlying group characteristics. ρ ensures that $\mathbf{x}_0 + \mathbf{a}_0$ gets closer to $\mathcal{S}_0$. Configuring $\rho = 0$ returns a traditional low-cost action and $\rho > 0$ leads $\mathbf{x}_0 + \mathbf{a}_0$ to be within a specified width of $\mathcal{S}_0$, for example, an ε-ball around $\mathcal{S}_0$.

Let $\hat{\mathbf{x}}_0 = \mathbf{x}_0 + \mathbf{a}_0$ be the post action feature profile of $\mathbf{x}_0$. $\operatorname{prox}(\hat{\mathbf{x}}_0, \mathcal{S}_0)$ estimates a plausibility score by capturing the proximity of $\hat{\mathbf{x}}_0$ to the closest exemplar set $\mathcal{S}_0$. Our choice is motivated by [37]'s definition of: (i) domain-consistency; (ii) density-consistency; and (iii) prototypical-consistency.

Group Plausibility. For a the post-action feature profile $\hat{\mathbf{x}}_0$ for an individual $\mathbf{x}_0$ from a group g, we characterize *plausibility score* using the proximity nature of $\operatorname{prox}(\hat{\mathbf{x}}_0, \mathcal{S}_0)$ as $\operatorname{plaus}\left(\hat{\mathbf{x}}_0, \mathcal{S}_g\right)$ of a post-action feature profile $\hat{\mathbf{x}}_0$ with respect to any corresponding

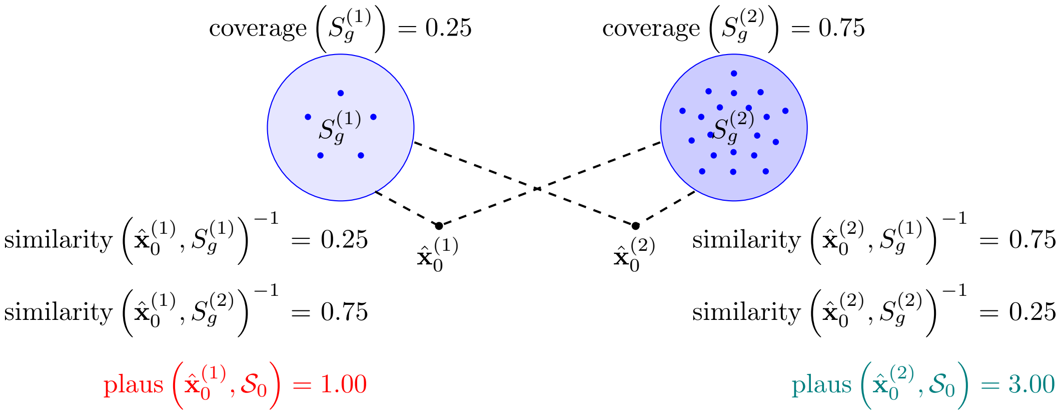

coverage $\left(S_g^{(1)}\right) = 0.25$
coverage $\left(S_g^{(2)}\right) = 0.75$
$S_g^{(1)}$
$S_g^{(2)}$
similarity $\left(\hat{\mathbf{x}}_0^{(1)}, S_g^{(1)}\right)^{-1} = 0.25$
$\hat{\mathbf{x}}_0^{(1)}$
$\hat{\mathbf{x}}_0^{(2)}$
similarity $\left(\hat{\mathbf{x}}_0^{(2)}, S_g^{(1)}\right)^{-1} = 0.75$
similarity $\left(\hat{\mathbf{x}}_0^{(1)}, S_g^{(2)}\right)^{-1} = 0.75$
similarity $\left(\hat{\mathbf{x}}_0^{(2)}, S_g^{(2)}\right)^{-1} = 0.25$
plaus $\left(\hat{\mathbf{x}}_0^{(1)}, \mathcal{S}_0\right) = 1.00$
plaus $\left(\hat{\mathbf{x}}_0^{(2)}, \mathcal{S}_0\right) = 3.00$

Figure 3.2: Demonstration of the effectiveness of plaus $(\hat{\mathbf{x}}_0, \mathcal{S}_0)$. $\hat{\mathbf{x}}_0^{(2)}$ has a high plaus $\left(\hat{\mathbf{x}}_0^{(2)}, \mathcal{S}_0\right)$ due to its high coverage $\left(S_g^{(2)}\right)$ and similarity $\left(\hat{\mathbf{x}}_0^{(2)}, S_g^{(1)}\right)^{-1}$, unlike $\hat{\mathbf{x}}_0^{(1)}$ which has a low plaus $\left(\hat{\mathbf{x}}_0^{(1)}, \mathcal{S}_0\right)$.

(approved) exemplar set $\mathcal{S}_g \in \mathcal{H}_g^+$, using:

$$\begin{aligned} &\text{plaus}\,(\hat{\mathbf{x}}_0, \mathcal{S}_g) \propto \text{density of } \mathcal{S}_g, \quad \text{and} \\ &\text{plaus}\,(\hat{\mathbf{x}}_0, \mathcal{S}_g) \propto \text{similarity with } \mathcal{S}_g \end{aligned} \tag{3.3}$$

We now define group plausibility using the patch proximity index [28] used to quantify the spatial context of a patch in relation to its neighbors. In our context, we define the proximity of $\hat{\mathbf{x}}_0$ with respect to any resulting neighbors set $\mathcal{S}_g^{(i)} \in \mathcal{S}_g$.

Definition 3.2.1 (Group Plausibility)**.** For any individual $\mathbf{x}_0$ in group $g \in \mathcal{G}$, we measure the *group-level recourse plausibility* plaus $(\hat{\mathbf{x}}_0, \mathcal{S}_g)$ of post-action features $\hat{\mathbf{x}}_0$ using:

$$\text{plaus}\,(\hat{\mathbf{x}}_0, \mathcal{S}_g) := \max\left\{\text{coverage}\left(\mathcal{S}_g^{(i)}\right) \times \text{similarity}\left(\hat{\mathbf{x}}_0, \mathcal{S}_g^{(i)}\right) : \mathcal{S}_g^{(i)} \in \mathcal{S}_g\right\} \tag{3.4}$$

where coverage $\left(\mathcal{S}_g^{(i)}\right)$ measures the fraction of data points covered by $\mathcal{S}_g^{(i)}$ and similarity $\left(\hat{\mathbf{x}}_0, \mathcal{S}_g^{(i)}\right)$ provides a score of how similar $\hat{\mathbf{x}}_0$ is with respect to $\mathcal{S}_g^{(i)}$, respectively.

We maximize the proximity score of the resulting post-action features with respect to any $\mathcal{S}_g^{(i)} \in \mathcal{S}_g$. The resulting $\hat{\mathbf{x}}_0$ must be closer to any of the exemplar profile clusters irrespective of the proximity score with other clusters.

Alternatively, *mean* based proximity score $\sum_{S_g^{(i)} \in S_g} \text{coverage}\left(S_g^{(i)}\right) \times \text{similarity}\left(\hat{\mathbf{x}}_0, S_g^{(i)}\right)$ fails in the following scenario in our formulation. Let $\text{plaus}\left(\hat{\mathbf{x}}_0, \mathcal{S}_0\right) = 2.0$ with two clusters having $\text{coverage}\left(\mathcal{S}_g^{(1)}\right) \times \text{similarity}\left(\hat{\mathbf{x}}_0, \mathcal{S}_g^{(1)}\right) = 2.0$ and $\text{coverage}\left(\mathcal{S}_g^{(2)}\right) \times \text{similarity}\left(\hat{\mathbf{x}}_0, \mathcal{S}_g^{(2)}\right) = 2.0$. Here, the resulting profile is not specifically closer to any of the exemplar sets.

3.3 Equalizing Recourse across Plausible Groups

In this section, we introduce *exemplar* set, our proposed metric to measure the plausibility of a post-action feature profile, and introduce a notion of plausibility bias. Then, we propose an optimization based model training technique to alleviate such bias caused at the group level.

3.3.1 Specifying an Exemplar Set

Action plausibility does not rely on the traditional cost of actions due to its prototypical nature [37]. This is unlike the traditional model decision boundary based low-cost actions. This provides degrees of freedom to capture individual action costs. For example, a low-density cluster signals profiles that are more likely to be outliers, which are possible to attain but *peculiar or atypical* for most individuals from that group.

The proposed plausibility metric captures the individual's group-level desirability of the actions. Identification of $\mathcal{G}$ should be done with care to ensure that it will not lead to inadvertent discrimination across protected groups.

We are motivated by the fact that an individual is more likely to enact actions that have led to approval for individuals in their exemplar group. We define groups based on the prevalence of feature values. We start by clustering the approved profiles of the group g from the training dataset into c clusters $S_g = \left\{\mathcal{S}_g^{(1)}, \ldots, \mathcal{S}_g^{(c)}\right\}$, where c is a hyperparameter selected by a domain expert. The details of the main procedure are as follows:

1. We estimate the *density* of each cluster $\mathcal{S}_g^{(i)} \in \mathcal{S}_g$ using the training dataset. We cluster approved data samples from the training dataset and associate a coverage score coverage $\left(\mathcal{S}_g^{(i)}\right)$ to each cluster. The choice of clusters must satisfy:
   1) *Positive Coverage:* coverage $\left(\mathcal{S}_g^{(i)}\right) > 0 \ \forall \ \mathcal{S}_g^{(i)} \in \mathcal{S}_g$,
   2) *Total coverage:* $\sum_{\mathcal{S}_g^{(i)} \in \mathcal{S}_g}$ coverage $\left(\mathcal{S}_g^{(i)}\right) = 1$.

2. The number of clusters c is domain dependent and can influence the average plaus $(\cdot)$ score. Please note that any choice of c should be identical across all the groups for consistency of plaus $(\cdot)$.

   For both the special cases of $c = 1$, and of $c = |\mathcal{D}_g^+|$ where $|\mathcal{D}_g^+| = |\mathcal{D}_{g'}^+| : \forall g, g' \in \mathcal{G}$, we have plaus $(\cdot) \propto$ similarity $(\cdot)$. In the former scenario, we have 1 cluster per group, and in the latter scenario, we have $|\mathcal{D}_g^+|$ clusters for every $g \in \mathcal{G}$.

3. Similarity score similarity $\left(\hat{\mathbf{x}}_0, \mathcal{S}_g^{(i)}\right)$ of the post-action feature profile $\hat{\mathbf{x}}_0$ with respect $\mathcal{S}_g^{(i)}$ can be approximated using any ℓ_p norm based distance metric. We choose ℓ_2 norm-based distance metrics to estimate the similarity score for our

experiments.

3.3.2 Measuring Plausibility Bias

Our formulation of plausibility draws on group level information, which requires a closer look at differences across groups. Existing literature focuses on equalizing recourse costs across groups [27]. However, fairness in terms of the traditional *cost* function, which is approximated using a distance metric from the factual profile, may not capture the unfairness in plausibility. To address this blind spot, we propose to capture a straightforward notion of *group-level* unfairness in plausibility. We start with a measure of the group-level plausibility-based unfairness measure for a classifier f.

Definition 3.3.1 (Expected plausibility)**.** The expected plausibility of recourse for a classifier $f : \mathcal{X} \rightarrow \{\pm 1\}$ over $\mathcal{H}^-$ is: $\overline{\text{plaus}}_{\mathcal{H}^-}(f) = \mathbb{E}\left[\mathcal{H}^-, \mathcal{D}^+\right] \text{plaus}\left(\hat{\mathbf{x}}_0, \mathcal{S}_0\right)$, where $\hat{\mathbf{x}}_0$ is the post-action feature profile resulting from solving the optimization problem in (3.2).

Definition 3.3.2 (Group plausibility bias)**.** The *group-level plausibility unfairness* of a classifier f for a dataset $\mathcal{D}$ is measured as: $\Delta_{\mathcal{P}} := \max_{g,g' \in \mathcal{G}} \left| \overline{\text{plaus}}_{\mathcal{H}_g^-}(f) - \overline{\text{plaus}}_{\mathcal{H}_{g'}^-}(f) \right|$.

where $\overline{\text{plaus}}_{\mathcal{H}_g^-}(f)$ is the group average of $\overline{\text{plaus}}\left(\hat{\mathbf{x}}_0, f\right) : \forall\ \mathbf{x}_0 \in \mathcal{H}_g^-$.

This chapter advocates for equalized plausibility across protected groups. We propose an optimization-based modeling procedure we call “Fair Feasible Training” (FFT) to train a model with an additional bias constraint. We now alleviate the effects of plausibility bias. [27] equalizes recourse action costs across groups, while we propose to

		Standard Training						**Proposed Training**					
Model	**Method**	**Succ. Rate**	**Avg. Tim.**	**Con. Vio.**	**Red.**	**Pro.**	**Spa.**	**Succ. Rate**	**Avg. Tim.**	**Con. Vio.**	**Red.**	**Pro.**	**Spa.**
N.N.	GS	1.00	0.03	0.00	2.63	1.14	4.97	1.00	0.03	0.00	2.10	1.24	5.05
	Wachter	1.00	0.05	2.00	3.20	1.25	6.94	1.00	0.05	2.00	1.77	1.42	6.95
	AR(-LIME)	0.51	1.72	0.00	0.00	1.34	1.60	0.76	1.94	0.00	0.00	1.31	1.50
	CCHVAE	1.00	0.11	3.73	7.82	3.11	8.64	1.00	0.28	3.74	7.80	3.13	8.64
	FACE	1.00	4.37	4.81	6.63	4.35	7.84	1.00	4.46	4.66	6.39	4.34	7.79
L.R.	GS	1.00	0.02	0.00	2.30	1.50	5.32	1.00	0.02	0.00	2.19	1.74	5.59
	Wachter	1.00	0.05	2.00	2.00	1.38	6.94	1.00	0.06	2.00	1.43	1.69	6.92
	AR	0.80	1.84	0.00	0.00	1.81	1.98	0.80	2.14	0.00	0.00	1.52	1.64
	CCHVAE	1.00	0.17	3.74	8.78	3.77	9.29	1.00	0.22	3.71	3.33	3.91	9.41
	FACE	1.00	4.29	4.72	6.57	4.42	7.87	1.00	5.83	4.69	6.11	4.49	7.71

Table 3.1: Overview of recourse actions for models trained using baseline methods and our approach on the Adult Income dataset.

