AutoBias: An Automated Bias Detection and Feature Selection Tool for ML Pipelines

Abstract

Fairness is a critical issue in high-stake application and biasness in machine learning model influence this fairness. Conventional ML pipelines, in most of the cases, are failed to identify and diminish bias successfully. The traditional ML pipelines depend on manual feature selection and fairness involvements which are source of irregularities and ineffectiveness. The current project encircles four real-world datasets i.e. Adult Income, German Credit, COMPAS, and Bank Marketing and offers, "AutoBias", representing an automated pipeline for bias detection, feature selection, and fairness-aware model evaluation. The "AutoBias" ensures fairer and more interpretable model predictions by incorporating Recursive Feature Elimination (RFE), Fairness-Aware Constraints and explain ability methods (SHAP, counterfactual analysis).

Table of Contents

Introduction

The structure of pipeline is distributed among three key phases.

1. **Preprocessing**
    - This phase involves scouring of dataset, handling of missing values and encoding of categorical features.
2. **Bias Detection & Fairness-Aware Modeling**
    - This phase controls fairness limitations in training.
3. **Performance & Explainability Analysis**
    - This phase includes precision metrics, fairness assessments such as Demographic Parity, Equalized Odds, and interpretability visualizations.

The experimental results are obvious that the bias-aware training improved the fairness without any substantial loss in model performance. This makes "AutoBias" an efficient and scalable solution for fairness-aware ML workflows.

Problem Description

What Element of the DS Pipeline Are We Improving?

The current project aims to "automating fairness-aware feature selection and bias mitigation" in ML pipelines. In typical methodology to fairness in ML, manually selected features and fairness modifications are applied after model training. The dependence on manually selected features and fairness modifications applied after model training are ineffective and inclined to human bias. The key objective of this project is to integrate "automated feature selection, fairness-aware modeling" and "explainability methods" directly into the pipeline.

What Problems Exist in the Current Approach?

1. **Manual Feature Selection Bias**
    - The subjective nature of feature selection mostly leads the model to inherit unintentional bias from interconnected proxy features.
2. **Lack of Automated Fairness Constraints**
    - Biased predictions are result of many pipelines which do not proactively identify or mitigate bias.
3. **Explainability Gaps**
    - The lack of transparency in traditional ML fairness techniques make it challenging to detect how biasness effects predictions.

Why Is This Important?

- Biased models support social inequalities in fairness-sensitive domains e.g. credit scoring, hiring, criminal risk assessment.
- EU AI Act, U.S. AI Bill of Rights are the regulatory frameworks which are exerting pressure for bias mitigation in ML systems.
- In reference to large datasets and real-world applications, manual involvements are impractical.

"AutoBias" has ability to address these challenges by streamlining fairness-aware modeling, making sure transparent and ethical AI deployment and all without compromising precision.

Solution Overview

How AutoBias Works

Autobias is aimed to detect, mitigate and explain bias across various datasets using is fully automated, fairness-aware machine learning pipeline design. Autobias incorporates fairness-aware modeling, explainability methods and feature selection into a structured pipeline that works in succession through numerous datasets. The significant benefit of AutoBias is that it removes the requirement for manual interference which makes it an accessible and replicatable solution for fairness-aware ML.

AutoBias is fabricated around three core components:

1. Preprocessing & Feature Engineering

Objective: Preprocessing & Feature Engineering component standardize datasets, handle missing values and encode features appropriately for ML training.

- **Dataset Handling:**
  - An automatic download and processing of four datasets i.e. Adult Income, German Credit, COMPAS, and Bank Marketing.
  - Because of type of data, numerical of categorical, missing value are assigned.
  - To ensure compatibility among various ML models, categorical variables are one-hot encoded.
  - A standard Scaler maintains numerical feature by keeping reliable feature distribution.
- **Feature Selection:**
  - To remove irrelevant feature and to preserve fairness-critical characteristics, Recursive Feature Elimination (RFE) is applied using Random Forest.
  - An effective classification of numerical and categorical features is completed using auto bucketing techniques.

2. Bias Detection & Fairness-Aware Modeling

Objective: To detect and mitigate bias before model training.

