Quarter 2 Report Checkpoint

Lina Battikha
lbattikha@ucsd.edu

Sai Poornasree Balamurugan
s2balamurugan@ucsd.edu

Abstract

Our work this quarter focuses on latent variable discovery of fair labels. This fairness algorithm and its accuracy will be evaluated on a variety of datasets, including common fairness datasets from UCI Adults, German Credit Scores, and Compas (recidivism data). The objective of this algorithm is to reduce unfair biases that are affected by sensitive features within the data. These biases can significantly impact predictive models, leading to unfair outcomes that disproportionately affect certain groups. Since these models often rely on sensitive attributes rather than unbiased patterns, ensuring fairness in predictions remains a critical challenge.

Website: https://s2balamurugan.github.io/fairness-application-website/
Code: https://github.com/linabat/fairness-application

1 Introduction

Bias in artificial intelligence is an extremely important problem to be explored and addressed as the results of such models can greatly impact an individual's life, especially those who are of minority groups. One common reason that bias arises is due to sensitive features that are included in the dataset, like gender or race. The inclusion of such sensitive features can lead the observed outcome label to pick up on unfair patterns and dependencies that will lead such labels to be unfair. To address such problems, our project focuses on discovery of fair latent variables. Latent variable discovery involves the identification and characterization of hidden variables that influence the distribution and outcomes of observed data. These variables are not explicitly provided in the dataset but can be inferred by analyzing patterns and relationships among the observed features. Our work has been focused on building a model for latent variable discovery for fairness applicability. The latent variable that we are looking to discover is the true unbiased outcome label that isn't influenced by a sensitive feature within the dataset.

The importance of latent variable discovery is evident across three key application domains:

- Clustering: Latent variables help identify natural groupings in data, enabling meaningful segmentation even when group boundaries are not explicitly labeled. This is especially valuable in domains like personalized medicine, where clustering can reveal patient subtypes for tailored treatments
- Distribution shift: In real-world scenarios, data often deviates from the training distribution. Latent variable models facilitate out-of-distribution robustness by capturing underlying structures that remain constant, thus improving model performance on unseen data.
- Fairness: Addressing bias in decision-making systems requires understanding how latent variables interact with observable features. By accounting for unobserved confounders, these models can enhance fairness, ensuring equitable outcomes for all subpopulations.

For our project, we focuses on the third key application: fairness. Specifically, we use tabular datasets where we identify features of the data, the senstive feature that can be resulting in on fair observed labels, and the observed label itself. Discovery of fair latent variables aim to promote fair results for all groups that are a part of sensitive features.

1.1 Literature Review

Our work is inspired by the methodology outlined in Group Fairness by Probabilistic Modeling with Latent Fair Decisions by YooJung Choi, Meihua Dang, and Guy Van den Broeck Choi, Dang and den Broeck (2020) as well as Scalable Out-of-distribution Robustness in the Presence of Unobserved Confounders by Parjanya Prashant, Seyedeh Baharan Khatami, Bruno Ribeiro, and Babak Salimi Prashant et al. (2024).

The Scalable Out-of-distribution Robustness in the Presence of Unobserved Confounders paper addresses the challenge of training machine learning models when there are unob-

served variables, known as latent confounders, which can affect the relationship between the covariates (input features) X and the target variable Y. These unobserved confounding variable Z can be a cause for distribution shift of $P(Y|X)$. To handle this, the paper proposes using a proxy variable S, which is a known variable that helps approximate the latent confounder Z. The proxy variable is crucial in managing sub-population shifts, where the distribution of subgroups in the data changes between training and testing. By using the proxy variable S, the model can infer the latent confounder Z, which is unobserved during both training and testing. In this setup, the model has access to all covariate input data X and the target variable Y during training, but the latent confounder Z is not directly observed. One of the stages proposed in the methods of this paper is the Encoder-Decoder model. In the first stage, the Encoder-Decoder model is used to estimate the latent confounder distribution by factoring the joint distribution of the input features X and the proxy variable S. The encoder part of the model learns how the input features X are related to the latent confounder Z, while the decoder models the relationship between S and Z. This factorization allows the model to infer the distribution of the latent confounder, $P(Z|X)$, given the input data.

