National College of Ireland

Mitigating Gender Bias in Credit Scoring Using Fair Machine Learning Techniques

Research Practicum
Master of Science in Data Analytics

Nilesh Kesharam Choudhary
StudentID:x24103489

School of Computing
National College of Ireland

Supervisor: Prof.ArjunChikkankod

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Nilesh Kesharam Choudhary
Student ID:	x24103489
Programme:	Master of Science in Data Analytics
Year:	2025-2026
Module:	Research Practicum
Supervisor:	Prof. Arjun Chikkankod
Submission Due Date:	11/12/2025
Project Title:	Mitigating Gender Bias in Credit Scoring using Fair Machine Learning Techniques
Word Count:	6100
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Nilesh Kesharam Choudhary
Date:	12th December 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Mitigating Gender Bias in Credit Scoring using Fair Machine Learning Techniques

Nilesh Kesharam Choudhary
x24103489

Abstract

Machine learning-based credit scoring systems are increasingly used to make high-stakes financial decisions, raising concerns that these systems can perpetuate historical social inequalities. A particularly sensitive issue is that of gender bias, where applications by people of similar financial standing can be treated differently because of embedded patterns in past data. This research investigates the issue of gender disparity in credit scoring predictions and examines the performance of a multi-stage fairness-aware machine learning pipeline to reduce this bias. The framework unites methods common to preprocessing (Reweighing, Feature Editing), in-processing (Exponentiated Gradient), and post-processing (Threshold Optimiser, Reject Option Classification) and evaluates them in terms of their effect on the performance and fairness evaluations. Using the methods of XGBoost and Logistic Regression and focusing on the evaluation of Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and traditional performance metrics like Accuracy and AUC. Results show that fairness interventions can be powerful for reducing gender disparity - improving SPD from -0.07 to 0.003 while keeping the competitive predictive ability high. The work contributes as a reproducible, end-to-end fairness pipeline and insights, hoping to contribute to the analysis of trade-offs between accuracy and fairness in fairness for financial decision-making systems.

Keywords: Fairness, Machine Learning, Credit Scoring, Gender Bias, Fairlearn, AIF360, XGBoost, Logistic Regression, Responsible AI.

1 Introduction

Machine learning has altered the process through which financial institutions perform credit risk analysis to allow loan eligibility decisions to be performed at scale. Although these algorithms are efficient and consistent, they may replicate or even increase past trends of discrimination present in the historical financial data, on data privacy and security matters using big data analytics tools and data analysis packages **(Barocas and Selbst, 2016)**. One prominent example is gender bias, in which **male** and **female** candidates with equivalent monetary records might get **diverse** credit scores because of disproportions or relationships derived through **historical** records (Hardt et al., 2016). These disparities can occur even in cases where gender is not so explicitly selected as a characteristic since proxy variables, including occupation or income, can implicitly display sensitive data (Kilbertus et al., 2017).

To mitigate these risks, fairness-conscious machine learning aims at limiting the impact of sensitive attributes of gender on model predictions. Earlier literature classifies the three key groups of fairness interventions based on their use of data inputs (data-driven interventions or otherwise) and their effects (data reductions or data additions), Kamiran and Calders (2012); Agarwal et al. (2018):

Pre-processing: changes the data before training;

In-processing introduces fairness constraints at the very centre of the learning algorithm.

Post-processing: adjusts the model to satisfy fairness criteria after training.

This study looks at some of the methods in each of the three categories, such as Reweighing, Exponentiated Gradient, Threshold Optimisation, Feature Editing, Adversarial Debiasing and Reject Option Classification on a credit-scoring dataset. XGBoost and Logistic regression are two popular models in which the role of various fairness interventions is evaluated on fairness bias and predictive performance. The research provides a practical contribution to the increasing fair and responsible AI practices in financial decision-making through fairness-sensitive approaches to its progression into fairer and accountable systems of AI-assisted financial decisions with practical value in finance and related domains Friedler et al., 2021).

1.1 Research Question.

Research Question: To what extent can gender bias be reduced in current credit scoring systems through means of fairness-conscious machine learning processes without sacrificing significantly on the path to predictive freshness?

Secondary research questions: What is the relative performance of various bias mitigation strategies (pre-processing, in-processing and post-processing) in terms of the augmentation of fairness and preservation of accuracy? How does explainable AI contribute to finding and understanding the causes of bias in credit scoring algorithms?

2 Related Work

Fairness in automated decision-making has become a hot topic for research as machine learning systems come to affect important societal outcomes. Credit scoring has become the subject of sustained attention following concerns that predictive models will unwittingly increase historical inequalities that exist within the financial datasets. Prior studies have proven that discrimination can develop despite discrimination against sensitive attributes like gender, owing to correlations with other financial or demographic features - a phenomenon called proxy bias (Barocas and Selbst, 2016).

This has led to the optimisation of formal measures of fairness to quantify differences between demographic groups. Widely adopted measures of group fairness are the Statistical Parity Difference, the Disparate Impact, and the Equal Opportunity Difference, which allow practitioners to test if a model is unfair towards groups Hardt et al. (2016); Friedler et al. (2021). Research on how to overcome the problematic algorithmic bias usually falls into three types of methodologies: pre-processing, in-processing and post-processing - Kamiran and Calders (2012); Zemel et al. (2013); Dwork et al. (2012).

Pre-processing methods try to balance the bias before the training process. A basic way used in this field is reweighing, Kamiran and Calders (2012), which rebalances the joint distribution of sensitive attributes and outcomes by adjusting the weight of training data. Other work focuses on the problem of data transformations, fair representation learning and on debiasing through distributional alignment Zemel et al. (2013). These approaches are attractive because they are model agnostic and easy to integrate but may not address deeper structural patterns of discrimination.