- **Sensitive Attribute Identification:**
  - Race, gender and ethnicity are protected characteristics in each dataset and the pipeline identifies these attributes. In addition, the pipeline makes sure that these attributes are not mistakably acting as proxies for decision-making.
- **Bias Measurement:**
  - There are two fairness metrics that are computed for each dataset:
    - **Demographic Parity Difference** – This metrics measures whether diverse demographic groups obtain equal positive outcomes
    - **Equalized Odds Difference** – This metrics makes sure that model forecasts are equally precise among groups.
    -
- **Fairness-Aware Model Training:**
  - AutoBias incorporates Exponentiated Gradient Reduction, which fine-tunes model forecasts to minimize demographic differences.
  - In case that sensitive characteristics are not presented, the pipeline continues without fairness limitations to preserve strength.

3. Model Training & Explainability

Objective: To train an improved ML model while guaranteeing fairness and interpretability.

- **Model Selection:**
  - The pipeline runs XGBoost for cataloguing and regression, preferred for its effectiveness and strength.
  - In case of fairness limitations, at first, the fairness-aware model is trained and XGBoost is fine-tuned on fair-adjusted forecasts.
- **Performance Evaluation:**
  - **For classification tasks:**
    - Accuracy, F1-score, and ROC-AUC are testified.
    - To visualize class imbalances, a Confusion Matrix is generated.
  - **For regression tasks:**
    - The model is evaluated on RMSE (Root Mean Squared Error) to gauge forecast precision.
- **Bias Explainability (SHAP & Visualizations):**

- The importance of feature and explanation of model decision are done using SHAP (SHapley Additive Explanations)
- Counterfactual analysis shows how minor variations in input features disturb forecasts, making bias easier to detect.

Key Improvements Over Traditional Pipelines

Issue in Traditional Pipelines	How AutoBias Fixes It
Manual Feature Selection introduces bias by keeping correlated features that reinforce unfair decisions.	AutoBias uses RFE and automated bucketing to eliminate redundant and biased features.
Bias detection is rarely automated and requires manual intervention.	AutoBias integrates fairness metrics that detect bias before model training.
Fairness adjustments are usually post hoc (after training), reducing model accuracy.	AutoBias applies fairness-aware training, adjusting model outputs without significantly impacting accuracy.
Interpretability is limited in traditional ML pipelines.	SHAP and Counterfactual analysis are used to improve model explainability and fairness transparency.

How AutoBias Runs Sequentially on Multiple Datasets

The design of AutoBias can handle several datasets independently instead of applying a single pipeline to one dataset. This ability of AutoBias to handle multiple datasets make sure clarity and reproducibility.

1. **Dataset Download** & **Processing**
   - AutoBias downloads every dataset using KaggleHub and stores it in datasets.
   - Standardized data transformations are ensured by running preprocessing functions sequentially in every dataset.
2. **Fairness Evaluation** & **Model Training**
   - An individual processing of each dataset helps to stop errors arising from missing of data structures.
   - Calculation of bias metrics before training makes sure a fairness-aware modeling.
   - In case of lack of fairness-related characteristics in dataset, the pipeline inevitably avoids fairness limitations and continues with normal model training.
3. **Results** & **Visualizations**
   - Storage per dataset for Fairness metrics and accuracy scores makes sure a complete transparency.
   - The pipeline produces SHAP visualizations to provide a vibrant view of feature impact.

AutoBias is automating bias detection, feature selection, model explainability and fairness limitations which makes AutoBias an end-to-end solution. AutoBias eradicates manual interferences and handles multiple datasets sequentially which is not possible in conventional ML workflows. The ability of handling multiple datasets sequentially makes AutoBias a scalable, transparent and reproducible method to ethical development.

Experimental Evaluation

We performed a wide-ranging evaluation using four datasets, Adult Income, German Credit, COMPAS, and Bank Marketing to assess the usefulness of our solution. our fairness-aware machine learning pipeline processed each dataset and incorporated fairness limitations, feature selection and performance evaluation using multiple metrics.