The Group Fairness by Probabilistic Modeling with Latent Fair Decisions paper also focuses on latent variable discovery, but specifically in regards to group fairness. Given a sensitive feature and an observed variable, the authors suggest a methodology where demographic parity can be established with a latent fair label. Demographic parity is where each group within the sensitive feature is experiencing similar label classification distribution; $P(D_f|S = 0) = P(D_f|S = 1)$.The authors introduce features X, sensitive feature S, observed label D and a latent fair variable, D_f , representing true, unbiased decisions. Knowing such variables, it is possible to achieve demographic parity by enforcing certain independencies. An independency that is enforced is that the latent variable D_f is independent of S. With that, the fair label should not have dependencies on the sensitive feature, which allows for demographic parity.

1.2 Datasets Implemented

Our first dataset is the UCI Adult dataset, a popular tool for determining if a person makes more than $50,000 annually and is obtained from the U.S. Census Bureau. It includes information about age, education, occupation, work class, marital status, and weekly hours worked, among other demographic and employment variables. The dataset, which represents actual income distribution inequalities, is often utilized in fairness and prejudice research. Due to its structured nature, it serves as a benchmark for classification models in machine learning.

The UCI Machine Learning Repository provides the German Credit dataset, which is used to evaluate credit risk by categorizing people as ”good” or ”bad” credit risks. It contains information on an applicant’s home situation, work status, financial history, and personal characteristics such as age and sex. The data set is frequently used to examine bias and fairness in financial decision making because of its importance in lending choices, which

helps to assess moral issues with credit scoring models.

Criminal justice data used to determine the probability of recidivism (reoffending) based on demographic and criminal history criteria may be found in the COMPAS dataset, which is gathered by ProPublica. Features including age, sex, race, past crimes, and COMPAS risk scores are among them. Studies have revealed that the COMPAS risk assessment tool includes racial discrepancies, making it a crucial resource for assessing ethical difficulties in judicial decision-making and predictive policing. As a result, this dataset has been at the center of arguments on algorithmic fairness and bias.

2 Methods

Our approach uses a similar encoder-decoder methodology as used in the Scalable Out-of-distribution Robustness in the Presence of Unobserved Confounders Prashant et al. (2024) while establishing independence between the sensitive feature and fair observed label as mentioned in the Group Fairness by Probabilistic Modeling with Latent Fair Decisions Choi, Dang and den Broeck (2020) paper. Given a data set that has an identifiable sensitive feature S, an observed label Y, and non-confounding features X, we treat Y as the proxy variable that will be used in the model to find the fair label Y_f.

Our model will use an encoder-decoder model with adversarial biasing used as a regularizer in it so that the the independence condition between S and Y_f is established. Adversarial biasing is a machine learning concept used to reduce bias in predictive models by implementing an adversarial network. The main idea is to train two competing networks: a predictor that learns to classify data while the adversary tries to identify sensitive attributes (e.g., race, gender). The predictor is trained to minimize classification loss while simultaneously manipulating the discriminator. This reduces the correlation between predictions and sensitive attributes. This makes sure that predictions are both accurate and fair, preventing models from making biased decisions based on protected attributes.

To do this, the model can broken down into 3 main sections: Encoder, Adversarial, and Decoder.

1. **Encoder:** Forward pass of retrieving $pseudo\_Y$, given X. $pseudo\_Y$ will be the intermediate label that will be used to establish independence with S.
2. **Adversarial Branch:** This branch will work to establish the independence between $pseudo\_Y$ and S. $pseudo\_Y$, will then be used to predict S. If the adversary can predict S from $pseudo\_Y$, then that is indication that independence has not yet been established yet. To discourage such dependence from occurring, the encoder will need to remove dependency on S from $pseudo\_Y$. This is done through the Gradient Reversal Layer (GRL), which act as a regularizer to flip the gradient so that the encoder doesn't encode S from psuedo Y. The adversary will work to minimize the loss between S_pred and S from $pseudo\_Y$. However, GRL will flip the gradient of this loss, leading the encoder to maximize it. As a result, the encoder removes dependencies between $psuedo\_Y$ and S, leading $psuedo\_Y$ to be independent of

S. This adversarial training process happens during the back-propagation between the encoder and adversary. GRL will minimize accuracy of adversary by causing $psuedo\_Y$ to lose info about S. The encoder would be working against the adversary, and as a result, the adversary will fail if $psuedo\_Y$ is independent of S, which is the desired outcome.