Reference– Succ. Rate: Success Rate, Avg. Tim.: Average Time, Con. Vio.: Constraint Violations, Red.: Redundancy, Pro.:Proximity, Spa.: Sparsity.

train models that equalize recourse across latent groups by including $\Delta_{\mathcal{P}}$ as part of the model training procedure.

Definition 3.3.3 (Fair Feasible Training)**.** Given a dataset $\mathcal{D} = \left\{\left(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right)\right\}_{i=1}^{n}$ and $\epsilon > 0$, we train a feasibly fair classifier f by solving the following optimization problem:

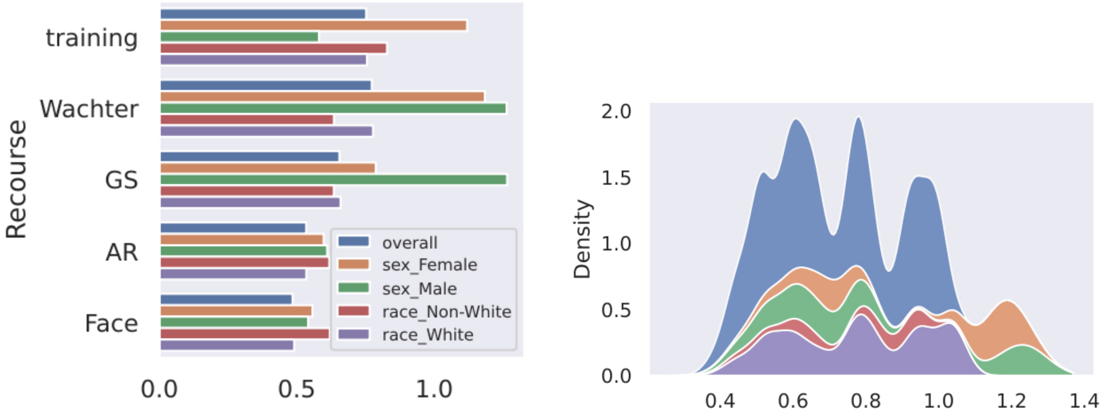
$$\begin{aligned} \min \quad & L\left(x, y\right) \\ \text{s.t.} \quad & \max_{g, g' \in \mathcal{G}} \left| \overline{\text{plaus}}_{\mathcal{H}_g^-}\left(f\right) - \overline{\text{plaus}}_{\mathcal{H}_{g'}^-}\left(f\right) \right| \leq \epsilon. \end{aligned} \tag{3.5}$$

where $L\left(x, y\right)$ is overall loss aggregated across $\mathcal{D}$ and we approximate $\overline{\text{plaus}}_{\mathcal{H}_g^-}\left(f\right)$ using $\overline{\text{plaus}}_{\mathcal{D}_g^-}\left(f\right)$ during the training process. $\overline{\text{plaus}}_{\mathcal{D}_g^-}\left(f\right)$ measures the mean distance of denied individuals of group g to their approved group counterparts, using the training dataset.

The main idea for this approximation is to equalize the spread between approved and denied sub-populations across groups during model training. With the proposed optimization, any existing recourse methodologies can be used to achieve equalized group-level plausibility across groups. An alternate approach of post-training based technique carries the risk of increased recourse costs for disadvantaged groups.

3.4 Experiments

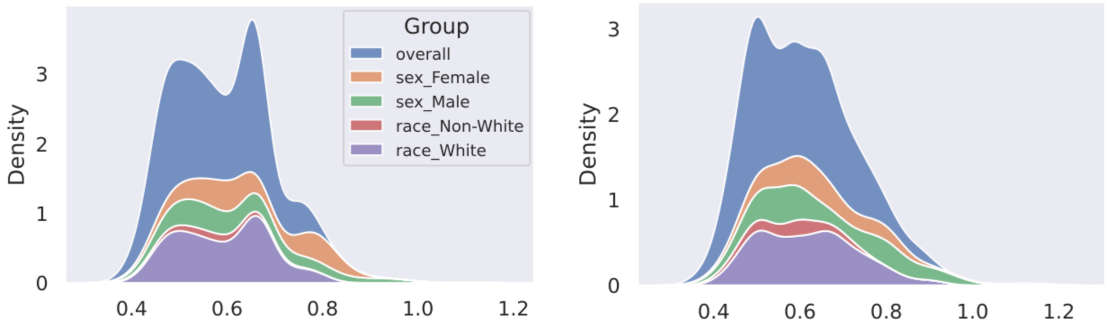
In this section, we present empirical results to show that the traditional approaches for recourse provision lead to plausibility bias and that our proposed approach (FFT) can mitigate these effects.



(a) Comparison chart of $\overline{\text{plaus}}_{(\cdot)}(f)$

(b) Stacked distribution of $\overline{\text{plaus}}_{\mathcal{D}^+_{(\cdot)}}(f)$

Figure 3.3: $\overline{\text{plaus}}_{(\cdot)}(f)$ of various recourse techniques for *gender* and *race* groups. For reference, $\overline{\text{plaus}}_{\mathcal{D}^+_{(\cdot)}}(f)$ for training data is also shown in image (a). Image (b) visualizes distributional differences of $\overline{\text{plaus}}_{\mathcal{D}^+_{(\cdot)}}(f)$ across immutable groups.



(b) Standard training

(c) Fair Feasible Training

Figure 3.4: Stacked distribution of $\overline{\text{plaus}}_{(\cdot)}(f)$ illustrates the distribution of plausibility scores across groups.

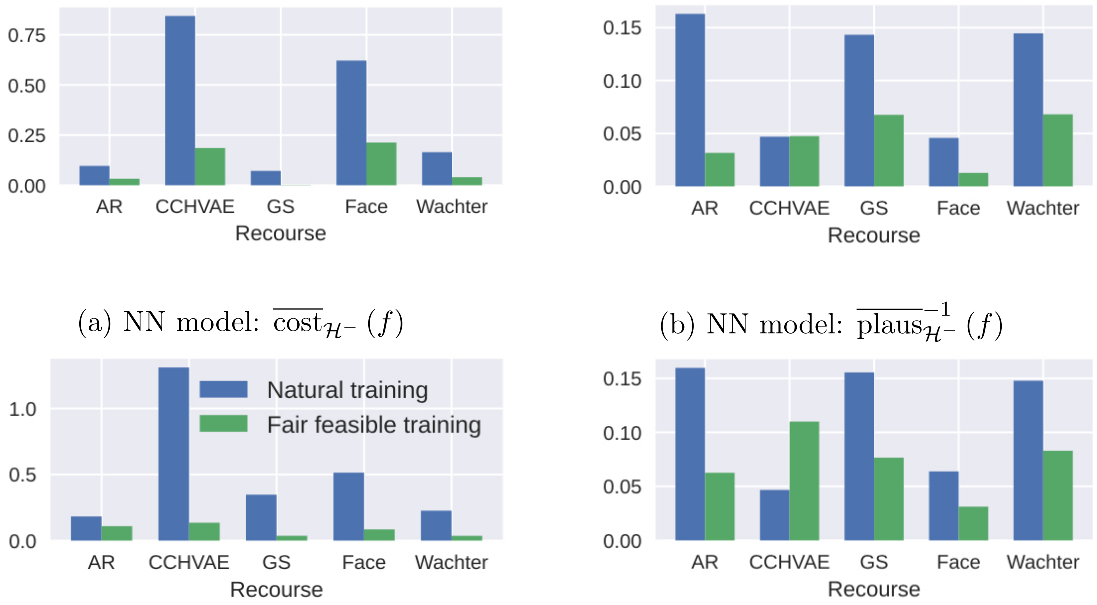
3.4.1 Setup

We train two kinds of classification models on the *Adult Income* dataset: Neural Networks (NN) and Logistic Regression (LR). For each model class, we fit a model using a baseline algorithm that optimizes cross-entropy loss and another using our proposed risk minimization in (3.3.3) utilizing the `Male` and `Female` sub-populations as the constraining groups. The NN models contain three layers of $[18, 9, 3]$ nodes with $ReLU$ activation functions, a standard drawn from the CARLA [60] recourse package. All models achieved comparable accuracy on the holdout set: the standard and constrained NN models denoted by θ_{nn}^{base}, θ_{nn}^{fair} saw 78.8% and 79.4% accuracy, respectively. While the standard and constrained LR models denoted by θ_{lr}^{base} and θ_{lr}^{fair} saw 79.2% and 78.6% accuracy, respectively. We chose *sex_ Female* as our protected group for our experiments.

Recourse methods. Although our experiments focus on one protected group, we note that the selection of groups can be parameterized to capture all the necessary groups. For all models, we then calculated a variety of recourse options on a sample of 500 adversely impacted individuals. Recourse Methods used for our experiments are: Wachter [86], Growing Spheres (GS) [49], Actionable Recourse (AR) [82], Feasible and Actionable Counterfactual Explanations (FACE) [63] and CCHVAE [62].

3.4.2 Results & Discussion

We provide evidence of several forms of plausibility bias. For instance, we identify that a particular *feature* distribution of a population categorized by strategically



(a) NN model: $\overline{\text{cost}}_{\mathcal{H}^-}(f)$

(b) NN model: $\overline{\text{plaus}}^{-1}_{\mathcal{H}^-}(f)$

(c) LR model: $\overline{\text{cost}}_{\mathcal{H}^-}(f)$

(d) LR model: $\overline{\text{plaus}}^{-1}_{\mathcal{H}^-}(f)$

Figure 3.5: Feasibility performance metrics for NN and LR models across a variety of recourse methods.

identifying protected groups shows idiosyncrasies across these groups.

On Group Level Effects Firstly, we show that feature distributions vary significantly at the immutable feature level. The distribution of `age`, `education-num` and `hours-per-week` for the *Adult Income* [21] dataset, when stratified by group shows the distributional uniqueness of individual protected groups (corresponding figures are included in the Appendix). For example, we observe twin peaks for *single woman* in `education-num`, which suggests that any recourse that lands the individual in the low-frequency region may not be actionable. The similar small second peak for *single woman* can be observed for *hours-per-week* feature.

Recourse Performance Metrics. Our results in Table 3.1 show that performance is remarkably consistent for FFT. Although FFT often incurs longer recourse generation times (seeing an average 24.6% increase in run time across recourse methods), it consistently identifies recourse that shows lower redundancy (an average 31.7% reduction). This is somewhat surprising; although we hypothesize that FFT learns more separable data representations, which may impact the ultimate redundancy of generated recourse. We observe that overall proximity costs are not significantly affected by FFT. Rather, FFT constrains the ultimate recourse to be feasibly fair to protected groups. Although recourse proximity fairness is not explicitly included in the cost function, we suspect the ultimate gains in proximity fairness result from learning a max-margin classifier on underlying fair representations.

Standard Training Exhibits Plausibility Bias. Figure 3.3 (a) shows $\overline{\text{plaus}}_{(\cdot)}(f)$ for the recourse actions generated by Wachter, GS, AR, FACE. Figure 3.3 (b) further shows the distributional differences of $\overline{\text{plaus}}_{(\cdot)}(f)$ at an individual level for the raw dataset.

FFT moderates Plausibility Bias. We compare the $\text{plaus}(\cdot)$ distributional differences across individuals based on their prediction, group, true label, and model. We observe from Figure 3.4 that the proposed training induces a consistent uni-modal $\text{plaus}(\cdot)$ distribution across groups, while standard training results in bimodal feasibility scores where female individuals in particular, see higher feasibility costs. To assess the fairness performance of FFT, we compare:

- *Expected recourse cost of a classifier [82],* $\overline{\text{cost}}_{\mathcal{H}^-}(f)$: measured as the average ℓ_2 distance $\hat{\mathbf{x}}_0$ and $\mathbf{x}_0$, of the protected groups used to constrain the training process.
- *Expected Plausibility of a classifier (Definition 3.3.1),* $\overline{\text{plaus}}_{\mathcal{H}^-}(f)$: measured as the inverse average ℓ_2 distance of $\hat{\mathbf{x}}_0$ and corresponding exemplar set $\mathcal{S}_0$ of its associated positive group.

Our findings are shown in Figure 3.5. We observe that for both model families, FFT consistently provides recourse that is fairer in terms of plausibility and the overall cost.

3.5 Concluding Remarks

In this chapter, we outlined a new approach to account for latent groups in applications where we wish to provide recourse. In particular, we developed machinery to identify such groups from data and studied the implicit disparity in plausibility across these groups. For example, suggesting naive and arguably famous recourse action of increasing the working hours to a *single parent* is not feasible. We proposed a method to train classifiers to mitigate these effects and demonstrated their capacity in practice.

Limitations. Group-level plausibility may not ensure individual actionability [45]. Our proposed approach may also exacerbate the cost of recourse. Our study raises the question of whether it is sufficient for a recourse to change the model's decision or whether a recourse improves the affected individual's overall group-level profile.

Related Work. This work is related to a previous study [40] where the authors referred to it as *believability or realizability* of recourse and refers to the likeness of the counterfactual profile resulting from the suggested set of actions. [37] refers plausibility as (i) domain-consistency; (ii) density-consistency; and (iii) prototypical-consistency. Providing recourse based on *manifold learning* [62] motivates us to utilize underlying group distributions for suggesting group-level data-dependent recourse that accounts for group-level actionability patterns [92]. Manifold-based CCHVAE [62] generates high-density counterfactuals using a latent space model. However, there is often no guarantee that the *what-if* scenarios identified are attainable. Another line of research [37] leverages causal knowledge [40] to identify *recourse via minimal interventions.* Taking causal knowledge is beneficial for identifying a recourse; however, the true underlying structural causal model is often unavailable [41].

Density-based *soft constraints* are essential for capturing group-level feasibility signals. FACE [63] follows high-density paths to produce *feasible counterfactual explana-*

tions, establishing the necessary condition of density for a feasible recourse. However, such *feasible paths* may not exist for certain groups if the approved and denied sub-populations are significantly farther apart than other groups. Other studies that learn from the dataset's underlying structure include REVISE [34] and CRUDS [19]. However, existing literature does not consider the distributional differences across groups while suggesting a recourse leading to *plausibility bias* across groups. We differ from existing literature, which prioritizes distance to the decision boundary by evaluating the actionability of recourse with respect to the distance to $\mathcal{H}^+$.