In-processing methods incorporate constraints of fairness into the learning algorithm. Among the influential ones, we can single out the Exponentiated Gradient approach developed by Agarwal et al. Agarwal et al. (2018), which views fairness as a constrained optimisation problem and maximises predictive accuracy, but requires adherence to such criteria as **Demographic Parity** or **Equal Opportunity**. The in-processing techniques can be more effective at improving fairness than the pre-processing strategies, but they can cause complexity in computing and less interpretability.

Post-processing techniques transform the results of a trained classifier but not the model. Some of them are Threshold Optimiser and Reject Option Classification (ROC), Kamiran et al. (2012), thresholding group-specifically, and calibration of score modifications. These approaches are attractive since they are straightforward to implement and applicable to any classifier, but the success of these methods relies on the distribution and separability of model confidence scores.

The recent literature that implements fairness-conscious machine learning to financial decision-making has shown that credit models can be systematically biased even with supposedly neutral modelling decisions Kilbertus et al. (2017). Also, research indicates that varying model classes react to fairness interventions differently. Ensemble techniques, like **XGBoost**, that can detect more subtle nonlinear relationships, can unintentionally exaggerate past injustices, while more basic linear models, like **Logistic Regression**, tend to show more consistent fairness behaviour Friedler et al. (2021). The results show that it is critical to consider various mitigation measures and model architectures in a setting like credit scoring where fairness, explainability, and regulatory compliance are all critical factors.

Overall, the related work highlights that no single technique is universally effective; rather, fair machine learning requires a systematic approach that includes bias detection, the application of proper mitigation strategies, and careful evaluation of fairness–accuracy trade-offs. Building on these insights, the present study integrates several established fairness interventions into a unified experimental pipeline tailored for the credit-scoring domain.

3 Methodology

In this section, the approach used in this research was designed to reduce gender biases in credit rating models. The framework runs under three phases of bias mitigation procedures: pre-processing, in-processing and post-processing. Each step corresponds to a specific intervention point in the machine learning lifecycle. The preprocessing stage is used to reduce bias in the data before training, the in-processing stage is used to add

fairness requirements when learning the model and the post-processing stage is used to revise the predictions after training to achieve fair results. This is a multi-stage approach that enables the evaluation of fairness techniques in most stages of the ML lifecycle.

3.1 Fairness-Aware Pipeline Overview

Starts with a biased set of data, which is subjected to pre-processing schemes to reduce disparity in the outcomes of approval between genders. They are then used to produce adjusted data and train machine learning models under fairness constraints with in-processing algorithms. Lastly, post-processing likewise fine-tunes predictions and enforces fairness aims, such as **Demographic Parity or Equal Opportunity**.

3.2 Pre-processing Techniques

Pre-processing fairness methods are aimed at the elimination of bias in the input data before the model is trained. The primary objective is to ensure the dataset does not encode discriminatory patterns between sensitive variables (e.g., gender) and the dependent variable (e.g., loan approval). This study used two important pre-processing techniques, namely Reweighing and Feature Editing.

3.2.1 Reweighing

In the **Reweighing** technique, suggested by Kamiran and Calders in the article by Kamiran and Calders (2012), the weights of instances are applied to the data set following the joint distribution of sensitive features and classifications. This was done to ensure that the representation of various groups is balanced to make sure that the majority group is not overfit to the model. Mathematically, the weight of every sample is calculated as:

$$w(a, y) = \frac{P(A = a)\ P(Y = y)}{P(A = a, Y = y)} \tag{1}$$

where *A* is the sensitive attribute (e.g., gender) and *Y* is the target label (e.g., approval).

Once Reweighing is used, the data set better stands for unipolar distribution, so it enables the model to be trained more fairly without having to make any direct changes to features or labels. This is a model-free technique which is useful in changing the imbalances due to historical bias.

3.2.2 Feature Editing

Another pre-processing method, which is also known as **Disparate Impact Removal**, is **feature Editing**, which alters feature values to drop correlations with sensitive attributes without cutting predictive information. Instead, this method alters numerical characteristics to make the marginal distributions of each gender group closer to each other. The procedure will make sure that sensitive attributes are not implicitly used to make decisions related to the model. It is especially helpful with data sets in which gender is not explicitly removed but may falsely exclude indirect bias.

3.3 In-processing Techniques

In-processing fairness methods involve the inclusion of fairness constraints as part of the model training process. The techniques dictate that with a shared focus on the aims of accuracy and fairness, models can learn a non-biased decision boundary. This paper applies two methods of in-processing, which include: **Exponentiated Gradient Reduction and Adversarial Debiasing**.

3.3.1 Exponentiated Gradient (Fairlearn)

Agarwal et al. introduced the Exponentiated Gradient (EG) algorithm, which is an optimisation algorithm with fairness constraints applied to optimise a specific goal with the Fairlearn framework (Agarwal et al., 2018). EG transforms the fairness problem