3. **Decoder:** The decoder will try to predict Y given S and $pseudo\_Y$. While we are trying to establish independence between Y and S by removing unfair dependencies of S on Y, some fair dependencies can still exist. For this reason, $psuedo\_Y$ and S are being used to predict Y so that the output includes accurate and fair results.

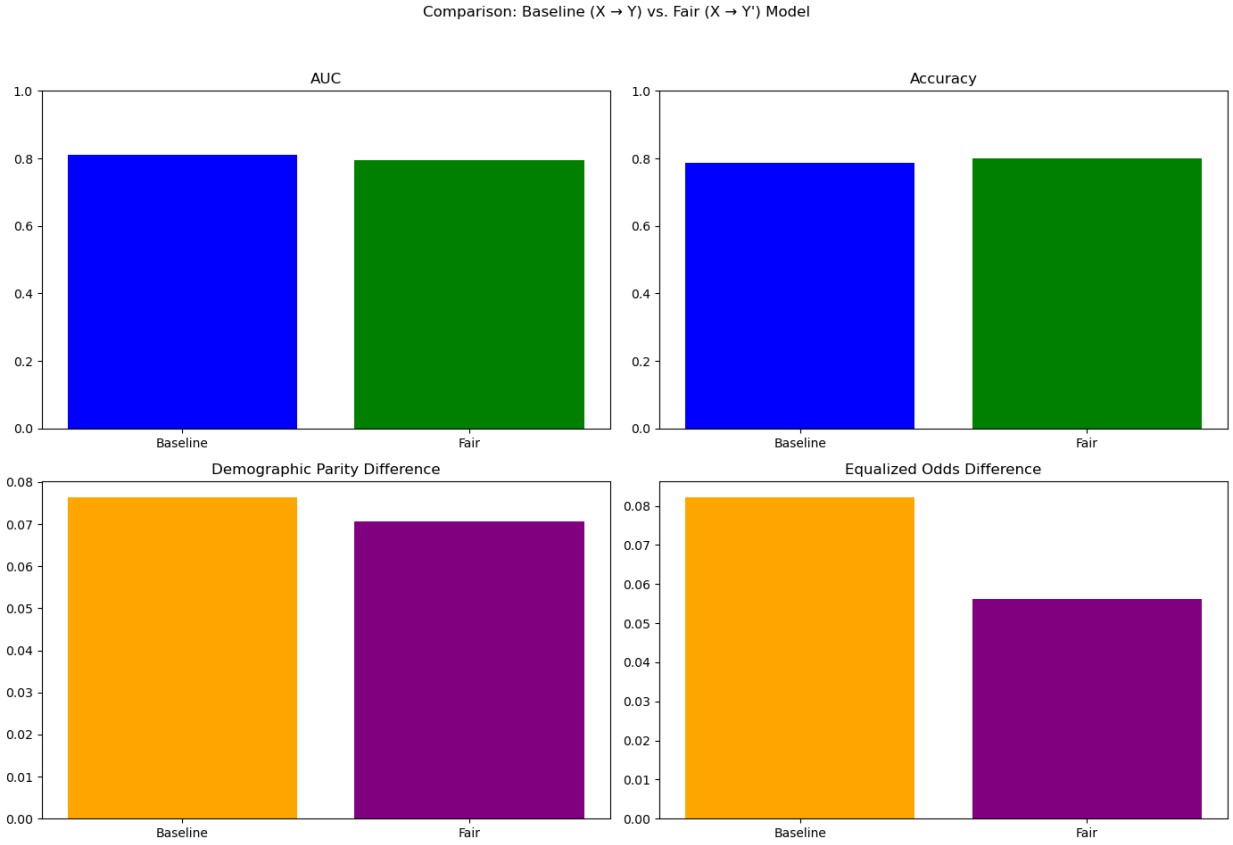
The loss functions that are being used are binary cross-entropy and categorical cross-entropy. Binary cross-entropy measures the difference between two probability distributions: the predicted probability distribution and the true distribution. It quantifies how well the predicted probabilities align with the actual class labels by computing the negative log-likelihood of the true class. It also penalizes incorrect predictions more harshly when the predicted probability is far from the true label. In binary classification, cross-entropy is used with the sigmoid activation function, while in multiclass classification, it is often combined with softmax. Lower cross-entropy values indicate better model predictions.