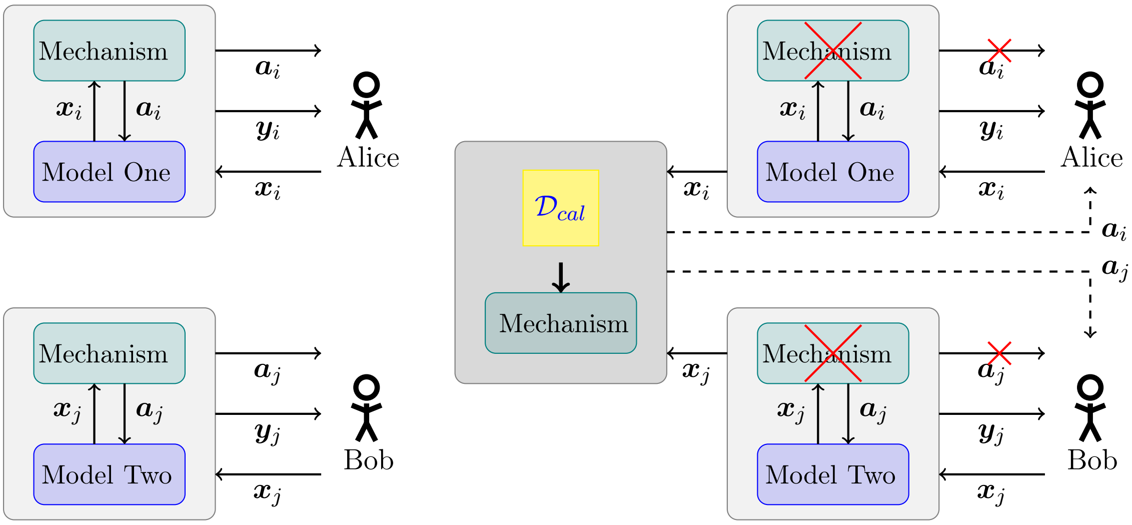
Chapter 4

Conformal Recourse Actions Framework

4.1 Introduction

In this chapter, we propose a novel framework to provide recourse actions set to an adversely affected individual with guarantees of user acceptance and additionally such recourses are model-agnostic. Decisions made by ML models have the potential to significantly impact the lives of individuals, and these decisions can often be undesired by a group of individuals, causing detrimental effects to the overall well-being of the society.

Generating a recourse is an expensive process with a variety of complexities [37], including, but not limited to, knowledge of the decision making model [35, 61], individual preferences [93] or group-level information [92], group-level recourse plausibility [94], extensive computations such as integer programming [82] for a low-cost recourse. Several critical issues are often overlooked by any state of art recourse generation mechanisms. Firstly, the pivotal dependency on the training datasets leading to *non-transferablility*



(a) Standard recourse framework

(b) Proposed recourse framework

Figure 4.1: Side by side comparison of the state-of-the-art recourse mechanisms and the proposed conformal recourse framework. Any two ML models in the same domain can have their own recourse mechanisms. We provide an essential framework to distinguish and provide a separate recourse mechanisms from ML models.

across models within the same domain. Secondly, there is an apparent inability to understand behavioral patterns in capturing the true cost of actions which is still an active research area. Third, these action selection patterns are dependent both on unpredictable macro-level economic policies and micro-level patterns of user action selection behaviors. Lastly, existing methods do not provide any guarantees of acceptance of an action by the user.

This chapter aims to address these issues with a model-independent entity which holds the responsibility of action suggestion. Specifically, with this chapter, we try to answer the following question:

- *Can we provide an audit-able model-agnostic set of recourse actions with guaranteed user acceptance?*

Here, *audit-ability* refers to the provision of an independent ethical agency that oversees the performance of the (model) suggested recourse for any domain. Note that this entity does not require any knowledge about the model parameters. The notion of an independent entity can solve training dataset associated bias for the models. Furthermore, from an anonymous survey, we identified that an overwhelming majority of approximately 67% of individuals trusted an independent entity to provide a recourse.

Although it is certainly possible to provide a set of actions that are both auditable and model-agnostic; guaranteeing user acceptance is a more challenging task that can be achieved by strategic consideration of a calibration dataset and leveraging the recent findings of conformal prediction literature in the domain of machine learning.

Conformal prediction [72, 2] provides a set of predictions that is guaranteed to contain the true prediction with high probability. In this chapter, we provide the foundational framework for the extension of conformal prediction to providing high probability recourse actions set.

This chapter discusses and addresses the foundational challenges for the implementation of conformal literature in the domain of algorithmic recourse. We focus primarily on identifying a set of actions $\mathcal{R}$ from a finite set of presupposed actions catalog A for any model specific to a domain. Our procedure deviates from existing techniques in terms of action acceptance guarantees. We provide the set $\mathcal{R}$ with high probability guarantees of user acceptance. We also include experimental results with real-world datasets, followed by an analysis and discussion section.

4.1.1 Preview of main results

The goal of this chapter is to provide a set of action recommendations to an individual adversely affected by a binary classifier model $f : \mathcal{X} \rightarrow \{-1, +1\}$ with guaranteed *probability of action acceptance*, where $\mathcal{X} \in \mathbb{R}^d$ denotes the input space. Each *action* $\boldsymbol{a}$ is a vector representing the modifications of the characteristics necessary for an individual represented by a *feature* vector $\boldsymbol{x} \in \mathcal{X}$.

For this chapter, we assume the availability of a calibration dataset $\mathcal{D}_{cal} = \{(\boldsymbol{x}_i, \boldsymbol{a}_i)\}_{i=1}^{\mathrm{n}}$ which consolidates *feature(denied)-action(approved)* pairs, where $f(\boldsymbol{x}_i) = -1$ and $\boldsymbol{a}_i$ is the action *accepted* or *desired* by the user, where $f(\boldsymbol{x}_i + \boldsymbol{a}_i) = +1$. We assume that $\boldsymbol{a}_i$ is both *valid* and *feasible*, where validity implies that f's decision was "flipped" and feasibility implies that $\boldsymbol{x}_i$ was able to perform the feature modifications suggested by $\boldsymbol{a}_i$.

In this work, we extend the conformal prediction to sets of recourse actions where emphasis is on the *desirability* of the action. Our proposed *Conformal Recourse AcTions Framework (CRAFT)* provides a set of recourse actions $\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right)$ to an adversely affected individual $\boldsymbol{x}_{\mathrm{n}+1}$. We intend to strategically design $\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right)$ such that it contains the *desired* action $\boldsymbol{a}_{\mathrm{n}+1}$ with high probability. $\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right) \subset \left\{\boldsymbol{a}^{(1)}, \ldots, \boldsymbol{a}^{(\mathrm{k})}\right\}$ is a function of n calibration data points which outputs an actions set.

CRAFT constructs actions set intended to find the *miscoverage* of an unseen feature-action combination $\left(\boldsymbol{x}_{\mathrm{n}+1}, \boldsymbol{a}_{\mathrm{n}+1}\right)$ with guarantee:

$$\operatorname{Pr}\left(\boldsymbol{a}_{\mathrm{n}+1} \notin \mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right) \mid \boldsymbol{x}_{\mathrm{n}+1}\right) \leq \alpha \tag{4.1}$$

Probability (4.1) is over the randomness of $n+1$ data points, obtained using the intuitiveness of the *conformal prediction* literature [2]. We further extend this result for a *miscoverage penalty* function $\Delta : \mathcal{R} \times \mathcal{X} \rightarrow \mathbb{R}$, which measures the penalty incurred due to not suggesting the optimal action in $\mathcal{R}$.

Formally, if $\boldsymbol{a}_{\mathrm{n}+1}$ is the desired (optimal) action for $\boldsymbol{x}_{\mathrm{n}+1}$ and $\boldsymbol{a}_{\mathrm{n}+1} \notin \mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right)$, then for any configurable lower bound z_1 and upper bound z_2 penalty for Δ, we provide guarantees for the penalty $\Delta\left(\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right), \boldsymbol{a}_{\mathrm{n}+1}\right)$ of the following form:

$$\operatorname{Pr}\left(z_1 < \Delta\left(\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right), \boldsymbol{a}_{\mathrm{n}+1}\right) < z_2\right) \leq \alpha \tag{4.2}$$

The significance of the result becomes evident from the following observation:

$$\operatorname{Pr}\left(\Delta\left(\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right), \boldsymbol{a}_{\mathrm{n}+1}\right) = 0\right) \geq 1 - \alpha \tag{4.3}$$

Remark 1. $\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right)$ takes in a feature vector $\boldsymbol{x}_{\mathrm{n}+1}$ which represents an individual who has received an unfavorable model decision, and outputs a set of actions.

Moreover, $\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right)$ is adaptive to the individual feature vector. A typical $\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right)$ becomes smaller as it becomes harder for $\boldsymbol{x}_{\mathrm{n}+1}$ to obtain an action and vice versa. We illustrate the conceptual idea using the right hand side diagram in Figure 4.1. Our contributions with this chapter include the following.

- We introduce the notion of *probability of action acceptance* $\Pr\left(\boldsymbol{a}|\boldsymbol{x}\right)$ and provide a straightforward technique for its implementation.
- We propose a general-purpose frequency-based approach to consolidate a representative action catalog $\mathcal{R}$.
- We develop a framework to provide *set of formal recourse actions* that guarantees the existence of the *desired* action with high probability.
- We provide empirical evidence for the efficacy of our approach using experiments with real-world datasets.

4.1.2 Related Works

Algorithmic Recourse. ML's proliferation into high stakes decision making domains such as banking, healthcare and recourse allocation has inspired the field of Algorithmic Recourse [82, 39], and Counterfactual Explanations [61, 87, 48].

Recourse mechanisms typically follow an optimization principle for action suggestion. For example, AR [82] follows the principle of identifying a *low-cost* feasible action and FACE [63] identifies a *high-density* feasible action, etc.

Given an individual with features $\boldsymbol{x}$ such that $f(\boldsymbol{x}) = -1$, AR returns an action $\boldsymbol{a}$ that achieves recourse by solving the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{a}} \quad & \boldsymbol{x}, \boldsymbol{a} \\ \text{s.t.} \quad & f(\boldsymbol{x} + \boldsymbol{a}) = +1, \\ & \boldsymbol{a} \in \mathcal{A}(\boldsymbol{x}) . \end{aligned} \tag{4.4}$$

Here, $\boldsymbol{x}, \boldsymbol{a} : \mathcal{A}(\boldsymbol{x}) \rightarrow \mathbb{R}^{+}$ is any cost function used to capture the difficulty of taking a set of actions $\boldsymbol{a}$ by an individual represented by $\boldsymbol{x}$ and $\mathcal{A}(\boldsymbol{x})$ is the set of feasible actions for $\boldsymbol{x}$.

The performance of such mechanisms is typically measured two fold: (i) *success rate* in terms of the fraction of adversely affected individuals who were provided with an action, and (ii) *average cost* which measures the average difficulty of the suggested actions.

Conformal prediction. The notion of *conformal prediction* introduced by conformal provides a distribution-free and statistically rigorous uncertainty quantification in algorithmic randomness. This mechanism has recently gained popularity [72, 2] due to its intuitive integration into algorithmic models.

Using the calibration data $\mathcal{D}_{cal}$ with n data points, conformal prediction constructs an uncertainty set C, that guarantees to include the true prediction with high probability:

$$\Pr\left(f(\boldsymbol{x}_{\mathrm{n}+1}) \in C \mid \boldsymbol{x}_{\mathrm{n}+1}\right) \geq 1 - \alpha \tag{4.5}$$

Recent studies have explained distribution-free reliability guarantees for a model [4] using conformal prediction, which is extended further to conformal risk control [3] for ordinal classification [89], gender-balanced conformal prediction [5] and class-conditional conformal prediction [18].

4.2 Methodology

We divide this section into four parts: the first subsection discusses the required preliminaries and background information, the second subsection specifies the procedure to quantify the heuristic notion of uncertainty within the acceptance of an action $\boldsymbol{a}$ for an individual $\boldsymbol{x}$ and the third subsection follows the conformal prediction framework to consolidate $\mathcal{R}(\boldsymbol{x})$. And finally in the fourth section, we provide theoretical guarantees of the proposed methodology.

4.2.1 Preliminaries

The two components discussed in this subsection are:

(i) Identifying a fixed *action catalog* A, and

(ii) Constructing the calibration dataset for uncertainty estimation.

Action catalog A**.** Traditionally, recourse actions are tailored specifically for an individual. For the scope of this chapter, we consider that these actions are selected from a presupposed set of meaningful actions which covers a broad range of individual profiles. These actions can vary from a minor change in a real valued feature to a major change in the categorical features.

We formulate the problem of action suggestion as a classification framework with a finite actions space A. The first challenge is to identify a fixed catalog of actions A which contains k actions, which can be generalized across various models. A domain specific standard A can be identified by a domain expert or simulated using historical data (if available).

For example, in credit lending domain, an action $\boldsymbol{a} \in A$ can be of the form "*reduce the loan amount by* \$500 *and close* 1 *existing loan*".

The catalog A can be selected either based on the popularity of certain actions or by post processing a random sample of a fixed set from all the previously observed suggested and accepted actions. Or, an alternate procedure is to capture A by selecting centroids from k-means clustering of the approved subspace.

Calibration data $\mathcal{D}_{cal}$. The sanctity of calibration dataset is pivotal to the conformal literature. The second challenge of our framework is gathering calibration data $\mathcal{D}_{cal}$. For our context of conformal action sets, $\mathcal{D}_{cal}$ is formally defined as:

$$\mathcal{D}_{cal} = \left\{ (\boldsymbol{x}_i, \boldsymbol{a}_i) : f(\boldsymbol{x}_i) = -1, f(\boldsymbol{x}_i + \boldsymbol{a}_i) = +1, \boldsymbol{a}_i \in A \right\}_{i=1}^{\mathrm{n}} \tag{4.6}$$

Here, for any data point $(\boldsymbol{x}_i, \boldsymbol{a}_i) \in \mathcal{D}_{cal}$, we assume $\boldsymbol{a}_i \in A$. In this chapter, we synthetically generate an action $\hat{\boldsymbol{a}}_i \notin A$ for every $\boldsymbol{x}_i$ and map $\hat{\boldsymbol{a}}_i$ to the closest $\boldsymbol{a}_i \in A$. For simplicity, we assume that $\mathcal{D}_{cal}$ remains fixed in this chapter, which can be easily extended for a dynamically evolving $\mathcal{D}_{cal}$ as discussed in conclusion section of this paper.