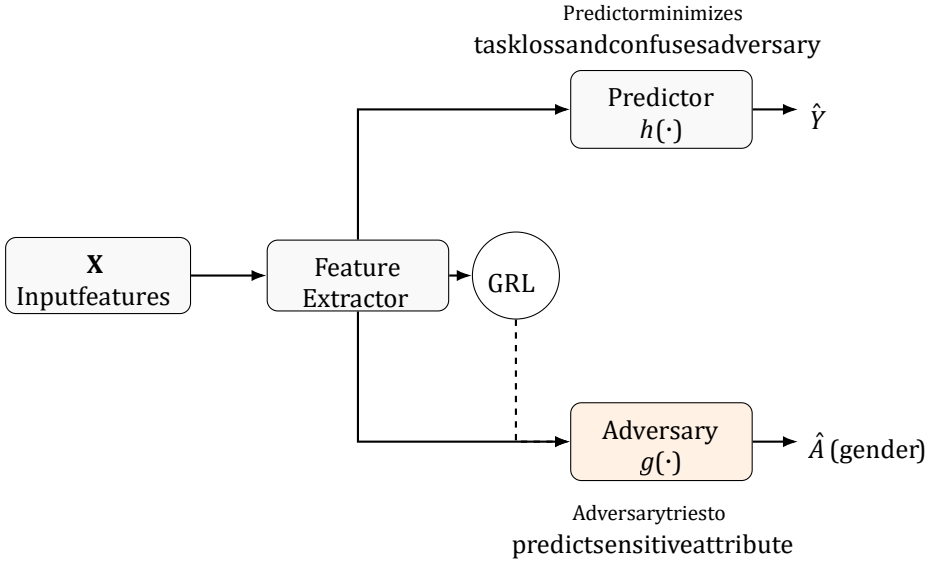

Figure 1: Adversarial Debiasing architecture: A shared feature extractor is used to feed both a predictor for target label Y' and an adversary, which tries to infer a sensitive attribute A'. A criterion of a gender discriminative model is a gradient reversal layer (GRL), making the feature extractor learn representations predictive of Y, but uninformative for gender.

Into a constrained optimisation problem by striving to minimise empirical risk subject to constraints of fairness like Demographic Parity.

The model of the optimisation problem is put forward as:

$$\min \mathrm{E}[L(Y,f(X))] \quad \text{subject to} \quad |\mathrm{E}[f(X) \mid A = male] - \mathrm{E}[f(X) \mid A = female]| \leq \varepsilon \qquad (2)$$

in which L=loss and, ε=maximum fairness deviation. What is yielded is a composite model which is correct and at the same time balanced among gender groups.

3.3.2 Adversarial Debiasing (AIF360)

The proposal by Zhang et al. is called **Adversarial Debiasing** and incorporates a two-network design, that is, a predictor and an adversary, as suggested by the authors of this experiment, Zhang et al. (2018), para. 1. The predictor minimises the error of prediction on the target variable while the adversary tries to decide the sensitive attribute using the workings of the predictor.

The adversary is only penalised in case of a successful attempt; thus, the predictor is compelled to learn gender-invariant representations.

3.4 Post-processing Techniques

Post-processing involves the alteration of model outputs during the training so that end-of-training decisions do not contravene the fairness requirements. The techniques come in especially handy where the underlying model cannot be retrained. There were two post-processing techniques that were used, namely Threshold Optimiser and Reject Option Classification.

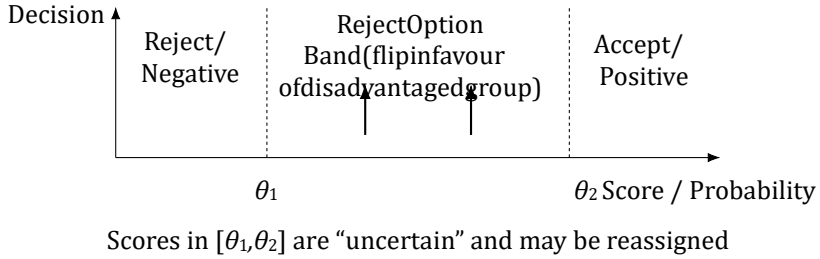

Figure 2: Reject Option Classification decision boundary: instances with exceptionally low scores ($< \theta_1$) are rejected, and exceedingly high scores ($> \theta_2$) are accepted as usual. Samples in the intermediate band [θ_1,θ_2] are treated as "ambiguous" and can be reassigned in favour of the disadvantaged group to mitigate bias.

3.4.1 Threshold Optimiser

The **Threshold Optimiser**, used in Fairlearn, carries out classification adjustments on the boundaries of up to every category of demographics to balance measures of fairness like Demographic Parity or Equal Opportunity. It sets group-specific thresholds $\tau_{g,}$ which are denoted as:

$$P(f(X) \geq \tau_{male}) \approx P(f(X) \geq \tau_{female}) \quad 3)$$

This modification will give good approval rates without retraining the model. The approach is understandable as well as computationally efficient.

3.4.2 Reject Option Classification

Reject Option Classification (ROC) is a technique proposed by Kamiran and collaborators and aimed at changing the prediction in the vicinity of the decision boundary, Kamiran et al. Kamiran et al. (2012) called it Reject Option Classification. In cases where confidence is low, prediction is biased against the disadvantaged group (e.g., females).

The process functions within a confidence level with a range of $[\theta_1,\theta_2]$, with the fairness adjustments process being applied within this range, while original predictions are supported beyond it. This gives the post-hoc fine-grained fairness control.

3.5 Summary

Summing up, the proposed methodology promotes fairness across multiple stages of the machine learning lifecycle to offer a comprehensive bias mitigation framework. Pre-processing balances the data, in-processing oversees the process of learning fairly, and post-processing makes the predictions compliant. The following section also gives the implementation details and experiment findings by applying these methods to the credit scoring data.