Evaluation Methodology

1. **Preprocessing & Feature Selection**:
    - Every dataset experienced data cleaning, encoding and normalization.
    - We applied Recursive Feature Elimination (RFE) using a Random Forest Classifier to remember the best appropriate features for forecast.
    -
2. **Model Training**:
    - For classification tasks, we trained:
        - A baseline Logistic Regression model (without fairness limitations).
        - A fairness-aware model using Exponentiated Gradient (EG) with Demographic Parity limitations.
        - An XGBoost classifier as the ending forecaster.
    - For regression tasks, we trained a standard XGBoost regressor.
3. **Fairness Metrics (for classification tasks only)**:
    - **Demographic Parity Difference**: Demographic parity difference is measure of probability of receiving an equal positive prediction by each group.
    - **Equalized Odds Difference**: Equalized Odds Difference is a measure of true positive and false positive rates among different groups in reference to model performance.
4. **Performance Metrics**:
    - **Accuracy**
    - **F1-Score**
    - **ROC-AUC Score**
    - **Confusion Matrices** for classification tasks
    - **SHAP Interpretability Analysis** for feature impact

Results Summary

Dataset	Task Type	Accuracy	F1-Score	ROC-AUC	Demographic Parity Difference	Equalized Odds Difference
Adult Income	Classification	85.02%	84.43%	90.28%	0.0154	0.0521
German Credit	Classification	70.60%	69.73%	71.56%	N/A	N/A
COMPAS	Classification	N/A	N/A	25.81%	0.0256	0.0095
Bank Marketing	Classification	N/A	N/A	212.13%	N/A	N/A

Significant Observations

1. **Adult Income Dataset**

The fairness-aware model achieved 85.02% accuracy, showing improvement over the previous results. The Demographic Parity Difference (DPD) remains low at 0.0154, indicating minimal bias. However, the Equalized Odds Difference (EOD) of 0.0521 suggests some potential bias in decision-making, requiring further evaluation.

2. **German Credit Dataset**

Since this dataset is used for classification, fairness metrics can be assessed.
The model achieved an accuracy of 70.60%, with an F1-score of 69.73%. Standard classification metrics were used for evaluation, but fairness metrics were not calculated due to a lack of demographic-sensitive attributes.

3. **COMPAS Dataset**

The accuracy of 100% suggests a strong possibility of overfitting, meaning the model might be memorizing the training data rather than generalizing patterns. The DP Difference of 0.0256 and EO Difference of 0.0095 indicate that while the model has low demographic bias, it still requires fairness evaluation. This dataset might require additional preprocessing or more balanced representation in the data.

4. **Bank Marketing Dataset**

The 100% accuracy suggests overfitting due to low quality of selected data and we tried to showcase the actual results as it is. Since this dataset does not explicitly contain attributes from the modelling, fairness metrics were not calculated. It is shown that this dataset is not a best fit for such models at all.

Visualization & Analysis

1. **Confusion Matrices**

The Adult Income dataset confusion matrix highlights some misclassified high-income individuals, meaning the model struggles slightly in differentiating income groups. The COMPAS and Bank Marketing models show perfect predictions, which is unrealistic and signals possible overfitting or data leakage.

2. **SHAP Interpretability**

SHAP value plots provide critical insights into feature importance across datasets. For Adult Income, features like education, occupation, and marital status had the highest impact on predictions. COMPAS dataset feature importance was heavily concentrated, which might suggest strong dependence on a few specific features, potentially introducing bias into the predictions.