4.2.2 Acceptance Probability

This subsection includes the essence of our work which is rooted in introducing the notion of *Acceptance Probability of an action* $\Pr(\boldsymbol{a} \mid \boldsymbol{x})$, which is defined as the probability that the action $\boldsymbol{a}$ is *preferred (desired) or accepted* by the individual $\boldsymbol{x}$. We know that:

$$\Pr(\boldsymbol{a} \mid \boldsymbol{x}) = \frac{\Pr(\boldsymbol{x} \mid \boldsymbol{a}) \Pr(\boldsymbol{a})}{\Pr(\boldsymbol{x})} \tag{4.7}$$

We begin by estimating the essential components of $\Pr(\boldsymbol{a} \mid \boldsymbol{x})$.

(i) $\Pr(\boldsymbol{x} \mid \boldsymbol{a})$**.** The first term in the numerator on the right hand side $\Pr(\boldsymbol{x} \mid \boldsymbol{a})$ represents the probability of a feature vector $\boldsymbol{x}$ conditioned on the action $\boldsymbol{a}$. An individual can select any action and an action can be preferred by a range of individuals. We know that if an individual $\boldsymbol{x}$ selects an action $\boldsymbol{a}$, then the **cost** of taking the action $\boldsymbol{a}$ is minimal.

Let $\mathrm{cost}(\boldsymbol{a} \mid \boldsymbol{x})$ define the difficulty of taking the action $\boldsymbol{a}$ suggested by the individual $\boldsymbol{x}$ with respect to the underlying population. $\mathrm{cost}(\cdot \mid \cdot)$ is typically measured using the *total log-percentile shift* introduced by [82], since it can accurately reflect the difficulty in taking actions in the target population and captures the notion of increasing the difficulty of action with features valued at a higher percentile.

For our presupposed $\mathrm{cost}(\boldsymbol{a} \mid \boldsymbol{x})$, we observe that:

$$\Pr(\boldsymbol{x} \mid \boldsymbol{a}) \propto -\ \mathrm{cost}(\boldsymbol{a} \mid \boldsymbol{x}) \tag{4.8}$$

Essentially, the farther the post action profile $\boldsymbol{x} + \boldsymbol{a}$ from $\boldsymbol{x}$, the lower the chance that the individual $\boldsymbol{x}$ chooses the action $\boldsymbol{a}$ and vice versa. This is due to the fact that a low cost action is always preferred by the user and if $\boldsymbol{a}$ is selected by her the chances of $\boldsymbol{x}$ being closer is higher. However, the other way around may not hold since several low cost actions are possible for $\boldsymbol{x}$.

Let $\overline{\text{quantile-shift}}\left(\boldsymbol{a} \mid \boldsymbol{x}\right)$ be estimated using any normalization based technique which transforms the raw percentile shift scores into probabilities, and we define:

$$\Pr\left(\boldsymbol{x} \mid \boldsymbol{a}\right) := -\ \overline{\text{quantile-shift}}\left(\boldsymbol{a} \mid \boldsymbol{x}\right), \tag{4.9}$$

where,

$$\overline{\text{quantile-shift}}\left(\boldsymbol{a} \mid \boldsymbol{x}\right) = \frac{\text{cost}(\boldsymbol{a} \mid \boldsymbol{x})}{\sum_{\boldsymbol{b} \in \mathcal{R}(\boldsymbol{x})} \text{cost}(\boldsymbol{b} \mid \boldsymbol{x})} \tag{4.10}$$

(ii) $\Pr\left(\boldsymbol{a}\right)$**.** The second term in the numerator on the right hand side in (4.7) refers to *marginal acceptance probability of action* $\Pr\left(\boldsymbol{a}\right)$ which fundamentally represents the prevalence of the action $\boldsymbol{a}$. $\Pr\left(\boldsymbol{a}\right)$ can be effortlessly approximated from the calibration dataset as:

$$\Pr\left(\boldsymbol{a}\right) \approx \frac{1}{\mathrm{n}} \sum_{i=1}^{\mathrm{n}} 1 \left\{(\boldsymbol{x}_i, \boldsymbol{a}_i) : \boldsymbol{a}_i = \boldsymbol{a}\right\} \tag{4.11}$$

$\Pr\left(\boldsymbol{a}\right)$ becomes critical in analyzing whether $\boldsymbol{a}$ is preferred or not, irrespective of its relative cost. For example, an action suggesting to increase the `MonthlyIncome` by $100 may seem easy to modify, but could be ineffective or infeasible (consider the case of a fixed salaried employees or people with fixed social benefits check).

(iii) $\Pr\left(\boldsymbol{x}\right)$**.** The third term $\Pr\left(\boldsymbol{x}\right)$ is estimated using a *probability density function* over the distribution of individual features $\mathcal{X}$. We can obtain $\Pr\left(\boldsymbol{x}\right)$ using a straightforward approach of approximating the density of the region around the individual profile $\boldsymbol{x}$ within the calibration data.

$$\Pr(\boldsymbol{x}) = \Pr(\mathcal{X} = \boldsymbol{x}) \tag{4.12}$$

Similar recourses suggested to different individuals are not equally actionable. This is exacerbated when sensitive (and often hidden) features are strongly correlated with other features. For example, [22] found that race has a profound correlation with the level of education of the individual. Such idiosyncrasies can be handled using $\Pr(\boldsymbol{x})$.

$\Pr(\boldsymbol{x})$ and $\Pr(\boldsymbol{a})$ are approximated using $\mathcal{D}_{cal}$, which can potentially be improved by a larger n and adaptive $\mathcal{D}_{cal}$. Now, substituting (4.9), (4.11) and (4.12) into (4.7) gives us $\Pr(\boldsymbol{a} \mid \boldsymbol{x})$. We are now equipped to generate $\mathcal{R}(\boldsymbol{x})$ in the next subsection.

4.2.3 Steps for Constructing action sets

$\Pr(\boldsymbol{a} \mid \boldsymbol{x})$ gives us the acceptance probabilities of all $\boldsymbol{a} \in A$ across all the data points in $\mathcal{D}_{cal}$. Notice that $\mathcal{D}_{cal}$ and $\{\boldsymbol{x}_{\mathrm{n}+1}, \boldsymbol{a}_{\mathrm{n}+1}\}$ have the principle of *exchangeability*, meaning that their joint distribution does not change for any permutation of the data points. In simple terms, the distribution does not depend on the order of the individuals in $\mathcal{D}_{cal}$. We now utilize the conformal predictions framework to form action sets, which we will discuss in the following steps.

(i) Quantifying uncertainty for recourse actions. $\Pr(\boldsymbol{a} \mid \boldsymbol{x})$ introduced in the previous subsection is an easy-to-understand concept designed to effectively convert the difficulty of the action into probabilities. Minimal percentile shift action resonates with a small intervention or an easy action, which is synonymous with a higher probability of acceptance. We leverage acceptance probability in this step, as the conformal literature [5] is not associated with any specific notion of *uncertainty quantification*.

(ii) Defining a nonconformity score function. A *nonconformity score* function $\mathrm{s} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ represents the error in the action suggestions. s can be straightforwardly obtained using:

$$\mathrm{s}(\boldsymbol{x}, \boldsymbol{a}) = 1 - \mathrm{Pr}\left(\boldsymbol{a} \mid \boldsymbol{x}\right) \tag{4.13}$$

We note that the *magnitude* of the scoring function does not have a meaning [5]. It is simply a measure of discrepancy of a new example from $\mathcal{D}_{cal}$. We will now use s to identify the empirical quantile of the required error rate.

(iii) Computing the calibration quantile. For any required error rate $\alpha \in [0, 1]$, the probability that the suggested *conformal action set* contains the desired action is almost exactly $1 - \alpha$. We determine an empirical quantile $\hat{\mathrm{q}}$ of the calibration scores $\mathrm{s}_1 = \mathrm{s}(\boldsymbol{x}_1, \boldsymbol{a}_1), \ldots, \mathrm{s}_\mathrm{n} = \mathrm{s}(\boldsymbol{x}_\mathrm{n}, \boldsymbol{a}_\mathrm{n})$ as.

$$\hat{\mathrm{q}} = \inf\left\{\mathrm{q} : \frac{|i : \mathrm{s}_1 \leq \mathrm{q}|}{n} \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right\} \tag{4.14}$$

(iv) Selecting actions set. Finally, utilizing the empirical $\hat{\mathrm{q}}$ for an unseen individual's feature vector $\boldsymbol{x}_{\mathrm{n}+1}$, we create the conformal action set as:

$$\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right) = \{\boldsymbol{a} : \mathrm{s}(\boldsymbol{x}_{\mathrm{n}+1}, \boldsymbol{a}) \leq \hat{\mathrm{q}}\} \tag{4.15}$$

Here, $\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right)$ includes all actions from the catalog with sufficient score, which guarantees the existence of the desired action with high probability. For example, with $\alpha = 0.05$, $\boldsymbol{a}_{\mathrm{n}+1}$ is guaranteed to be in $\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right)$ with probability at least 0.95.

4.2.4 Theoretical analysis

In this subsection, we provide guarantees for both the coverage of optimal action and the miscoverage penalty of conformal recourse for not suggesting the optimal action.

Remark 2 (Conformal recourse coverage)**.** Let $\{(\boldsymbol{x}_i, \boldsymbol{a}_i)\}_{i=1,\dots,n}$ and $(\boldsymbol{x}_{\mathrm{n}+1}, \boldsymbol{a}_{\mathrm{n}+1})$ be independent and identically distributed. For $\hat{\mathrm{q}}$ and $\mathcal{R}$ defined as (4.14) and (4.15) respectively, we have:

$$\Pr\left(\boldsymbol{a}_{\mathrm{n}+1} \notin \mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right) \mid \boldsymbol{x}_{\mathrm{n}+1}\right) \leq \alpha \tag{4.16}$$

Assumption 1. For an individual $\boldsymbol{x}_{\mathrm{n}+1}$, each feasible action $\boldsymbol{a}^{(i)} \in A$ is associated with a predetermined maximum percentile shift, that is, $\mathrm{cost}(\boldsymbol{a}^{(i)} \mid \boldsymbol{x}_{\mathrm{n}+1}) \leq \delta_i$.

Each action $\boldsymbol{a}^{(i)} \in A$ is associated with a certain cost of action for the individual $\boldsymbol{x}_{\mathrm{n}+1}$. (1) assumes that the cost of a particular action is bounded for any individual feature vector $\boldsymbol{x}$. An expensive action reduces the likelihood of its feasibility. Here feasibility implies that not the actions in $\mathcal{R}\left(\boldsymbol{x}\right)$ are feasible due to the fact that the actions are highly individual feature dependent.

Feasible cost amplification. We define *feasible cost* as the cost of feasible actions for $\boldsymbol{x}_{\mathrm{n}+1}$ in $\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right)$. Here, the set of feasible actions within the set $\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right)$ is denoted by $\mathcal{R}_{\mathcal{F}}\left(\boldsymbol{x}_{\mathrm{n}+1}\right)$. Our next result bounds the *feasible cost* amplification due to miscoverage.

We measure feasible cost amplification as the additional cost of actions an individual incurs due to miscoverage of the preferred action. For simplicity, we measure the increase in cost of action in terms of $\text{cost}(\cdot \mid \cdot)$ with respect to the actions in $\mathcal{R}_{\mathcal{F}}(\boldsymbol{x}_{\text{n+1}})$.

Definition 4.2.1 (Miscoverage penalty)**.** The feasible cost amplification $\Delta(\mathcal{R}(\boldsymbol{x}_{\text{n+1}}), \boldsymbol{a}_{\text{n+1}})$ of an action set $\mathcal{R}(\boldsymbol{x}_{\text{n+1}})$ is defined as:

$$\Delta(\mathcal{R}(\boldsymbol{x}_{\text{n+1}}), \boldsymbol{a}_{\text{n+1}}) = \min_{\boldsymbol{b} \in \mathcal{R}_{\mathcal{F}}(\boldsymbol{x})} \Big\{ \text{cost}(\boldsymbol{b} \mid \boldsymbol{x}_{\text{n+1}}) - \text{cost}(\boldsymbol{a}_{\text{n+1}} \mid \boldsymbol{x}_{\text{n+1}}) \Big\} \tag{4.17}$$

$\text{cost}(\cdot \mid \cdot)$ is any general purpose cost function to measure the action cost of $\boldsymbol{a}$ for $\boldsymbol{x}$. We also refer $\Delta(\mathcal{R}(\boldsymbol{x}_{\text{n+1}}), \boldsymbol{a}_{\text{n+1}})$ as the miscoverage penalty of the action set $\mathcal{R}(\boldsymbol{x}_{\text{n+1}})$ for not including the desired action $\boldsymbol{a}_{\text{n+1}}$. In line with the existing literature, we assume $\boldsymbol{a}_{\text{n+1}}$ to be the least cost recourses (arguably preferred by the user).

Theorem 1. The cost amplification $\Delta(\mathcal{R}(\boldsymbol{x}_{\text{n+1}}), \boldsymbol{a}_{\text{n+1}})$ due to *miscoverage* of desired action $\boldsymbol{a}_{\text{n+1}}$ is bounded by:

$$\Pr(\delta_{\min} < \Delta(\mathcal{R}(\boldsymbol{x}_{\text{n+1}}), \boldsymbol{a}_{\text{n+1}}) < \delta_{\max}) \leq \alpha \tag{4.18}$$

Proof. To prove Theorem 1, we start by observing the equality of the following two events:

$$\{\boldsymbol{a}_{\text{n+1}} \notin \mathcal{R}(\boldsymbol{x}_{\text{n+1}})\} = \{\Delta(\mathcal{R}(\boldsymbol{x}_{\text{n+1}}), \boldsymbol{a}_{\text{n+1}}) > 0\} \tag{4.19}$$

The left-hand term implies that the preferred low-cost action $\boldsymbol{a}_{\text{n+1}}$ is not covered by $\mathcal{R}(\boldsymbol{x}_{\text{n+1}})$ and the right-hand term implies that the cost amplification due to miscoverage is non-zero.

We note that the cost is only amplified when the preferred action is not covered by $\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right)$ and is 0 otherwise. We will now bound the term $\Delta\left(\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right), \boldsymbol{a}_{\mathrm{n}+1}\right)$.