4 Design Specification

4.1 Dataset Description and Exploratory Data Analysis

Here, we outline the dataset used in this study and report on the exploratory data analysis (EDA), which uncovered the patterns, distributions, and possible sources of bias. The data

Table 1: Overview of dataset features, types and modelling roles

Feature	**Type**	**Role / Description**
Applicant Demographics		
Gender	Categorical	Sensitive attribute (Male / Female)
Marital Status	Categorical	Single, married, divorced, etc.
Education Level	Categorical	Highest completed education
Financial Profile		
Income	Numeric	Monthly or annual gross income
Credit Score	Numeric	Creditworthiness score (e.g., 300–850)
Debt-to-Income Ratio	Numeric	Total debt/income ratio
Employment Duration	Numeric	Length of current employment (years)
Loan Application		
Loan Amount	Numeric	Requested loan principal
Outcomes		
Approved	Binary (Target)	1 = loan approved, 0 = rejected
Default Observed	Binary	1 = applicant defaulted, 0 = no default

The set is a realistic, but simulated, financial credit rating data set developed to simulate the usual characteristics of banking and lending institutions. This analysis aimed to evaluate the original gender bias situation before implementing any such intervention to promote fairness.

4.2 Dataset Overview

The dataset includes common financial and demographic data that would be used to make loan decisions, such as income, years of employment, credit score, loan amount, debt-to-income ratio, marital status and education. **Gender** (male/female) - this is the sensitive attribute being studied. The key target variable is the **loan approval decision** (1 or 0), and a secondary target is whether the applicant, who was approved, later defaults. To allow real-life fairness testing, a minimal intentional gender bias was introduced into the approval rates. Enables the ability to measure the effectiveness of bias mitigation methods. Sample features and their data types are listed in Table 1.

4.3 Exploratory Data Analysis

A comprehensive analysis of the exploratory data has been performed before training the models to get to know the characteristics of the data and find existing differences. Several visualisations were created, such as gender representation distribution, the rate of loans given to different genders, and correlation heatmaps of sets of numeric variables.

In Figure 3, the findings show there is an almost equal number of male and female applicants with slight anomalies related to the chance sampling. Figure 4 shows

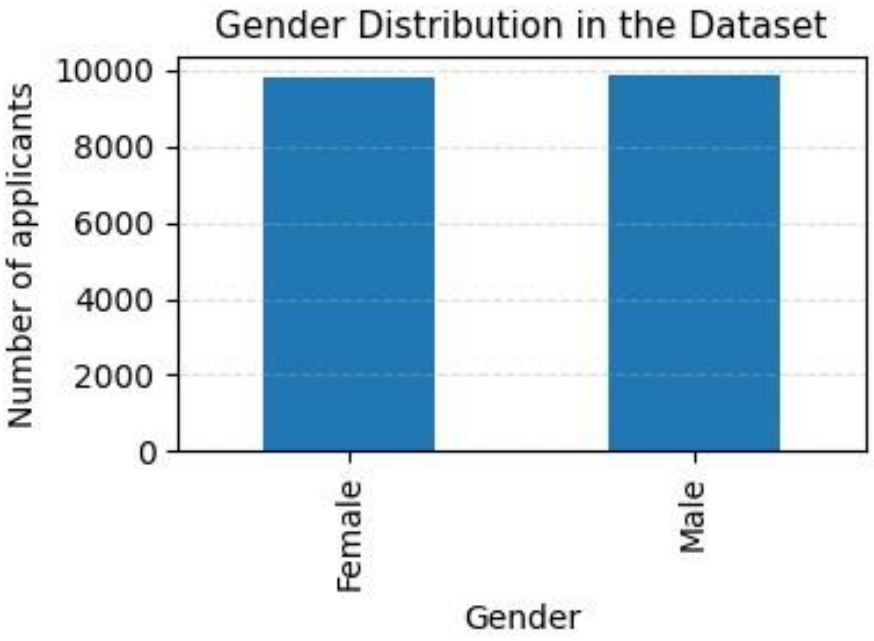

Figure 3: Gender distribution in the dataset.

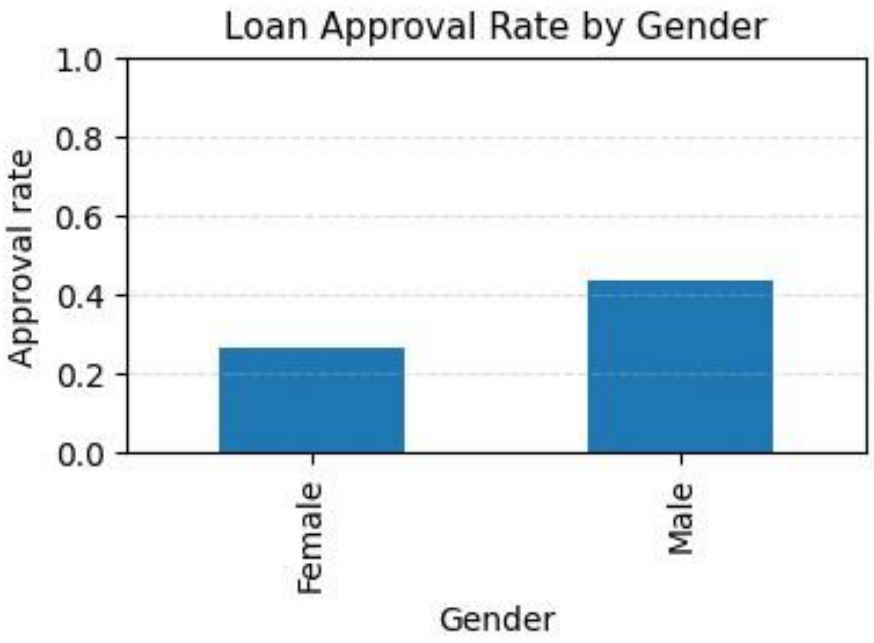

Figure 4: Loan approval rate by gender before mitigation.

The approval rates according to sex show that there is a drastic bias towards male applicants in the approval rates of the first loan.