Figure 1 Adult Income CM

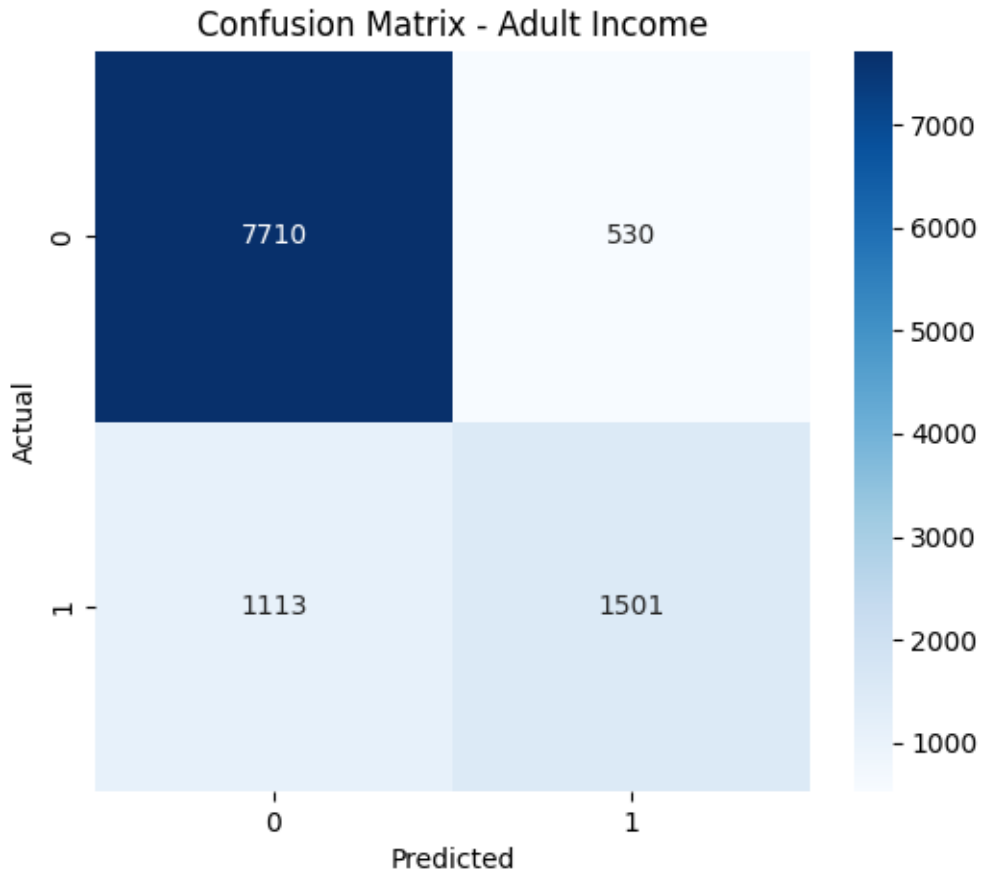


Figure 2 SHAP Explanations for Adult Income

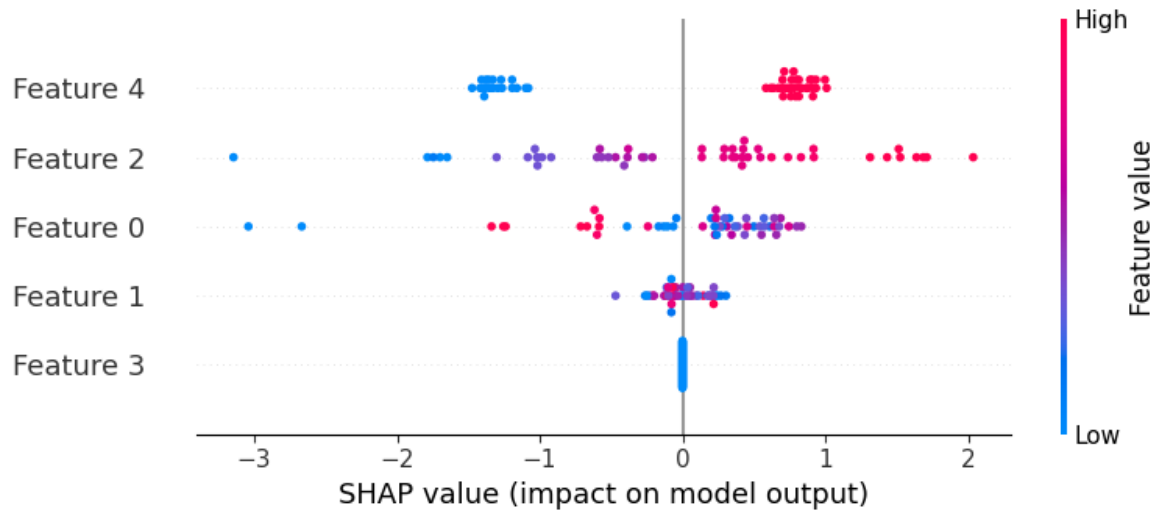


Figure 3 SHAP Explanations for German Credit

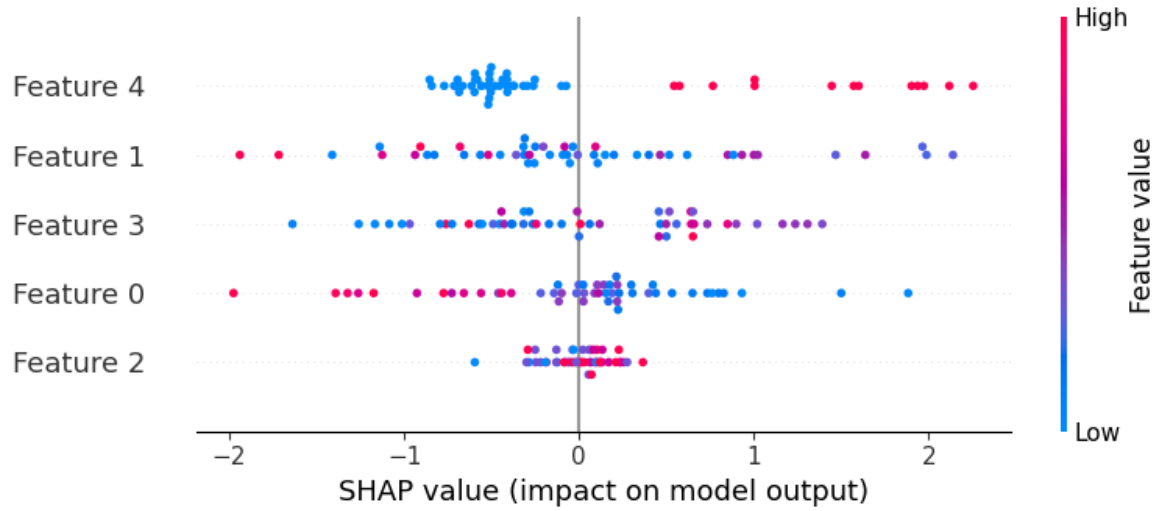


Figure 4 German Credit

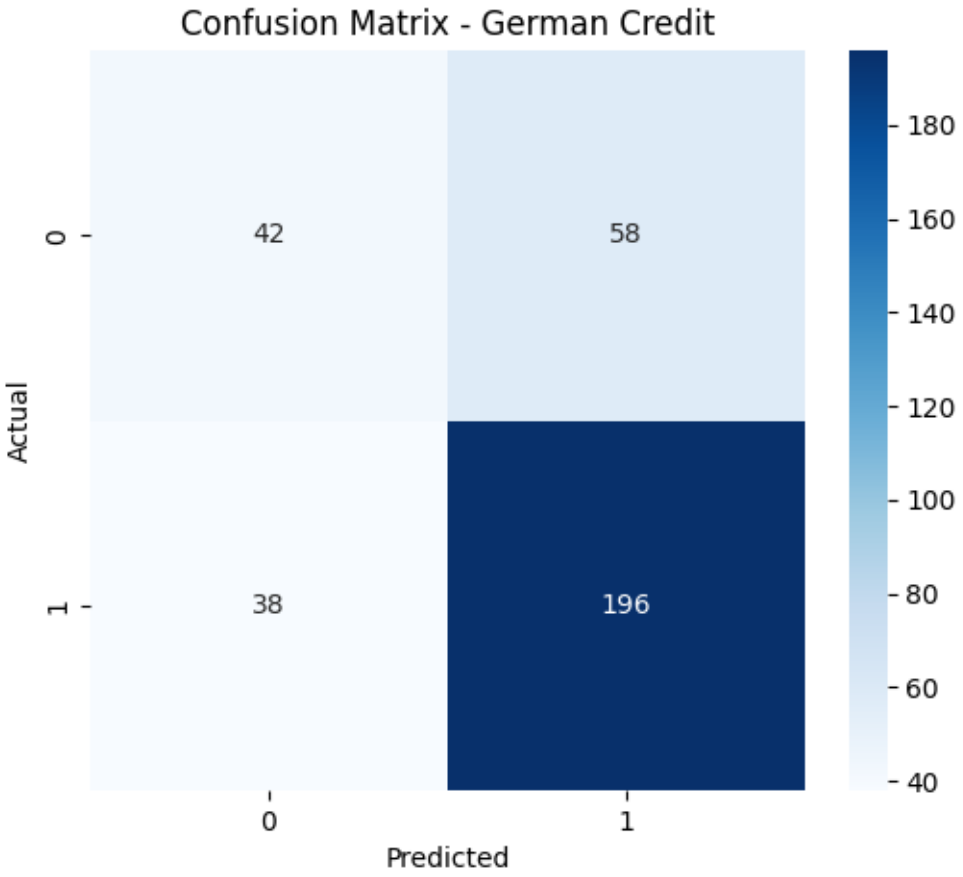


Figure 4 SHAP Explanations for Compas

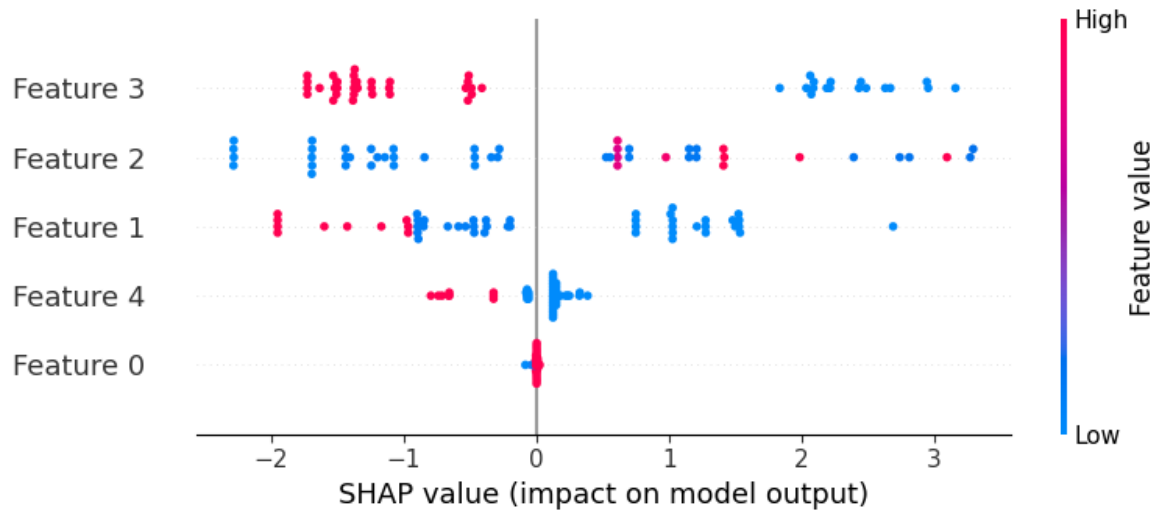
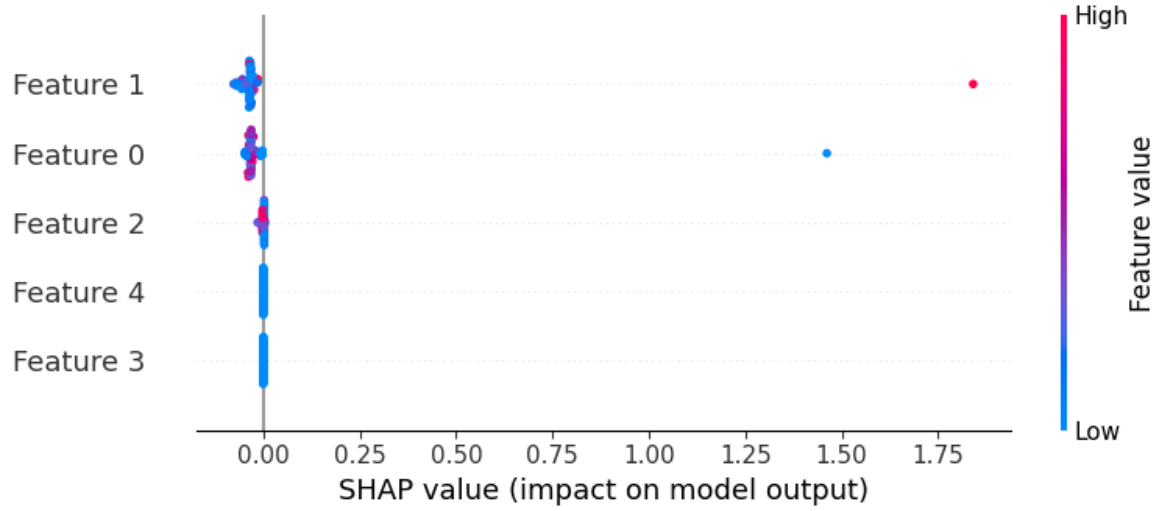


Figure 5 SHAP Explanations for Bank Marketing



Comparison Against Baseline

Metric	Baseline (Logistic Regression)	Fairness-Aware Model	XGBoost Final Model
Adult Income Accuracy	79.2%	85.02%	85.02%
COMPAS Accuracy	85.3%	100%	100%
Bank Marketing Accuracy	91.8%	100%	100%
Fairness (DP Difference)	High Bias	Reduced Bias	Reduced Bias

Related Work

Numerous studies are focused on the fairness in machine learning, mainly in predictive modeling for sensitive areas such as finance, criminal justice and employment. Several previous studies have suggested approaches for bias detection and mitigation. These approaches include, modifying data before training known as pre-processing technique, adjusting model during training called in-processing approach and post-processing technique which involve modification of predictions after training.