The first edge case event to get a lower bound occurs when $\operatorname{cost}(\boldsymbol{a}_{\mathrm{n}+1} \mid \boldsymbol{x}_{\mathrm{n}+1}) = \delta_{\mathrm{min}}$. And using the definition of $\Delta\left(\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right), \boldsymbol{a}_{\mathrm{n}+1}\right)$, we have:

$$\Delta\left(\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right), \boldsymbol{a}_{\mathrm{n}+1}\right) > \delta_{\mathrm{min}} \tag{4.20}$$

Moreover, the other edge case scenario happens when $\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right)$ contains an infeasible action and $\boldsymbol{a}_{\mathrm{n}+1}$ is actually a high cost action then:

$$\Delta\left(\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right), \boldsymbol{a}_{\mathrm{n}+1}\right) \leq \delta_{\mathrm{max}} - \delta_{\mathrm{min}} \tag{4.21}$$

$\Delta\left(\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right), \boldsymbol{a}_{\mathrm{n}+1}\right)$ can now be upper bounded by the maximum quantile shift for any action denoted by δ_{max}, i.e.,

$$\Delta\left(\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right), \boldsymbol{a}_{\mathrm{n}+1}\right) < \delta_{\mathrm{max}} \tag{4.22}$$

Now, combining both (4.20) and (4.22), we get the following:

$$\delta_{\mathrm{min}} < \Delta\left(\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right), \boldsymbol{a}_{\mathrm{n}+1}\right) < \delta_{\mathrm{max}} \tag{4.23}$$

From (4.19) and (4.23), we can equate the two events as follows:

$$\begin{aligned} &\left\{\boldsymbol{a}_{\mathrm{n}+1} \notin \mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right)\right\} = \\ &\qquad \left\{\delta_{\mathrm{min}} \; < \; \Delta\left(\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right), \boldsymbol{a}_{\mathrm{n}+1}\right) \; < \; \delta_{\mathrm{max}}\right\} \end{aligned} \tag{4.24}$$

Now using Remark 2, we obtain Theorem 1. □

Additionally, from the definitions of δ_{min} and δ_{max}, we have a direct observation following Theorem 1:

$$\Pr\left(\Delta\left(\mathcal{R}\left(\boldsymbol{x}_{\mathrm{n}+1}\right), \boldsymbol{a}_{\mathrm{n}+1}\right) = 0\right) \geq 1 - \alpha. \tag{4.25}$$

Table 4.1: An EASY instance $\boldsymbol{x}_i$ with a lower value for `PastDue`. In this example, note that $\boldsymbol{a}^{(6)}$ and $\boldsymbol{a}^{(7)}$ are infeasible for $\boldsymbol{x}$. $\boldsymbol{a}^{(1)}$ is the optimal choice.

Features	**Current Values**	**Actions set**						
		$\boldsymbol{a}^{(1)}$	$\boldsymbol{a}^{(2)}$	$\boldsymbol{a}^{(3)}$	$\boldsymbol{a}^{(4)}$	$\boldsymbol{a}^{(5)}$	$\boldsymbol{a}^{(6)}$	$\boldsymbol{a}^{(7)}$
`PastDue`	5	−2	−2	−3	−4	-	−6	-
`Income`	$8,883	+$592	-	+$690	+$50	+$3,570	-	+$6,539
`Credits`	14	-	+1	-	-	-	+1	+1
`Loans`	0	-	-	-	-	-	-	−1

Result (4.25) implies that our framework provides action sets that guarantee non-amplification of feasible cost with high probability.

4.3 Analysis and Results

This section presents empirical evidence to show the effectiveness of the proposed framework for conformal action sets. We perform our experiments on three real-world datasets which focus on *banking* or *lending* domain. Our experimental goals are as follows:

- To analyze the empirical coverage of the action sets generated from the proposed framework on real world datasets.
- To show that the generated conformal action sets are adaptive across the population

Table 4.2: A HARD instance $\boldsymbol{x}_j$ with a higher value for PastDue. $\boldsymbol{a}^{(1)}$ is the optimal choice. **Reference** for this example is shared with Table 4.1.

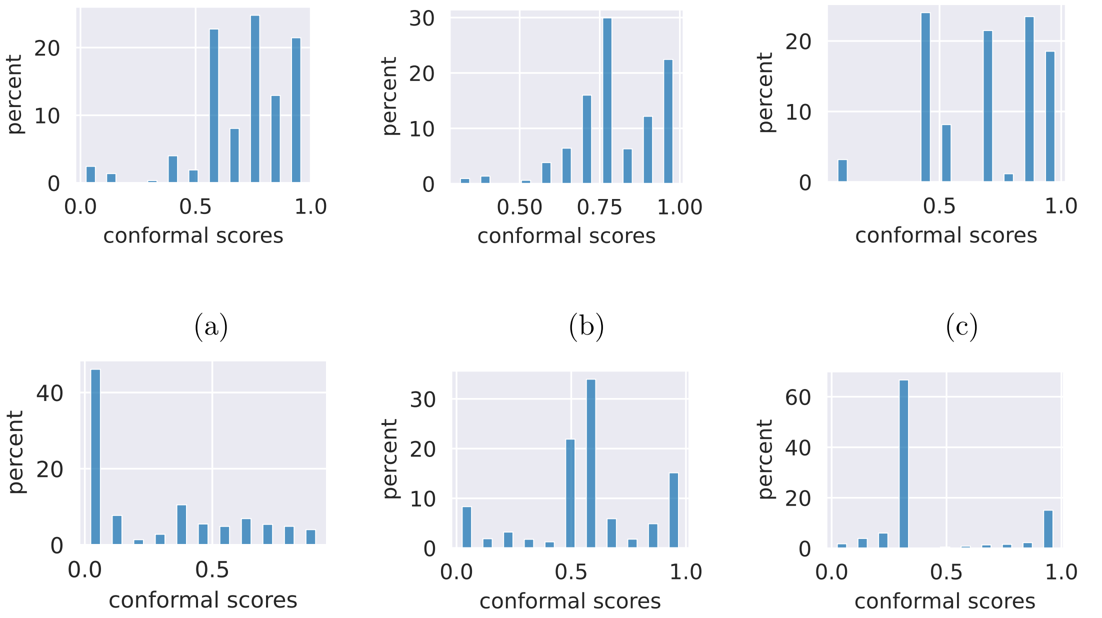
Features	**Current Values**	**Actions set**		
		$\boldsymbol{a}^{(1)}$	$\boldsymbol{a}^{(2)}$	$\boldsymbol{a}^{(3)}$
PastDue	7	−4	−6	−5
Income	\$6,265	+\$50	-	+\$70
Credits	14	-	+1	-

and also across sub-populations.

- To analyze the adaptability of action sets using qualitative examples.

4.3.1 Setup

Datasets. We train Logistic Regression models for a binary classification task on processed versions [82] of the three datasets: (i) german [9], (ii) credit [91] and (iii) givemecredit [36]. Here GoodCustomer, NoDefaultNextMonth and SeriousDlqin2yrs are the corresponding target labels for german, credit and givemecredit datasets, respectively. For GoodCustomer and NoDefaultNextMonth, 1 is the desired positive prediction whereas, for SeriousDlqin2yrs, 0 is the desired prediction; an individual is adversely affected if their prediction is otherwise.

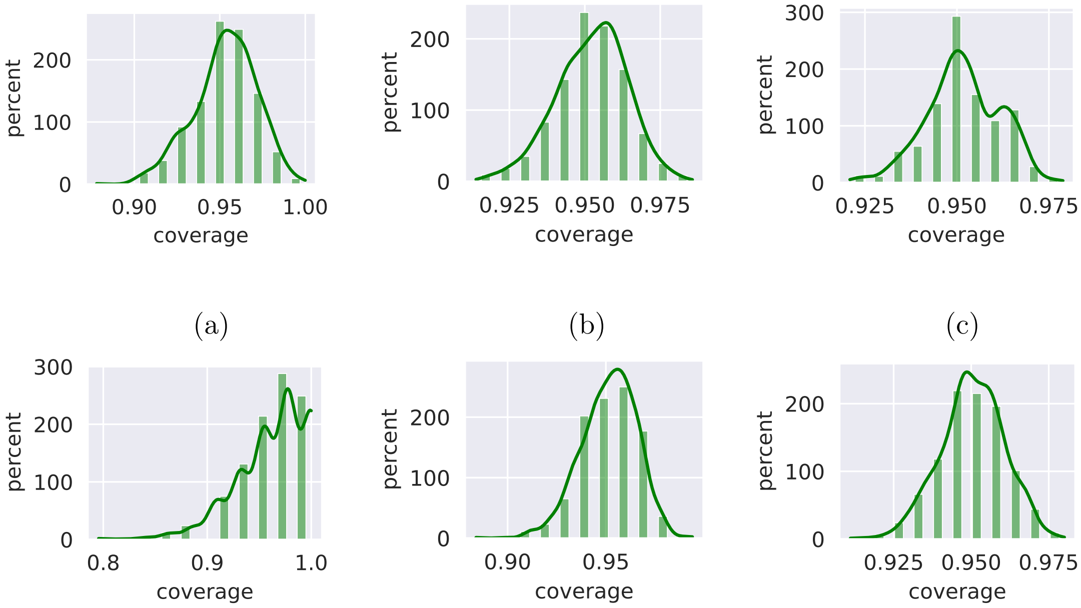


(a) (b) (c)

(d) (e) (f)

german givemecredit credit

Figure 4.2: Histogram of the Conformal scores. The top (a), (b), (c) sub-graphs are for the LR model, and the bottom (d), (e), and (f) sub-graphs are for the RF model.

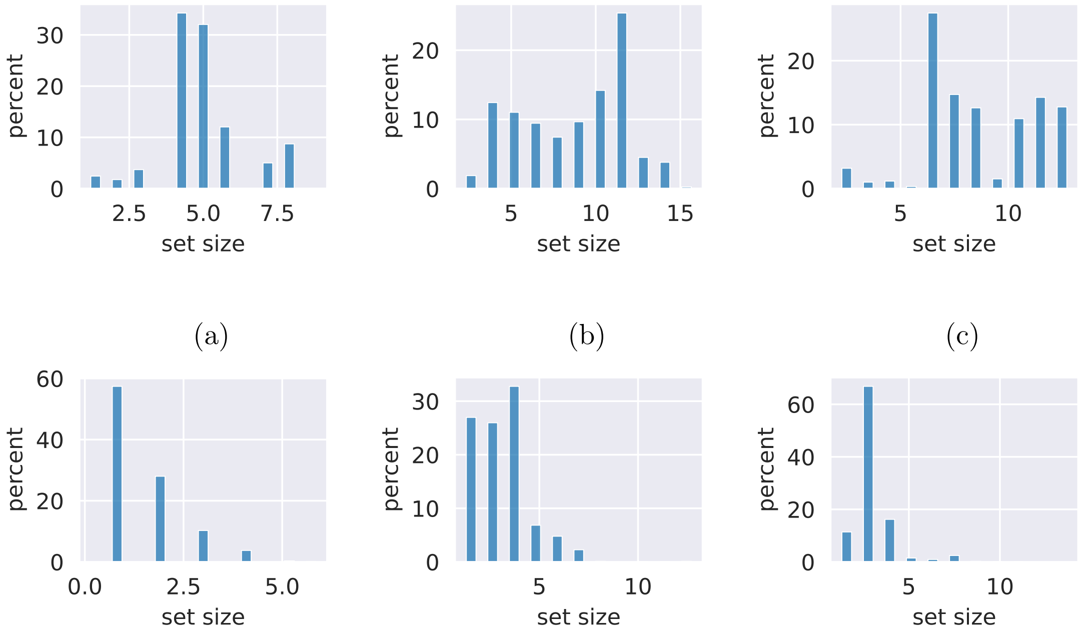


(a) (b) (c)

(d) (e) (f)

`german` `givemecredit` `credit`

Figure 4.3: Distribution of average empirical coverage across 1000 independent runs. The top (a), (b), (c) sub-graphs are for the LR model, and the bottom (d), (e), and (f) sub-graphs are for the RF model.

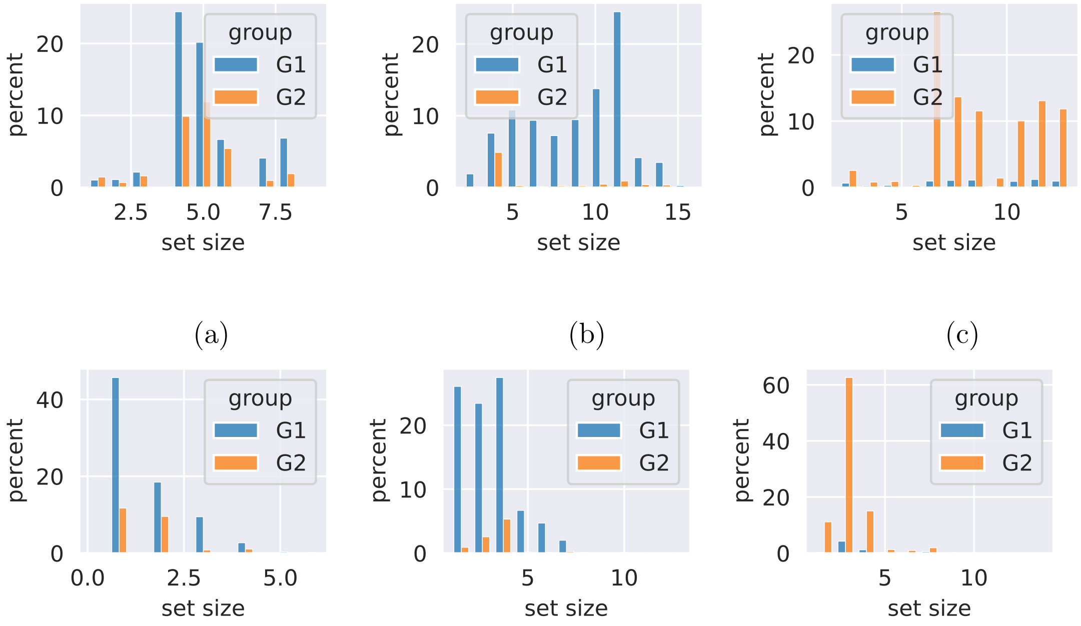


(a) (b) (c)

(d) (e) (f)

`german` `givemecredit` `credit`

Figure 4.4: Conformal Recourse Set sizes over a sample of 1000 independent random data splits for `german`, `givemecredit` and `credit` dataset. The top (a), (b), (c) sub-graphs are for the LR model, and the bottom (d), (e), and (f) sub-graphs are for the RF model.