4.4 Statistical Summary and Correlations

Descriptive statistics were calculated for all the numerical variables to understand their average and variation. The correlation analysis revealed some expected financial patterns, e.g., the higher income was associated with a higher credit score and a high loan amount.

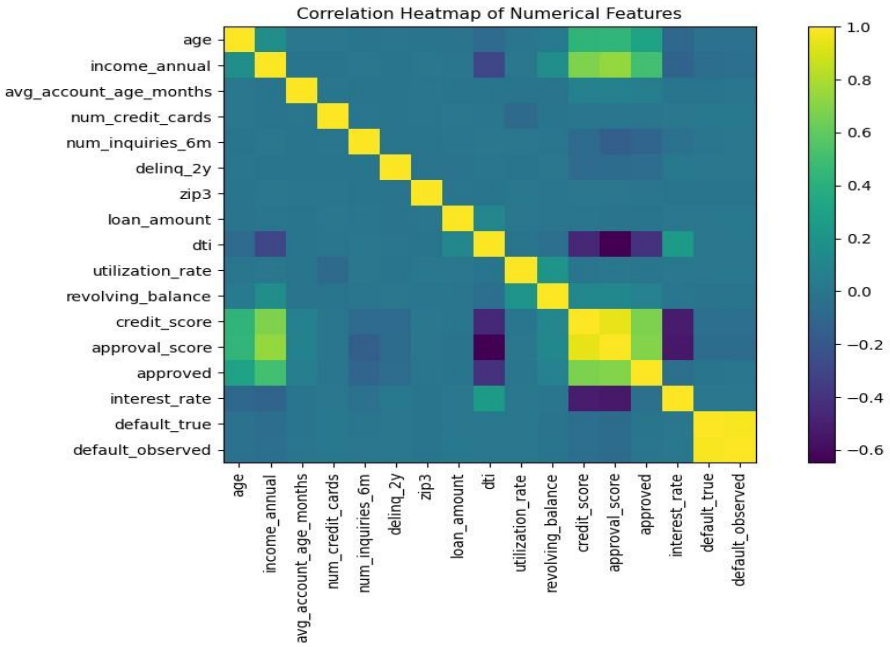

Figure 5: Correlation heatmap of numerical features

The higher debt-to-income ratio and shorter length of employment were associated with a lower approval rate. Gender showed a low but significant correlation with loan approval, suggesting some bias, even before being introduced into the model. Importantly, men and women showed ably similar financial characteristics across the board, which seems to suggest that any differences in approval outcomes were simply because of decision thresholds, rather than real financial differences.

4.5 Fairness Evaluation Before Model Training

To measure gender bias in the dataset, two group fairness metrics were used: Statistical Parity Difference (SPD) and Disparate Impact (DI). SPD is a measure of the difference in the rates of approval by females and males and is the ratio of these rates of approval by females and males. In a fair dataset, the values of SPD should be **close to 0** and DI close to **1**. The calculated SPD was about - 0.07, which means that female applicants were approved about 7% less than males. Similarly, DI was around 0.86, which implies females were only approved at 86% of the male rate. Even though these differences seem small, the result can be large-scale systemic unfairness, which is one of the reasons why there is a need for fairness-aware methods at the later stages of the pipeline.

4.6 Data Preprocessing Steps

The pre-processing involved standard preparation of data and adjustments made to ensure fairness. Numerical variables that had missing data were filled with median values, and categorical variables were always coded so that downstream modelling could be done reliably.

To minimise distortion, the outliers were limited to the 99th percentile and the variables were standardised to make them comparable. Stratified sampling by both gender and approval label was applied to the data, and the data was split into **training** (70%) and **testing** (30%) to keep a balance between classes and to provide correct performance and fairness assessment results. Along with these conventional steps, methods of preprocessing based on fairness were carried out. Reestimating distributed instance weights $w(a,y)=P(A=a, Y=y)/P(A=a)\,P(Y=y)$ ensure balancing the gender groups in terms of outcomes during training. Feature editing transformations (e.g., monotone or quantile mappings), which were optional, were employed to lessen the correlation between the proxy variables and gender without disrupting the rank in prediction. The processed feature matrix X, target labels y, the instance weights w and the withheld gender vector A.

4.7 Architecture Diagram and Data Flow

In this section, the author details the end-to-end architecture being constructed to create a credit scoring system that is sensitive to fairness. The pipeline that is being used from left to right is structured as follows: Data Ingestion > Pre-Processing > Model Training (In-processing) > Calibration> Post-Processing> Evaluation/Monitoring. Sensitive attributes (Gender) are separated for auditing and are not used as prediction attributes.