A loss function used in multiclass classification problems, where each sample is a member of multiple classes, is categorical cross-entropy. By calculating the negative log-likelihood of the proper class, it assesses how closely the projected probability distribution matches the actual class labels. The loss for a single sample is computed given a predicted probability distribution y-hat and a one-hot encoded true label y. Since the model gives the right class a higher probability, lower numbers correspond to better predictions.

In detail our model optimizes three different loss functions. The loss ensures that the pseudo-label Y closely matches the true label Y. This classification loss helps the model learn meaningful feature representations. Using the pseudo-label Y, the adversarial branch aims to forecast the sensitive characteristic S. Because the gradient reversal layer makes sure that this loss function optimizes the error in forecasting S, bias is reduced because Y must be independent of S.

3 Results - Binary Label

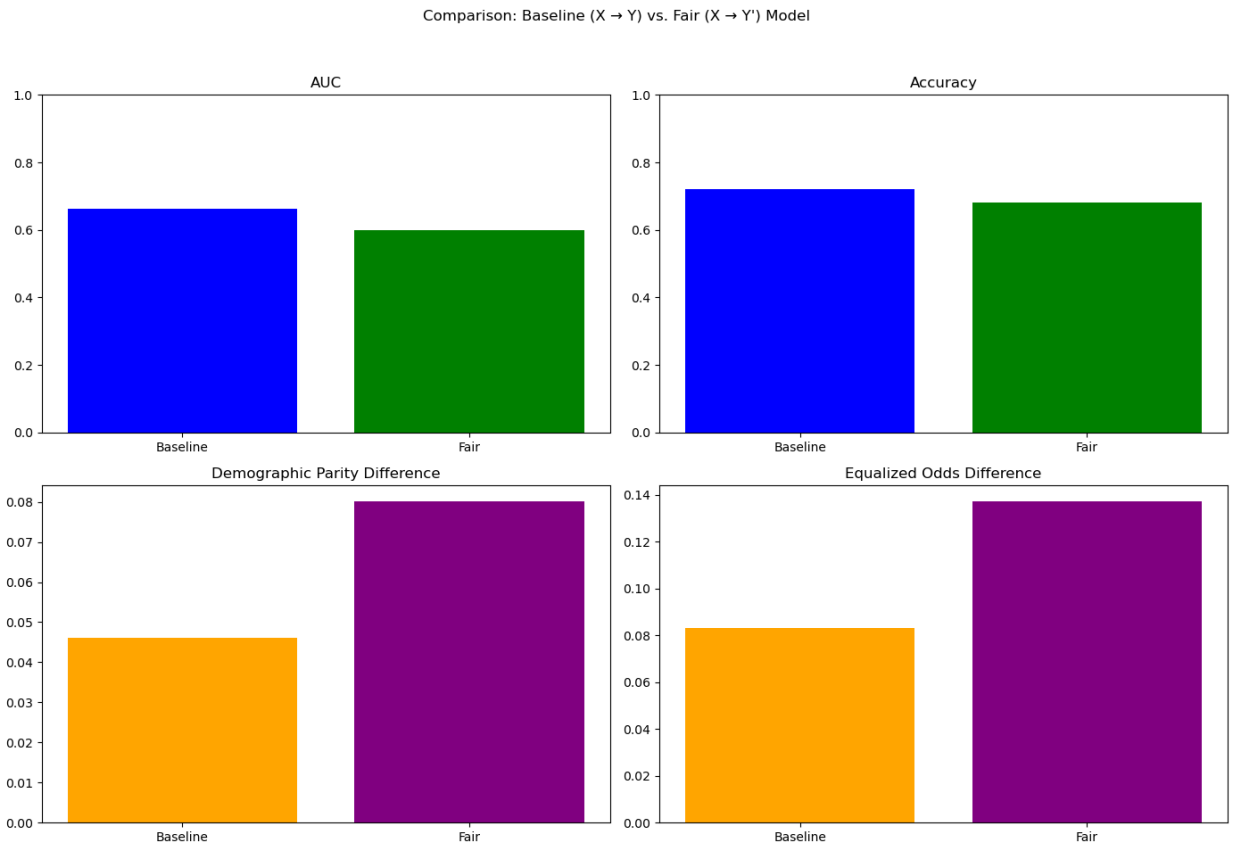
UCI Adults



Baseline Results:
AUC: 0.8095, Accuracy: 0.7877
Fairness metrics:
Demographic Parity Difference: 0.0764
Equalized Odds Difference: 0.0822
Selection Rate: Class 0: 0.07996, Class 1: 0.15637
Group Accuracy: Class 0: 0.8716, Class 1: 0.7470

Fair Model Results:
AUC: 0.7952, Accuracy: 0.7989
Fairness metrics:
Demographic Parity Difference: 0.0706
Equalized Odds Difference: 0.0561
Selection Rate: Class 0: 0.1068, Class 1: 0.1773
Group Accuracy: Class 0: 0.8617, Class 1: 0.7684