(a) (b) (c)

(d) (e) (f)

`german` `givemecredit` `credit`

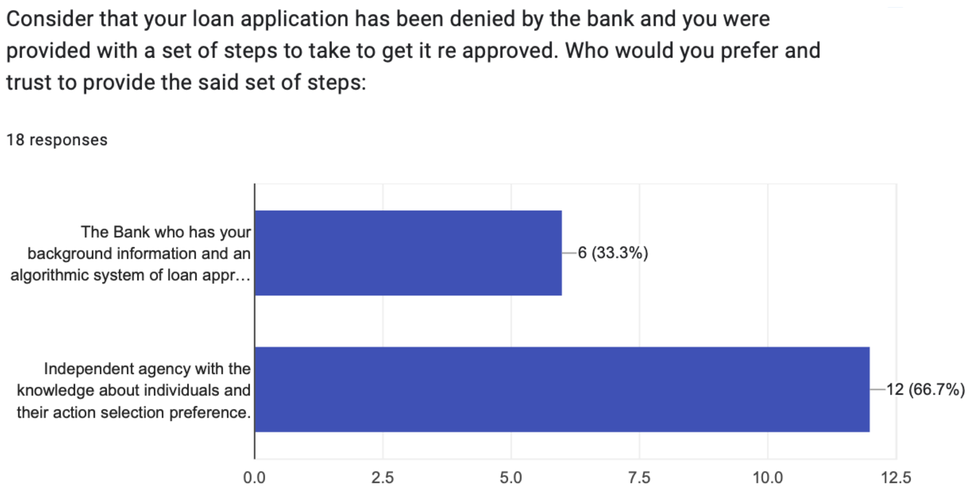
Figure 4.5: Conformal Recourse Set sizes stratified based on gender for `german`, age (≥ 25) for `givemecredit` and age (< 25) for `credit` datasets respectively. The top (a), (b), (c) sub-graphs are for the LR model, and the bottom (d), (e), and (f) sub-graphs are for the RF model.

Baseline models. We perform our analysis on the Logistic Regression (LR) and Random Forest (RF) classifiers. Each type of model is trained with three different datasets discussed above. The results provided differs only in the score function while estimating $\Pr(\boldsymbol{x} \mid \boldsymbol{a})$, with an additional criterion of the action being *valid* i.e., $f(\boldsymbol{x} + \boldsymbol{a}) = +1$

Nonconformal scores. The number of individuals adversely affected by the models decision are 864, 2369 and 1678 for `german`, `credit` and `givemecredit` respectively. Integer Programming technique of [82] is used to secure a low-cost recourse action for our models, which we consider to be the desired action. A fixed action catalog A with 40 actions is *randomly* selected for the three datasets separately. The choice of $k = 40$ to denote the size of A is made using human judgement using some domain knowledge. The size n of $\mathcal{D}_{cal}$ for each dataset is set to 70% of their corresponding number of denied individuals. And the remaining 30% denied individuals are used to evaluate our framework. The results shown in this section are an average of 1000 independent random data splits across the datasets, and the results are reported for $\alpha = 0.05$.

4.3.2 Analysis

Coverage. We show the distribution of the conformal scores to capture the empirical quantile in Figure 4.2. The average coverage across the three datasets for 1000 independent runs is 95% with a standard deviation of 2%. An illustration of the distribution of coverage across the runs is shown in Figure 4.3 and we notice that the coverage follows a normal distribution with the empirical average centered around $1 - \alpha$.



Consider that your loan application has been denied by the bank and you were provided with a set of steps to take to get it re approved. Who would you prefer and trust to provide the said set of steps:

18 responses

Figure 4.6: User preference analysis chart from an anonymous survey. The survey concludes that most people trust an individual entity to provide actionable recourse regardless of the existing ML model

.

Adaptive sets. Figure 4.4 shows the adaptability of the set sizes using a histogram plot. Although there are 40 actions in A, our mechanism provides fewer individuals with larger set sizes. We notice a high frequency of smaller set sizes across the datasets.

Table 4.1 illustrates an *easy* instance from the `givemecredit` dataset and the corresponding actions set. The action $\boldsymbol{a}^{(1)}$ highlighted in green is the true desired action. On the contrary, Table 4.2 contains a *hard* instance with a relatively smaller actions set. The hardness of the instance can be observed in the higher value of 7 for the feature `NumberOfTime30-59DaysPastDueNotWorse`.

Please observe that not all suggested actions are feasible for an individual. However, our high-probability guarantees ensure the coverage of true action in the suggested actions set. With a finite space A, the actions set $\mathcal{R}\left(\boldsymbol{x}\right)$ suggested for $\boldsymbol{x}$ can contain *infeasible* actions. For example, in Table 4.2, actions $\boldsymbol{a}^{(6)}$ and $\boldsymbol{a}^{(7)}$ are infeasible. Here in $\boldsymbol{a}^{(6)}$, `NumberOfTime30-59DaysPastDueNotWorse` can not be lower than 5 and in $\boldsymbol{a}^{(7)}$, `NumberRealEstateLoansOrLines` can not be negative.

Group-stratified adaptive sets. In addition to the previous results, we show the adaptability of recourse set size and further show group stratified set sizes. The sub populations selected to analyze the performance of set size adaptability are: (i) `male` and `female` groups for `german` dataset, (ii) `age` ≥ 25 and otherwise for `givemecredit` dataset and (iii) `age` < 25 for the `credit` dataset.

For further analysis, we plot Figure 4.5, which illustrates the distribution of set sizes using group-stratified histogram plots.

4.4 Discussion and Conclusion

With ever evolving modeling and decision making environments, the assumption of a fixed $\mathcal{D}_{cal}$ over a period of time may become impractical in several critical domains. This can be addressed using a strategically evolving dynamic $\mathcal{D}_{cal}$, which gets updated by one of the following design choices: (i) replacing an oldest entry with a recent successful $(\boldsymbol{x}, \boldsymbol{a})$ pair, or (ii) frequency weighted score function to capture the desirability via popularity of any denied profile and recourse combination.

Distributional shift in A. Calibration dataset is initially synthesized using in-house techniques. However, as test instances are obtained, the calibration data can be re-evaluated in the same manner as distributional drift. After ℓ test samples with available user recourse selection $\mathcal{D}_{cal} = \{(\boldsymbol{x}_i, \boldsymbol{a}_i)\}_{i=\ell}^{\mathrm{n}+\ell}$. We can either pick a rolling window of m or a domain specific smooth decay:

$$\mathrm{w}_i^{(1)} = 1\{i \geq m - \ell\} \quad \text{or} \quad \mathrm{w}_i^{(2)} = 0.99^{m-i+1} \tag{4.26}$$

Validity of suggested Actions. The idea of an independent entity suggesting a recourse raises a critical question of whether the recourse is valid across models. Here, validity refers to the fact that if the individual acts on the recourse, the model provides a favorable decision.

Future work in this direction must consider a *cost tolerance threshold* δ for each model, which means that the models must be flexible enough to accommodate δ while making a favorable decision to an individual.

Our proposed framework can also help reduce modeling bias by segregating the recourse mechanism from a model. To enforce the models to be fairer, δ can be bounded for each model and a model exhibiting a higher δ signals modeling bias.

Anonymous User Survey. To understand the choice of recourse mechanism entity for building trust for ML models into the society, we conducted an anonymous one question survey.

The goal of this survey was to analyze if the proposed framework of an independent entity can help build trustworthy framework of accountability and user acceptance. We asked participants if they preferred the model provider or an independent entity to provide actionable recourse. Specifically, the survey included the following question and choices:

Consider that your loan application has been denied by the bank and you were provided with a set of steps to take to get it re-approved. Who would you prefer and trust to provide the said set of steps?

- *The Bank who has your background information and an algorithmic system of loan approval.*
- *Independent agency with knowledge about individuals and their preference for action selection.*

From the one-question survey, we observed that an overwhelming majority of approximately 67% of the participants preferred an independent (ethical) agency to oversee the data-driven recourse generation mechanism. The snapshot of the results of our anonymous survey is illustrated in Figure 4.6.

Broader Impact statement. Our framework supports the notion of an independent entity that supervises and audits the ethical aspects of the recourse mechanism for ML models. Furthermore, such an entity can take the responsibility of providing a truly model-agnostic set of actions, specifically calibrated for an individual with guarantees of user actionability.

The proposed framework can accommodate action preferential effects due to dynamic macro- and micro-level policy changes. These effects are intuitively captured by user action selections and dynamic calibration dataset. For example, with increasing interest rates, individuals would prefer reducing the loan amount to getting a co-borrower. Such action selection would be reflected in the proposed approach, but not with a fixed policy of low-cost recourse.

Concluding remarks. We provide an essential toolkit for providing action sets with guarantees of high probability of action acceptance. This system is essential in a black-box setting of an ML model, where obtaining high quality model information or individual preferences is challenging. Even with lack of such information, our novel, yet intuitive framework can assist in providing high quality action sets with additional independent audit benefits.

The action sets provided can also be used to understand the difficulty of identifying an action for an individual. An individual closer to a decision boundary will have a larger $\mathcal{R}$ compared to the individual farthest from the decision boundary. With this chapter, we bring a new perspective to the generation and auditability of recourse actions for ML models.

Chapter 5

Parallelism of Counterfactuals to Multifaceted Reformulations in e-Commerce

5.1 Introduction

This chapter discusses other unexplored arenas where counterfactuals are applicable for improving trustworthiness. We start this chapter by introducing the notion of Loser search queries in an e-commerce domain and its parallelism to the counterfactual literature.

The performance of search engines in e-commerce domain is measured by several crucial business metrics like *purchases* and *click-through rate*. A drop in these metrics indicates that user intent is not captured accurately. Here, *precision* measures how many items retrieved are relevant and *recall* measures how many relevant items were retrieved.

In this setup, there is an apparent vocabulary gap between sellers and buyers; making retrieval problem extremely complicated. Improving user experience in by capturing user's search intent and retrieving relevant items from the inventory catalog is a core e-commerce [70] problem.

A search engine retrieval system often employs a complex combination [81] of Machine Learning (ML) models and human-defined rules that aim to return the list of items that best interpret the user query. In a major e-commerce company *eBay*, approximately 25% queries over a span of 6 weeks result in zero or low *relevant inventory set* in the *Search Result Page* (SRP). These are referred to as N&L queries for the rest of the paper and suffer from a deteriorating user engagement.

Typically, a naive baseline N&L recovery [54, 81, 31] triggers a *heuristic* recovery model, which matches empirically determined fraction of query tokens with the catalog item titles and drops the rest of query tokens or terms (since matching all the tokens provides insufficient *recall set*, which represents the set of items retrieved). This technique can result in a better recall set while compromising on precision in query-item relevance. For example, an N&L query, *old flower decoration tea pot set* might retrieve items that match two tokens with the item titles regardless of the linguistic signals, which may not even show the core source query intent of *tea pot set*.

Furthermore, we have identified multiple category interpretations for a significant portion of search queries from user search logs at *eBay*. This suggests that deriving multiple reformulations for a query and using it for retrieving additional relevant recall can be highly beneficial when the original query is N&L. With diversity measured in terms of a category taxonomy derived from *eBay* inventory, we have identified that: "*Approximately* 29% *of user-issued reformulations (of the same source query, by the same user, in the same session) belong to different item categories*". For example, *womens gothic clothing* can be interpreted as both *womens gothic dresses* and *womens gothic skirts*, the items of which belong to different categories. At a high level, the goal of the proposed novel method is bifold:

1. *Providing multiple alternative reformulations to enhance relevant recall for N&L queries.*

2. *Ensuring that the reformulations are diverse to enhance user engagement.*

Our key contributions with this chapter include:

1. We identify and consolidate the conceptual relations between *counterfactual explanations* and N&L query reformulations. To the best of our knowledge, this is the first work to study these similarities.

2. Formulating the N&L reformulation problem into an NMT-based framework, designing a Multi-Seq2Seq model, and proposing a *diversity-inducing optimization* function.

5.1.1 Related Work

Although seemingly straightforward, the search domain comes with several inherent research problems such as *query intent detection*, *personalized recommendations*, *query expansion*, *query reformulation* etc. Search engine also suffers from noisy buyer input such as misspellings, over-specification [47], under-specification, token reordering, and unpopular synonyms. A recent study [81] provides the architecture of the search engine. Query reformulation captures user intent and retrieves relevant items from the database. An in-depth analysis of query reformulation based on query logs of a search engine is studied by [30] and studies [32, 33, 57, 53, 24, 8] provides relevant literature on query reformulation.

Statistical Machine Translation viewpoint for query rewriting has been explored by [68], where the authors formulate the query reformulation problem as a monolingual translation problem. Minimum risk training (MRT) [73] has been widely used to train machine translation models, which aims to minimize the expected loss of the training data. Recent works have also actively explored machine translation for query rewriting [65, 11]. A similar approach is also previously studied for N&L query reformulation [79], where the authors proposed a system to provide multiple reformulations to improve the relevant recall performance of N&L queries. The widespread popularity and effectiveness of NMT [76, 10] are also explored for query reformulation. However, these studies lack the flexibility to capture inherent diversity in the reformulations.

An alternate field of study to identify a counterfactual [56, 84] instance (for a given instance) to obtain an alternate desired decision from a machine learning model has recently gained popularity for both robustness and explanation properties. To the best of our knowledge, this is the first study to identify the optimization and conceptual similarities between N&L reformulation and counterfactual explanation.

5.2 Problem Formulation

Given a Search Engine (SE) that takes a user query q and retrieves the relevant items from the database. With a fixed *item index space* and *retrieval mechanism*, we identify q as N&L query if the number of items retrieved for q is less than a threshold τ. Let $SE\left(q\right) = \{i_1, i_2, \ldots, i_{n_q}\}$ be the set of items recovered and let f be a classifier to identify if q is N&L or not such that: $f\left(SE\left(q\right)\right) = 1$ if $n_q < \tau$ and 0 otherwise. If $f\left(SE\left(q\right)\right) = 1$, a query reformulation procedure is triggered for q to retrieve items from a reformulated query r with the minimum user intent disparity.