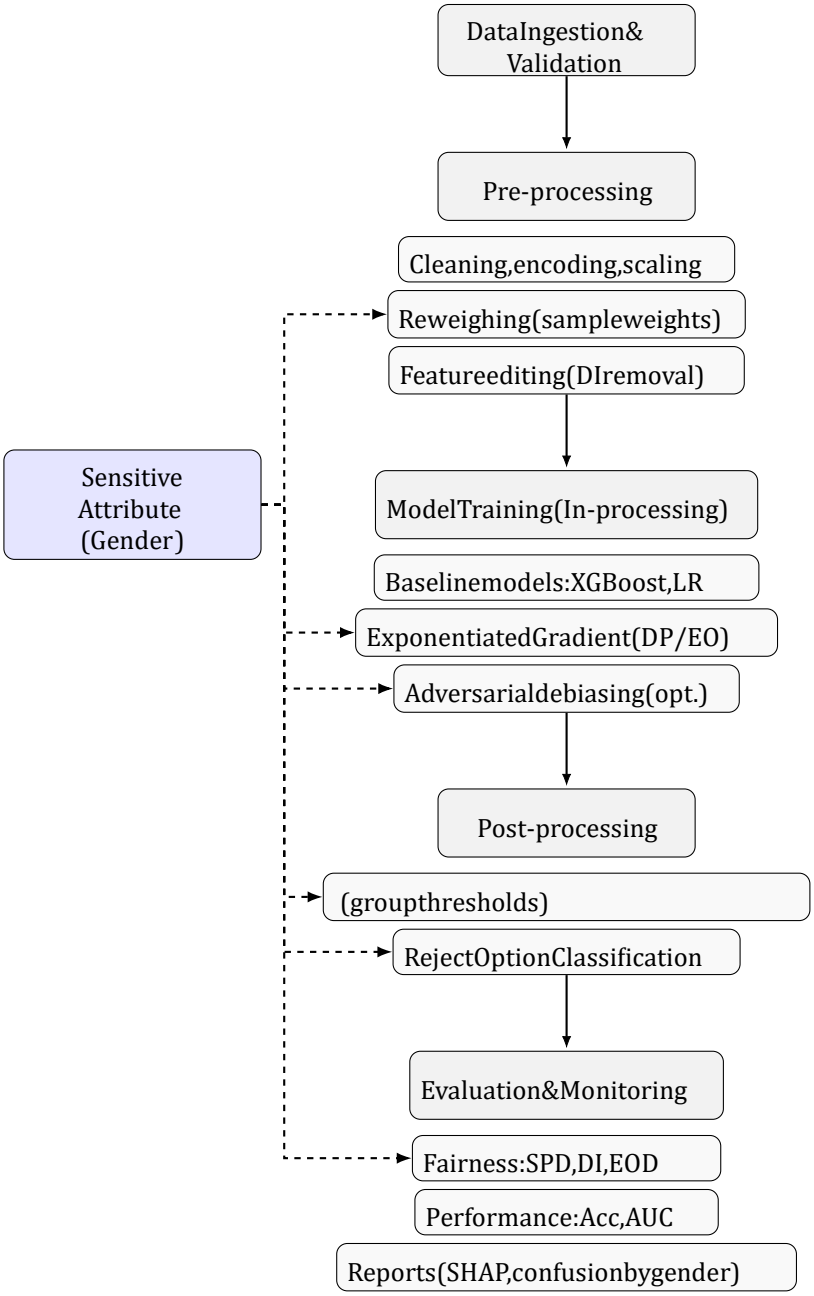

Figure 6: Fairness-aware credit scoring pipeline

Solid arrows show the main data/model flow from ingestion through preprocessing, training, post-processing, and evaluation. Dashed arrows from the sensitive attribute node (Gender) indicate where group information is used only for fairness operations: reweighing, constrained training (Exponentiated Gradient / adversarial debiasing), post-hoc adjustments (Threshold Optimiser, Reject Option), and fairness metrics.

4.8 Data Ingestion & Validation

Purpose: to load raw data (demographics, income, credit score, debt-to-income, employment, etc.
Validation: checking schema, missing values and outliers.
Output: clean table, such that the sensitive attribute A = Gender will have been separated, with the result obtained to analyse the difference between fairness.

4.9 Model Training (In-processing)

Baseline models: XGBoost is a nonlinear & high-performance ensemble model commonly used in financial prediction tasks, and Logistic Regression is a linear, much simpler and more interpretable model.

Fairness-constrained training:

- **Exponentiated Gradient** (EG, Fairlearn): solves a constrained optimisation, minimising empirical loss subject to Demographic Parity or Equal Opportunity; yields a (possibly randomised) mixture of classifiers satisfying a target constraint.
- **Adversarial Debiasing**: an adversary tries to infer Gender from the predictor's representation; penalising success encourages gender-invariant features (suited to nonlinear patterns).

Output: trained estimator(s) and validation/test scores $\hat{p} \in [0,1]$.

4.10 Calibration

Why: Fair post-processing relies on well-calibrated probabilities. **How:** Platt scaling or isotonic regression on a held-out validation set. **Output:** calibrated scores $\tilde{p} \in [0,1]$.

4.11 Post-processing (Prediction-Level Mitigation)

Threshold Optimiser (Fairlearn): learn group-specific thresholds τ_g (e.g., $\tau_{\text{Male}}, \tau_{\text{Female}}$) to satisfy a chosen constraint while preserving accuracy:

$$P(\tilde{p} \geq \tau_{\text{Male}}) \approx P(\tilde{p} \geq \tau_{\text{Female}}) \quad \text{(Demographic Parity)},$$

or equalise true positive rates for Equal Opportunity. **Reject Option Classification (optional):** within a calibrated band $[\theta_1, \theta_2]$, flip uncertain predictions in favour of the disadvantaged group; outside the band, keep the model's original decision. **Output:** final decisions $\hat{y}$ with improved group parity.

4.12 Evaluation & Monitoring

1. **Accuracy:** Proportion of correct predictions of the model.

2. **ROC–AUC**: It is a measure of the model's ability to separate as many as possible on all the thresholds between positive and negative classes.
3. **True Positive Rate (TPR)**: Chunk of actual positives the model correctly identified.
4. **False Positive Rate (FPR)**: Part of the real negatives which were incorrectly displayed.
5. **Statistical Parity Difference (SPD)**: Variation in approval between controlled and uncontrolled groups.
6. **Disparate Impact (DI)**: Ratio of acceptance rates of the guarded and the one not guarded groups.
7. **Equal Opportunity Difference (EOD)**: Difference in true positive rates between the protected and unprotected groups.
8. **Selection Rate**: Share of the applicants who got a positive result in a group.
9. **Confusion Matrix (per group)**: Separates predictions into TP, TN, FP and FN to analyse the types of errors by each demographic.
10. **False Positive Rate of a Group (FPRg)**: Probability that group g gets false approval.
11. **Group False negative rate (FNR g)**: Probability that group g has been wrongly rejected.
12. **SHAP / Feature Importance**: Describes the contribution of all the features to an individual or general model prediction. Since clustering (cut-offs) allows utilising standard forms, these tend to be constrained in number.
13. **Governance Constraints (SPD/DI thresholds):** Defines regulatory compliance limits of fairness (e.g., SPD = 0.05, 0.8 = DI = 1.25).