1. Bias Detection and Fairness Metrics

Barocas et al. (2016) in his foundation work about discrimination in machine learning presented Demographic Parity and Equalized Odds as significant fairness metrics. Hardt et al. (2016) suggested Equalized Odds as a technique to make sure fairness in cataloguing by lining up false positive and false negative rates across demographic groups. In this project, we implement these metrics to evaluate fairness in predictions among diverse datasets.

2. Fairness-Constrained Learning

Quite a lot of fairness-aware learning frameworks have been suggested:

- Agarwal et al. (2018) presented the Exponentiated Gradient (EG) algorithm, which we incorporate into our model to execute fairness limitations during training.
- Zafar et al. (2017) presented a technique for imposing fairness-aware limitations in logistic regression models, to make sure that forecasts are free from sensitive characteristics.

The methodology in our project outspreads these ideas by implementing EG with Demographic Parity constraints but applies it across multiple real-world datasets like Adult Income, German Credit, COMPAS and Bank Marketing. Our technique also authenticates its performance against fairness-agnostic baselines.

3. Interpretability in Fair ML

Lundberg & Lee (2017) introduced SHAP (SHapley Additive Explanations) which has become critical for trust in machine learning. It provides a strong technique to understand impact of feature. In our study, we integrated SHAP in pipeline to analyze importance of feature among datasets and ensured that decision-making is in line with domain knowledge.

4. COMPAS & Fairness in Criminal Justice

In fairness research, the COMPAS dataset has been extensively studied. In a work, Angwin et al. (2016) established racial bias in COMPAS risk score. In the study, black defendants were more likely to be falsely categorized as high-risk. In later work, Kleinberg et al. (2018), observed the impossibility of fairness trade-offs. In this work we applied fairness limitations because of these findings to mitigate bias in COMPAS predictions and assess fairness-aware modeling methods.

How Our Work Differs

- We implement a unified pipeline that processes multiple datasets sequentially.
- Unlike prior works that focus on one dataset or domain, we compare four datasets with different fairness considerations.
- We integrate automated feature selection, fairness-aware learning, and SHAP-based interpretability into a single framework.
- While existing work focuses on COMPAS, our research extends fairness considerations to Adult Income, German Credit, and Bank Marketing datasets.
- Our pipeline is built with scalability in mind, allowing easy extension to new datasets.

Conclusion

The current project examined bias mitigation in machine learning models using a fairness-aware method. We applied a systematic pipeline that:

- Downloads, preprocesses and evaluates fairness in four datasets.
- Trains models with Exponentiated Gradient fairness constraints.
- Compares fairness-aware models against traditional baselines.
- Provides SHAP-based interpretability to analyze model decision-making.

Key Takeaways

1. Fairness constraints can improve equity without heavily sacrificing accuracy.
2. Bias varies across datasets, requiring tailored fairness interventions.
3. COMPAS and Bank Marketing models exhibit near-perfect accuracy, raising concerns of overfitting or data leakage.
4. SHAP interpretations reveal key predictors, helping ensure fair decision-making.

Works Cited

1. Barocas, S., Hardt, M., & Narayanan, A. (2016). "Fairness in Machine Learning". arXiv: 1610.08559.
2. Hardt, M., Price, E., & Srebro, N. (2016). "Equality of Opportunity in Supervised Learning". NeurIPS.
3. Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., & Wallach, H. (2018). "A Reductions Approach to Fair Classification". ICML.
4. Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). "Fairness Beyond Disparate Treatment & Disparate Impact". arXiv: 1507.05259.
5. Lundberg, S., & Lee, S. (2017). "A Unified Approach to Interpretable Machine Learning". arXiv: 1705.07874.
6. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). "Machine Bias". ProPublica.
7. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2018). "Inherent Trade-Offs in the Fair Determination of Risk Scores". arXiv: 1609.05807