Let $intentDisparity\left(q_1, q_2\right)$ mimic *user intent disparity* between two queries $q_1,\ q_2$, and $\Gamma(q)$ be the set of all *plausible* reformulations for q. Plausibility refers to acceptable reformulation behavior such as term dropping or replacement. Traditional N&L reformulation aims to identify r with minimum intent disparity [32] with the user-specified query q.

A conservative solution is to recursively obtain r by dropping tokens from q until $n_q \geq \tau$, or until a pre-determined fraction of tokens has been dropped. However, such a greedy approach makes it impractical, since we might lose semantic meanings of the original query. A better solution is to train a deep learning model to learn user-intent-driven behavior.

Nevertheless, these models still lack the capability of capturing the prevailing ambiguity in N&L user query interpretations. This study aims to address such ambiguities and improve user SRP experience for an N&L query by diversifying the items retrieved.

Building on the existing deep-learning-based solution, we aim to obtain (at least) two diverse reformulation by solving the following optimization:

$$\begin{aligned} \operatorname{argmin} \{r1, r2\} \quad & \sum_{i \in \{1,2\}} \mathit{intentDisparity}\,(ri, q) && \text{(feasibility)} \\ s.t. \quad & f\,(r1) = 0, \quad f\,(r2) = 0 && \text{(validity)} \\ & r1, r2 \in \Gamma\,(q) && \text{(plausibility)} \\ & \lambda_{r1,r2} \geq \lambda^{*} && \text{(diversity)} \end{aligned}$$

where $\lambda_{r1,r2}$ is the diversity score between two reformulations $r1$ and $r2$. Here, λ^{*} is the pre-determined threshold representing minimum required diversity score. We consider: $\lambda_{q1,q2} = 1 - \mathit{Jacc}\,(q1, q2)$

Note that the definition of $\lambda_{r1,r2}$ is domain dependent and our selection of the Jaccard-based metric is motivated by ease of understanding. We use K *grams* (with $K = 3$) based Jaccard similarity score defined as:

$$\mathit{Jacc}\,(q1, q2) = \text{3grm}\,(q1) \cap \text{3grm}\,(q2) / \text{3grm}\,(q1) \cup \text{3grm}\,(q2) \tag{5.1}$$

where $\text{3grm}\,(q)$ is the consecutive set of 3 character words from all the tokens in q. A higher $\lambda_{q1,q2}$ implies greater diversity in items retrieved between the two queries.

5.3 Lessons from Counterfactual Explanations

A related field of study, *Counterfactual Explanations* [86, 23] provides the basis for obtaining a desired prediction from an ML model. Counterfactual generation identifies an instance closer to the data point, which can alter the model's behavior. Consider a classification-based model $g(x)$ which classifies an instance x to belong to class C_1. [86] solves to find a counterfactual x' which belongs to class $C_2 \neq C_1$ for x using the optimization: $\operatorname{argmin} x' \;\; dist\,(x, x') \;\; \text{s.t.} \;\; g\,(x') = C_2$.

Our proposed optimization of N&L query reformulation is analogous to this field of study. We consolidate the shared optimization similarities between *counterfactual generation* and N&L query reformulation, prominent to this study:

1. **Feasibility**: If real-world implications of the ML model's decision adversely affect an instance, *actionable recourse* [83] provides the desired outcome from the ML model. Analogously, the search science domain intends to capture user intent for an N&L query and provide a *reformulation*. Feasibility of a reformulated query is determined by its closeness to the source query. The higher the similarity of r with q, the higher the feasibility of r.

2. **Validity**: A counterfactual, flips the model prediction from its prediction of the original instance. The reformulation for an N&L query is similarly aimed at improving the relevant recall set size to be at least τ. A valid reformulation improves the user experience of an N&L query by increasing the relevant recall set size.

3. **Plausibility**: A counterfactual is highly plausible with respect to the training data. In search science space, plausibility can be interpreted as the acceptable reformulation behavior. Examples of such behaviors are term dropping, synonym replacement, or misspelling correction.

4. **Diversity**: Given the randomness of counterfactual generation techniques, a wide range of diverse counterfactuals [56] are possible for a given instance. A similar observation of diverse user query interpretation is noticed between multiple user reformulations for the same query.

5.4 Dataset Generation Framework

5.4.1 Training data considerations

We gather user behavioral data from the historical search reformulations with improved user engagement. We extracted *six weeks* of search logs and constructed *three different versions* of datasets based on the reformulation behavior with the following steps.

5.4.1.1 SRP bursts

A user enters a *search query* q, then reformulates it to a *reformulated query* t_1 and reformulates it again if necessary. We assume a *user session* to last for about 10 minutes and consolidate all the search information in that session into an *SRP burst*. Every SRP burst signifies a sequence of successful user query transitions/reformulations along with user engagement signals. Making the session longer might diverge users' original intent. Thus by increasing the user burst size, we will only sample more 2 hop pairs. For sampling more pairs, we increase the dataset window, instead.

Data	Source query	Hop-1 query	Hop-2 query
TD	apple watch nike sport band	~~apple~~ watch nike sport ~~band~~	~~apple watch~~ nike sport band
	vintage gothic shirts tight	vintage gothic shirts ~~tight~~	~~vintage~~ gothic shirts tight
TDCR	apple watch band official	apple watch band nike	apple watch band ~~official~~
	womens gothic clothing	womens gothic dresses	womens gothic leggings
TDR	apple watch sport	apple watch nike	apple watch nike se 44
	vintage gothic	vintage gothic dress	vintage gown

Table 5.1: Real-world examples of user-reformulated targets from the training dataset extracted from (real world) raw data **Reference characteristics** — ~~Strikethrough~~: Dropped tokens, Red: Replaced tokens, Blue: Added tokens

5.4.1.2 Data generation

Every consecutive *hop-1* and *hop-2* reformulations are considered to be valid one neighbor and two neighbors (away) from the source query. For a user query q, let the corresponding user-reformulated targets be $t1$ and $t2$. Acceptability of $t1$ and $t2$ is established by an increase in user engagement, measured by a *User Engagement Score*[1], $ueScore(q)$ for q. We use successful user engagement as an approximation for ground truth from eBay's standpoint. A typical $ueScore(\cdot)$ is a *linear combination* of multiple signals like *user clicks*, *active time spent*, and actions like *add to cart.* A valid user query transition shows a minimum increase of 10% (established by a domain expert).

- *Independence of targets $t1$ and $t2$:* For simplicity, we consider that both $t1$ and $t2$ are

[1] This is a highly confidential score and can not be shared outside of eBay. Readers can synthetically estimate this for experimental purposes.

conditionally independent. This is due to the fact that both the targets are derived (with some minor modification) from the same source query q. Here, a user can go to either of the targets from q, implying that t_1 is not an intermediate query. This fact is also backed by our manual inspections of the training dataset and sample examples are illustrated in Table 5.1.

5.4.1.3 Capturing Diversity in user reformulations

With the criteria of a minimum $Jacc(\cdot)$ score between t_1, t_2 and q, both t_1 and t_2 are considered valid reformulations capturing the the user intent of q. However, we also observe from the training data that items retrieved using t_1 and t_2 come from multiple categories, indicating diversity in interpretations. For targets t_1, t_2 and model output $r1$, $r2$; our solution discussed below aims to *minimize* the *reformulation* loss between the pairs $(t_1,\ r1)$ and $(t_2,\ r2)$ along with *maximizing* the *diversity* loss between the pairs $(t_1,\ r2)$ and $(t_2,\ r1)$.

5.5 Diversity Induced model training

In this section, we propose our solution anchored on user behavioral data to provide multiple reformulations for a N&L query.

An instance in the training data is a triplet of $\langle q, t_1, t_2\rangle \in \mathcal{D}$ and for the rest of this study, we call user reformulated queries in training data as *targets* and model predictions as *reformulations*. With the established sanctity of the training data, we will now propose a solution to the optimization for N&L reformulation in Section 5.2 using an NMT-based approach.

5.5.1 Diversity Induction

Consider any source query $q = q_1, \ldots, q_k, \ldots, q_M$ consisting of M tokens, where q_k represents the k^{th} token. And let corresponding target queries be; $t\imath = t\imath_1, \ldots, t\imath_{j_1}, \ldots, t\imath_{N_1}$, and $t\mathit{2} = t\mathit{2}_1, \ldots, t\mathit{2}_{j_2}, \ldots, t\mathit{2}_{N_2}$. We model the target translation probabilities as:

$$\Pr\left(ti|q; f\right) = \prod_{j_i=1}^{N_i} \Pr\left(ti_{j_i}|q, ti_{<j_i}; f\right) : i \in \{1, 2\} \tag{5.2}$$

where f represents the model parameters and $t\imath_{<j_1} = t\imath_1, \ldots, t\imath_{j_1-1}$ is a partial translated query. As discussed in the previous section, we assume that the translation probability of hop-2 reformulation $t\mathit{2}$ is conditionally independent of hop-1 transition $t\imath$ i.e., $\Pr\left(t\mathit{2}|t\imath, q; f\right) \approx \Pr\left(t\mathit{2}|q; f\right)$

In other words, any reformulated query ri in the SRP burst for any N&L query q should have a high intent similarity with source query q irrespective of $t\imath$ while improving the performance of retrieved items in terms of recall and relevance.

We incorporate a *Diversity loss* $\mathcal{L}_{Div}$ component intended to *maximize* the diversity between the decoder predictions in conjunction with the traditional *Reformulation loss* $\mathcal{L}_{Ref}$ component intended to *minimize* the error between the training targets and model predictions. For $t\imath$ and $t\mathit{2}$, let $r1$ and $r2$ as the corresponding predicted reformulations by the model. Let $\ell\left(ri, ti\right)$ be any *loss function* to measure the disagreement between the model prediction ri and the training sample ti. For simplicity, we choose the *crossentropy* loss as a representative ℓ. For a given training dataset $\mathcal{D} = \{\langle q^{(s)}, t1^{(s)}, t2^{(s)} \rangle\}_{s=1}^{S}$, the training objective is to minimize the total loss:

$$\hat{f} = \operatorname{argmin} f \left\{ l(()) \, f \right\} \quad \text{s.t.} \quad l(()) \, f = \mathcal{L}_{Ref}\left(f\right) - \alpha \cdot \mathcal{L}_{Div}\left(f\right),$$

where:

$$\mathcal{L}_{Ref}\left(f\right) = \sum_{s=1}^{S} \sum_{i \in \{1,2\}} \ell\left(ri^{(s)}, ti^{(s)}\right), \tag{5.3}$$

$$\mathcal{L}_{Div}\left(f\right) = \sum_{s=1}^{S} \sum_{(i,j) \in \{(1,2),(2,1)\}} \ell\left(ri^{(s)}, tj^{(s)}\right). \tag{5.4}$$

5.5.2 Significance and estimation of α

The effect of the seemingly adversarial component $\mathcal{L}_{Div}$ is controlled by α, which represent the intended diversity score. We approximate α from the training data such that $\alpha \approx \lambda_{tr}$.The *training diversity score* λ_{tr} can be leveraged to tune the reformulation diversity by the model. We define *training diversity score* as: $\lambda_{tr} = 1 - \frac{1}{S} \sum_{s=1}^{S} Jacc\left(t1^{(s)}, t2^{(s)}\right)$.

The domain-specific diversity score can be defined as per the business needs. For instance, we recognize various types of *intent diversities* in the e-commerce domain like: (i) *Categorical diversity*: where the user targets correspond to items from different categories, and, (ii) *Aspect diversity*: where the user targets fetch items with different aspects/attributes.

5.5.3 Multi Sequence-to-Sequences (Multi-Seq2Seq) Model

An effective solution to language modeling using Machine Translation has found widespread usage in query rewrites. Building on the traditional transformer-based Sequence-to-Sequence (Seq2Seq) architecture [50, 76], we propose one encoder and two decoder approach with a shared loss function. Each sample in the training data consists of one source query and two user reformulation targets, and each decoder in the proposed architecture learns the user reformulation behavior for the corresponding target. The components of the proposed optimization function capture diversity-induced translation behavior. For each of the three dataset versions, we train a Multi-Seq2Seq model.

The loss function from Section 5.5 is shared between the two decoders and aims to *minimize reformulation error and maximize weighted diversity error*. The model is trained offline, and when an N&L query is encountered, the *Inference phase* is triggered. Reformulated queries $r1$ and $r2$ are predicted for input query q. These queries are then used to fetch item recall set by the SE. For each dataset version, we learn model parameters and let θ_{td}, θ_{tdcr}, and θ_{tdr} be the learned models on the training datasets TD, TDCR, and TDR, respectively.

5.6 Conclusion and future work

With this study, we introduced the necessity for diverse reformulations for a N&L query in the e-commerce domain. The diversity between reformulations is captured by considering the targets from multiple hops within a user session. This can be extended

Model	**Source query**	**Reformulated queries**
θ_{td}	*iphone 13 mini armor case shockproof*	iphone 13 mini armor case iphone 13 mini shockproof case iphone 13 mini case shockproof iphone 13 mini armor shockproof
θ_{tdcr}	*iphone 13 pro max case casetify*	iphone 13 pro max case iphone 13 pro max case otterbox iphone 13 pro max case spigen iphone 13 pro max case cute
θ_{tdr}	*iphone 13 pro max*	iphone 13 pro max case iphone 13 pro max unlocked iphone 13 pro max

Table 5.2: Model reformulations consolidated from both decoders

using other domain-specific diversities between the reformulations. In the e-commerce domain, one can also consider targets from different categories as a diversity signal.

Our work intends to motivate further study into exploring and capturing user diversity behavior for multifaceted reformulation for bad-performing user queries. We show that counterfactual literature and N&L query reformulation shares conceptual properties, and our study motivates further research to bring these fields closer.

Chapter 6

Conclusion

6.1 Summary and Takeaways

This thesis motivates other arenas of exploring principles of trustworthiness of an algorithmic model. With consistent evolving of machine learning model capabilities, their performance in terms of building trust needs to be re-imagined with ever evolving principles. In this section we discuss a few of the ideas for further research within this context.

Contemporary research has often delved into using standard performance metrics such as recourse cost and success rate. However, from the perspective of trustworthiness, these metrics may not be the best judge. A high success rate does not give a clear picture in case of a class-biased dataset. Similarly, the cost of the recourse often does not capture the individual difficulty in acting on the suggested recourse. One possible research direction is measuring cost of action in terms of time taken to update a feature. This approach has the potential to provide adversely affected individuals with the flexibility to act on immutable characteristics.