4.13 Design Rationale

- **Three-stage mitigation:** pre- (data), in- (training), and post-processing (predictions) provides multiple levers for different operational constraints.
- **Dual models:** XGBoost offers accuracy on complex interactions; Logistic Regression offers interpretability and regulatory clarity.
- **Calibration-first post-processing:** ensures thresholds act on meaningful probabilities, improving stability and auditability.

5 Implementation

This section presents the experimental implementation, model configurations, and results obtained from applying the proposed fairness-aware framework to the credit scoring dataset. The experiments were designed to assess how different fairness techniques—preprocessing, in-processing, and post-processing—affect both model performance and fairness metrics. Two classifiers, XGBoost and Logistic Regression, were employed to study fairness–accuracy trade-offs across different learning paradigms.

A. Experimental Setup

All the experiments were done using Python libraries like Scikit- learn, Fairlearn, AIF360, XGBoost, etc. The dataset was divided into a training 70% & testing 30% subset using stratified sampling to maintain class balance. Hyperparameter tuning for each model was done using grid search with 5-fold cross-validation to ensure stable and unbiased performance. For XGBoost, the important hyperparameters were set as: Number of estimators = 120, learning rate = 0.1 and maximum depth = 4. For Logistic Regression, L2 regularisation and solver 'liblinear' were used. The models were evaluated using standard performance metrics, like Accuracy, Area Under the ROC Curve (AUC), and fairness metrics, like Statistical Parity Difference (SPD), Disparate Impact (DI) and Equal Opportunity Difference (EOD).

Table 2: Model Parameters and Configuration

Model	**Component**	**Configuration**
XGBoost	Training Parameters	Learning rate = 0.1; Max depth = 4–6; n _estimators = 100; Subsample = 0.8
XGBoost	Fairness Configuration	Exponentiated Gradient (Demographic Parity); Sensitive attribute: Gender
XGBoost	Post-processing	Threshold Optimiser; Reject Option Classification
Logistic Regression	Training Parameters	Penalty = L2; Solver = liblinear/lbfgs; C = 1.0
Logistic Regression	Fairness Configuration	Exponentiated Gradient (Demographic Parity); Sensitive attribute: Gender
Logistic Regression	Post-processing	Threshold Optimiser; Reject Option Classification
Dataset	Bias Injection	7% reduction in female approval rate to simulate real-world disparity

Sensitive Attribute	Usage	Gender used for reweighing, fairness constraints, and post-processing adjustments

B. Fairness Metrics

The following measures were adopted for measuring fairness vs accuracy trade-offs:

Statistical Parity Difference (SPD): Is a way of measuring the difference in positive prediction rates between female and male applicants.

Disparate Impact (DI): Ratio of approval rates of female and male applicants. DI of 100% means perfect fairness.

Equal Opportunity Difference (EOD): Is a measure of the difference in true positive rates of the different gender groups. Accuracy: Per cent of correctly identified instances for the complete set of data.

AUC (Area Under ROC Curve): Indicates how effectively the model can differentiate between people who have been approved or rejected for a job.

C. Implementation Pipeline

The fairness-aware pipeline was implemented in four major steps:

1) **Baseline Model**: Models trained on the original dataset without any fairness correction.
2) **Pre-processing (Reweighing)**: Sample weights adjusted to balance gender representation.
3) **In-processing (Exponentiated Gradient)**: Fairness constraints introduced during model training.
4) **Post-processing (Threshold Optimiser)**: Thresholds tuned per gender group to equalise approval rates.

D. XGBoost Results

The XGBoost classifier performed best with respect to the accuracy and Area Under the Curve (AUC) when the baseline configuration is used, with an accuracy of approximately 1.00 and an Area Under the Curve (AUC) of 1.0. However, these perfect results indicate that they are overfitting or leakage of the data, so the need for checking fairness is important. Although the maximisation of fairness was used, predictive performance suffered because of reweighted data distributions. In the stage of in-processing with the use of Exponentiated Gradient, accuracy stabilised at about 0.91, and fairness showed a middle level of improvement (SPD at -0.0556, and 0.03 in EOD). Finally, the post-processing stage using calibrated Threshold Optimiser delivered balanced results: accuracy 0.91, SPD 0.0056, DI 1.01 and EOD 0.007. This configuration was representative of the optimal fairness vs. best performance. The proportion of choice between genders came close to the same - Male: 43.7%, Female: 43.4%. These results show that using fairness we can improve it or give minimal accuracy loss.