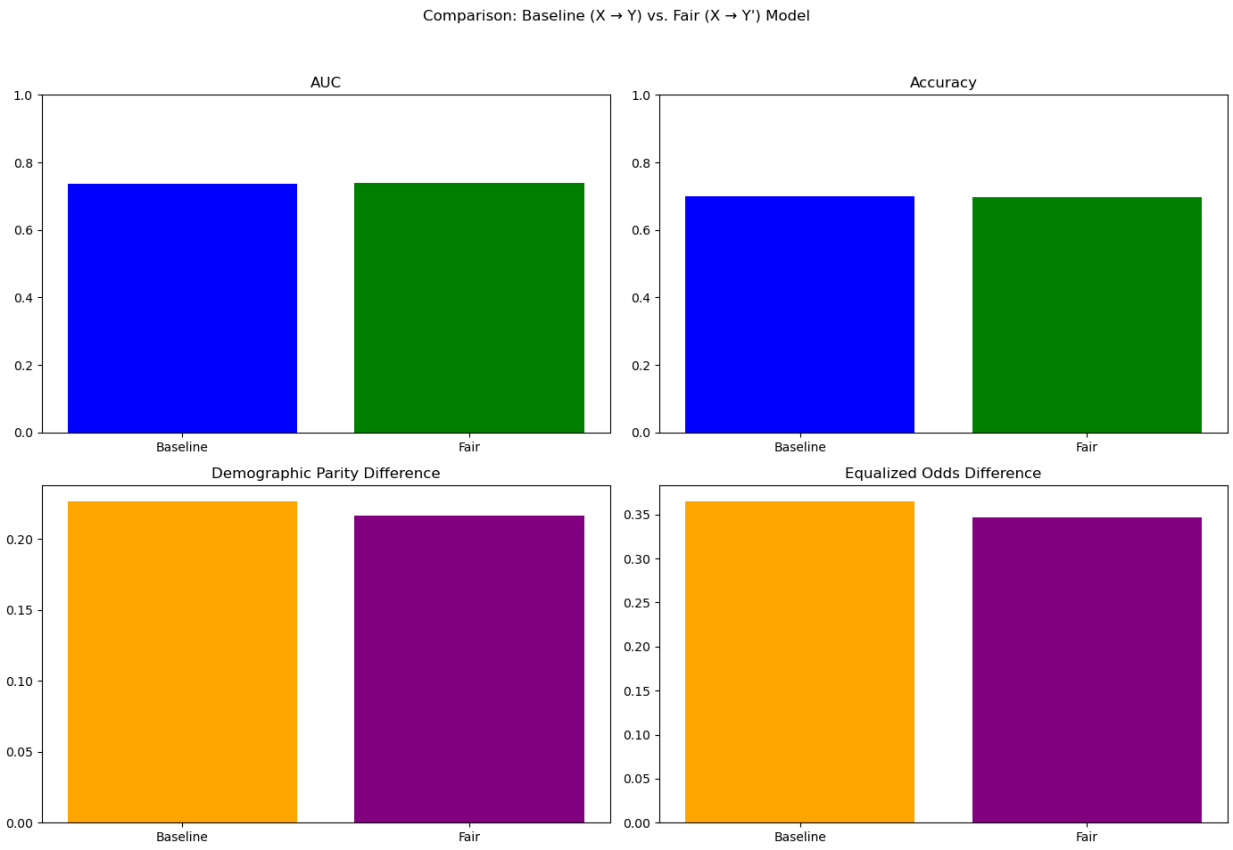
German Credit Scores



Baseline Results:
AUC: 0.6631, Accuracy: 0.7200
Fairness metrics:
Demographic Parity Difference: 0.0461
Equalized Odds Difference: 0.0833
Selection Rate: Class 0: 0.9588, Class 1: 0.9126
Group Accuracy: Class 0: 0.7526, Class 1: 0.6893

Fair Model Results:
AUC: 0.5983, Accuracy: 0.6800
Fairness metrics:
Demographic Parity Difference: 0.0802
Equalized Odds Difference: 0.1374
Selection Rate: Class 0: 0.8763, Class 1: 0.7961
Group Accuracy: Class 0: 0.7113, Class 1: 0.6505

COMPAS



Baseline Results:
AUC: 0.7372, Accuracy: 0.6992
Fairness metrics:
Demographic Parity Difference: 0.2266
Equalized Odds Difference: 0.3651
Selection Rate: Class 0: 0.2289, Class 1: 0.4555
Group Accuracy: Class 0: 0.7037, Class 1: 0.6949

Fair Model Results:
AUC: 0.7385, Accuracy: 0.6972
Fairness metrics:
Demographic Parity Difference: 0.2164
Equalized Odds Difference: 0.3463
Selection Rate: Class 0: 0.2528, Class 1: 0.4692
Group Accuracy: Class 0: 0.6938, Class 1: 0.7004

4 Appendix

4.1 Quarter 2 Report Proposal

Link: https://www.overleaf.com/read/ftkrvdmkgdng#b5f66b

4.2 Future Works

In order to ensure fairness across numerous groups, rather than only reducing bias between two categories, multiclass classification is essential and will be implemented on the observed label. A binary model might miss differences across subgroups since many real-world datasets contain attributes that have more than two observed label categories. By ensuring that bias mitigation is applicable to all subgroups and not just one particular binary split, this method makes the learned representation genuinely fair and invariant.

Additionally, we will be putting this strategy into practice by assessing the fairness of models forecasting health-related outcomes using the Drug Consumption and Hospital Readmission datasets. These datasets are evaluating how well adversarial training reduces bias across a variety of demographic groups because they include multiple labels for the observed feature.

5 Contributions

Lina Battikha:

- Abstract, Introduction, Literature Review, Methods, Results of report
- Cleaned up code + README.md for reproducibility
- Worked on changing code for multi-classification
- Found and cleaned Drug Consumption data to be tested for multi-classification

Sai Poornasree Balamurugan:

- Creating and started on website
- Creating and started initial process of cleaning report
- Introduction, Methods, Datasets implemented, and Future works of report
- Working on implementing multi-classification for Health Readmission datatset

References

Choi, YooJung, Meihua Dang, and Guy Van den Broeck. 2020. "Group Fairness by Probabilistic Modeling with Latent Fair Decisions." [Link]

Prashant, Parjanya, Seyedeh Baharan Khatami, Bruno Ribeiro, and Babak Salimi.

2024. “Scalable Out-of-distribution Robustness in the Presence of Unobserved Confounders.” [Link]