Building trust within the society is an extremely complex tax. Machine learning models are often blindsided by their performance on the test dataset in terms of selective performance metrics. However, I argue that a live machine learning models performance and trust can be significantly improved by bringing all the stakeholders in the decision making process. Having open discussions and providing realistic expectations to society about the capabilities of the deployed model will have a huge uplifting impact.

The end goal of any model should be focused towards improving the overall health and well-being of the society. This can open up a new direction of delayed performance evaluation and the taking of active feedback from both positively and negatively affected individuals by the model.

Another crucial takeaway with this research is to recall that one size or solution does not fit across all the domains of model deployments. Hence, domain specific evaluations and domain specific end user feedback loop can help build trustworthy models.

6.2 Concluding Remarks

In this study, we propose to capture different forms of user preferences and propose an optimization function to generate actionable recourse adhering to such constraints. We also provide an approach to generate a connected Laugel2019IssuesWP recourse guided by the user. We show how UP-AR adheres to soft constraints by evaluating user satisfaction in fractional cost ratio. We emphasize the need to capture various user preferences and communicate with the user in a comprehensible form. This work motivates further research on how truthful reporting of preferences can help improve overall user satisfaction.

In this work, we outline a new approach to account for latent groups in applications where we wish to provide recourse. In particular, we developed machinery to identify such groups from data and studied the implicit disparity in plausibility across these groups. For example, suggesting naive and arguably famous recourse action of increasing the working hours to a *single parent* is not feasible. We proposed a method to train classifiers to mitigate these effects and demonstrated their capacity in practice.

Limitations. Group-level plausibility may not ensure individual actionability [see e.g.,][]kothari2023prediction. Our proposed approach may also exacerbate the cost of recourse. Our study raises the question of whether it is sufficient for a recourse to change the model's decision or whether a recourse improves the affected individual's overall group-level profile.

Broader Impact statement. Our framework supports the notion of an independent entity that supervises and audits the ethical aspects of the recourse mechanism for ML models. Furthermore, such an entity can take the responsibility of providing a truly model-agnostic set of actions, specifically calibrated for an individual with guarantees of user actionability.

The proposed framework can accommodate action preferential effects due to dynamic macro- and micro-level policy changes. These effects are intuitively captured by user action selections and dynamic calibration dataset. For example, with increasing interest rates, individuals would prefer reducing the loan amount to getting a co-borrower. Such action selection would be reflected in the proposed approach, but not with a fixed policy of low-cost recourse.

Concluding remarks. Our study provides an essential toolkit for providing action sets with guarantees of high probability of action acceptance. This system is essential in a black-box setting of an ML model, where obtaining high quality model information or individual preferences is challenging. Even with lack of such information, our novel, yet intuitive framework can assist in providing high quality action sets with additional independent audit benefits.

The action sets provided can also be used to understand the difficulty of identifying an action for an individual. An individual closer to a decision boundary will have a larger $\mathcal{R}$ compared to the individual farthest from the decision boundary. With this study, we bring a new perspective to the generation and auditability of recourse actions for ML models.

With this study, we introduced the necessity for diverse reformulations for a N&L query in the e-commerce domain. The diversity between reformulations is captured by considering the targets from multiple hops within a user session. This can be extended using other domain-specific diversities between the reformulations. In the e-commerce domain, one can also consider targets from different categories as a diversity signal.

Our work intends to motivate further study into exploring and capturing user diversity behavior for multifaceted reformulation for bad-performing user queries. We show that counterfactual literature and N&L query reformulation shares conceptual properties, and our study motivates further research to bring these fields closer.

Bibliography

[1] Ifeoma Ajunwa, Sorelle Friedler, Carlos E Scheidegger, and Suresh Venkatasubramanian. Hiring by algorithm: predicting and preventing disparate impact. *Available at SSRN*, 2016.

[2] Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

[3] Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

[4] Anastasios N Angelopoulos, Karl Krauth, Stephen Bates, Yixin Wang, and Michael I Jordan. Recommendation systems with distribution-free reliability guarantees. In *Conformal and Probabilistic Prediction with Applications*, pages 175–193. PMLR, 2023.

[5] Anastasios Nikolas Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021.

[6] Julia Angwin, Lauren Kirchner, Jeff Larson, and Surya Mattu. Machine bias: There's

software used across the country to predict future criminals. and it's biased against blacks. 2016.

[7] Javier Antoran, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. Getting a {clue}: A method for explaining uncertainty estimates. In *International Conference on Learning Representations*, 2021.

[8] Ahmed H. Awadallah, Xiaolin Shi, Nick Craswell, and Bill Ramsey. Beyond clicks: Query reformulation as a predictor of search satisfaction. In *ACM International Conference on Information and Knowledge Management (CIKM)*, October 2013.

[9] Kevin Bache and Moshe Lichman. UCI machine learning repository, 2013.

[10] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[11] Jun-Wei Bao, De-Qvan Zheng, Bing Xu, and Tie-Jun Zhao. Query rewriting using statistical machine translation. In *2013 International Conference on Machine Learning and Cybernetics*, volume 2, pages 814–819. IEEE, 2013.

[12] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision*, pages 2722–2730, 2015.

[13] Furui Cheng, Yao Ming, and Huamin Qu. Dece: Decision explorer with counterfactual

explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1438–1447, 2020.

[14] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148. PMLR, 2018.

[15] Zhicheng Cui, Wenlin Chen, Yujie He, and Yixin Chen. Optimal action extraction for random forests and boosted trees. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 179–188, 2015.

[16] Giovanni De Toni, Paolo Viappiani, Bruno Lepri, and Andrea Passerini. Generating personalized counterfactual interventions for algorithmic recourse by eliciting user preferences. *arXiv preprint arXiv:2205.13743*, 2022.

[17] Chris DeBrusk. The risk of machine-learning bias (and how to prevent it). *MIT Sloan Management Review*, 2018.

[18] Tiffany Ding, Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Ryan J. Tibshirani. Class-conditional conformal prediction with many classes, 2023.

[19] Michael Downs, Jonathan L Chu, Yaniv Yacoby, Finale Doshi-Velez, and Weiwei Pan. Cruds: Counterfactual recourse using disentangled subspaces. *ICML WHI*, 2020:1–23, 2020.

[20] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[21] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[22] Lorelle L Espinosa, Jonathan M Turk, Morgan Taylor, and Hollie M Chessman. Race and ethnicity in higher education: A status report. 2019.

[23] Timo Freiesleben. The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines*, 32(1):77–109, 2022.

[24] Sreenivas Gollapudi, Samuel Ieong, and Anitha Kannan. Structured query reformulations in commerce search. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, page 1890–1894, New York, NY, USA, 2012. Association for Computing Machinery.

[25] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[26] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[27] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing recourse across groups. *ArXiv*, abs/1909.03166, 2019.

[28] Eric J Gustafson and George R Parker. Using an index of habitat patch proximity for landscape design. *Landscape and urban planning*, 29(2-3):117–130, 1994.

[29] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[30] Sharon Hirsch, Ido Guy, Alexander Nus, Arnon Dagan, and Oren Kurland. Query reformulation in e-commerce search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1319–1328, New York, NY, USA, 2020. Association for Computing Machinery.

[31] Chien-Kang Huang, Lee-Feng Chien, and Yen-Jen Oyang. Relevant term suggestion in interactive web search based on contextual information in query session logs. *J. Am. Soc. Inf. Sci. Technol.*, 54(7):638–649, may 2003.

[32] Jeff Huang and Efthimis N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. CIKM '09, page 77–86, New York, NY, USA, 2009. Association for Computing Machinery.

[33] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. Patterns of query reformulation during web searching. *J. Am. Soc. Inf. Sci. Technol.*, 60(7):1358–1371, jul 2009.

[34] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *Safe Machine Learning workshop at ICLR*, 2019.

[35] Vassilis Kaffes, Dimitris Sacharidis, and Giorgos Giannopoulos. Model-agnostic counterfactual explanations of recommendations. In *Proceedings of the 29th ACM conference on user modeling, adaptation and personalization*, pages 280–285, 2021.

[36] Kaggle. Give me some credit, 2011.

[37] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*, 2020.

[38] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys (CSUR)*, 2021.

[39] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Comput. Surv.*, 55(5), 2022.

[40] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 353–362, New York, NY, USA, 2021.

[41] Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems*, 33:265–277, 2020.

[42] Davinder Kaur, Suleyman Uslu, and Arjan Durresi. Requirements for trustworthy artificial intelligence–a review. In *International Conference on Network-Based Information Systems*, pages 105–115. Springer, 2020.

[43] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012.

[44] Aleksander Kołcz and Choon Hui Teo. Feature weighting for improved classifier robustness. In *CEAS'09: sixth conference on email and anti-spam.* Citeseer, 2009.

[45] Avni Kothari, Bogdan Kulynych, Tsui-Wei Weng, and Berk Ustun. Prediction without preclusion: Recourse verification with reachable sets. *arXiv preprint arXiv:2308.12820*, 2023.

[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[47] Giridhar Kumaran and Vitor R. Carvalho. Reducing long queries using query quality predictors. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 564–571, New York, NY, USA, 2009. Association for Computing Machinery.

[48] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, X. Renard, and Marcin

Detyniecki. Inverse classification for comparison-based interpretability in machine learning. *ArXiv*, abs/1712.08443, 2017.

[49] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Inverse classification for comparison-based interpretability in machine learning. *stat*, 1050:22, 2017.

[50] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015.

[51] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 349–358, New York, NY, USA, 2019. Association for Computing Machinery.

[52] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

[53] Saurav Manchanda, Mohit Sharma, and George Karypis. Intent term weighting in e-commerce queries. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2345–2348, New York, NY, USA, 2019. Association for Computing Machinery.

[54] Aritra Mandal, Ishita K Khan, and Prathyusha Senthil Kumar. Query rewriting using automatic synonym extraction for e-commerce search. In *eCOM@ SIGIR*, 2019.

[55] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021.

[56] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

[57] Graham Neubig. Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619*, 2017.

[58] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[59] Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 4574–4594. PMLR, 2022.

[60] Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. CARLA: A python library to benchmark algorithmic recourse and

counterfactual explanation algorithms. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021*, 2021.

[61] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pages 3126–3132, 2020.

[62] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, WWW '20, page 3126–3132, New York, NY, USA, 2020. Association for Computing Machinery.

[63] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

[64] Aleksandra Przegalinska. State of the art and future of artificial intelligence. *Policy Department for Economic, Scientific and Quality of Life Policies*, 2019.

[65] Yiming Qiu, Kang Zhang, Han Zhang, Songlin Wang, Sulong Xu, Yun Xiao, Bo Long, and Wen-Yun Yang. Query rewriting via cycle-consistent translation for e-commerce search. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 2435–2446, 2021.

[66] Kaivalya Rawal and Himabindu Lakkaraju. Beyond individualized recourse: In-

terpretable and interactive summaries of actionable recourses. *Advances in Neural Information Processing Systems*, 33:12187–12198, 2020.

[67] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[68] Stefan Riezler and Yi Liu. Query rewriting using monolingual statistical machine translation. *Computational Linguistics*, 36(3):569–582, 2010.

[69] Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 20–28, 2019.

[70] J Ben Schafer, Joseph A Konstan, and John Riedl. E-commerce recommendation applications. *Data mining and knowledge discovery*, 5:115–153, 2001.

[71] Leslie Scism. New york insurers can evaluate your social media use - if they can prove why it's needed, 2019. [Online; accessed January-2019].

[72] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9:371–421, jun 2008.

[73] Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1683–1692, 2016.

[74] Ravi Shroff. Predictive analytics for city agencies: Lessons from children's services. *Big Data*, 5(3):189–196, 2017. PMID: 28829624.

[75] Naeem Siddiqi. *Credit risk scorecards: developing and implementing intelligent credit scoring*, volume 3. John Wiley & Sons, 2012.

[76] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[77] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[78] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[79] Zehong Tan, Canran Xu, Mengjie Jiang, Hua Yang, and Xiaoyuan Wu. Query rewrite for null and low search results in ecommerce. In *eCOM@SIGIR*, 2017.

[80] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. Trustworthy artificial intelligence. *Electronic Markets*, 31(2):447–464, 2021.

[81] Andrew Trotman, Jon Degenhardt, and Surya Kallumadi. The architecture of ebay search. In *eCOM@SIGIR*, 2017.

[82] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear

classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

[83] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 10–19, New York, USA, 2019. Association for Computing Machinery.

[84] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E Hines, John P Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

[85] Julius Von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9584–9594, 2022.

[86] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law and Technology*, 31:841, 2017.

[87] Sandra Wachter, Brent Daniel Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Cybersecurity*, 2017.

[88] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L. Yuille, and Quoc V.

Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[89] Yunpeng Xu, Wenge Guo, and Zhi Wei. Conformal risk control for ordinal classification. In Robin J. Evans and Ilya Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2346–2355. PMLR, 31 Jul–04 Aug 2023.

[90] Prateek Yadav, Peter Hase, and Mohit Bansal. Low-cost algorithmic recourse for users with uncertain cost functions. *arXiv preprint arXiv:2111.01235*, 2021.

[91] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480, 2009.

[92] Jayanth Yetukuri. Individual and group-level considerations of actionable recourse. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 1008–1009, New York, NY, USA, 2023. Association for Computing Machinery.

[93] Jayanth Yetukuri, Ian Hardy, and Yang Liu. Towards user guided actionable recourse. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 742–751, New York, NY, USA, 2023. Association for Computing Machinery.

[94] Jayanth Yetukuri, Ian Hardy, Yevgeniy Vorobeychik, Berk Ustun, and Yang Liu.

Providing fair recourse over plausible groups. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21753–21760, Mar. 2024.

[95] Jayanth Yetukuri and Yang Liu. Robust stochastic bandit algorithms to defend against oracle attack using sample dropout. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5845–5854, 2022.

[96] Mary Frances Zeager, Aksheetha Sridhar, Nathan Fogal, Stephen Adams, Donald E. Brown, and Peter A. Beling. Adversarial learning in credit card fraud detection. In *2017 Systems and Information Engineering Design Symposium (SIEDS)*, pages 112–116, 2017.