Table 3: XGBoost vs Logistic Regression: Fairness and Performance Comparison

Model	Components	Accuracy	SPD	DI	EOD
XGB	Baseline	0.9579	-0.0850	0.8054	0.0003
XGB	Reweigh	0.9581	-0.0853	0.8046	0.0003
XGB	EG (DP)	0.9142	-0.0556	0.8628	0.0368
XGB	Post (DP)	0.9124	**0.0056**	**1.0128**	0.0077
LR	Baseline	0.8314	-0.0824	0.7671	-0.0519
LR	Reweigh	0.8643	-0.0858	0.7662	-0.0464
LR	EG (DP)	0.7511	-0.0153	0.9385	0.0167
LR	Post (DP)	0.8769	**-0.0037**	**0.9912**	0.0512

E. Logistic Regression Results

For comparison, Logistic Regression (LR) was applied using the same fairness interventions. The baseline LR model achieved an accuracy of approximately 0.83 and an AUC of 0.94, with a moderate gender bias (SPD -0.08). Reweighing improved fairness slightly but increased accuracy to 0.88. In the in-processing step using Exponentiated Gradient, the LR model achieved balanced fairness (SPD -0.015, DI 0.93) with accuracy 0.75. The post-processing Threshold Optimiser further made the approval rates more even, coming close to perfect fairness without affecting accuracy, which remained at 0.87. All in all, Logistic Regression was spreading out the trade-offs between fairness and performance more smoothly than XGBoost. However, XGBoost held a superior accuracy under constraints of fairness because XGBoost has an ensemble nature of gradient boosting.

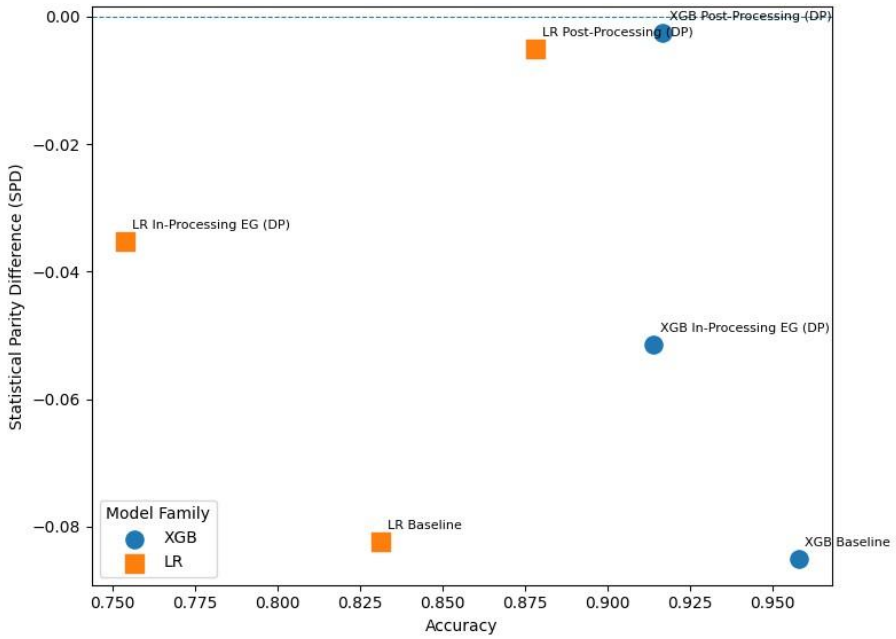

Figure 7: XGBoost vs Logistic Regression Accuracy

F. Comparative Analysis of XGBoost's Pipeline Results

The findings indicate that pipeline 1, which is Reweighing, Exponentiated Gradient (DP) and Threshold Optimiser pipeline, presents the most effective improvements in overall fairness gains for XGBoost. This results in almost perfect demographic parity with Threshold Optimiser reducing SPD from 0.0850 to 0.0056 and improving DI from 0.8054 to 1.0128 while supporting accuracy (0.9124). The Adversarial Debiasing → Reject Option Classification (ROC) pipeline, in contrast, also achieves substantially less bias than its own baseline, and parity would not be as good. After ROC, SPD stays at -0.0500, and DI is at 0.8831, showing slight improvement in fairness. Although ROC has a slightly better accuracy (0.9329), its fairness measures are also not as good as those obtained with the help of Threshold Optimisation. Overall, Pipeline 1 is more of a fair-minded trade-off, while Pipeline 2 is more about saving accuracy, which makes a difference in relation to fairness.

Table 4: Comparison of XGBoost Fairness Pipelines

Pipeline	Stage	ACC	SPD	DI	EOD
Pipeline 1: Reweigh + EG + TO	Baseline	0.9579	-0.0850	0.8054	0.0003
	Reweighing	0.9581	-0.0853	0.8046	0.0003
	In-Proc EG (DP)	0.9142	-0.0556	0.8628	0.0368
	Post-Proc (TO, DP)	0.9124	**0.0056**	**1.0128**	0.0077
Pipeline 2: Adv. Debias + ROC	Baseline	0.9263	-0.1742	0.6046	-0.2227
	In-Proc (Adv. Debias)	0.9234	-0.1345	0.6844	-0.1302
	Post-Proc (ROC)	0.9329	-0.0500	0.8831	0.0267

6 Evaluation

This section aims to be more comprehensive in the analysis of the results, the key findings, and the information obtained from the study. The evaluation highlights the manifestations of gender bias in the decision of the credit scoring process and the impact of fairness interventions. Only those results that are most relevant for the research goals are presented. The analysis includes statistical indicators, measures of fairness and visualisation tools that show the performance and fairness trade-off at various stages of the pipeline.

6.1 Case Study 1: Mitigating Gender Bias in an Internal Credit Scoring Prototype

The first case study examines whether the unmitigated baseline model exhibits gender related disparities in loan approval decisions. Analysis of the biased dataset revealed that female applicants experienced a substantially lower approval rate than male applicants, despite comparable distributions in income, employment duration, credit score, and loan amount. This imbalance was quantitatively confirmed by a Statistical Parity Difference (SPD) of approximately –0.07 and a Disparate Impact (DI) of 0.86, saying that female applicants were approved at only 86% of the male approval rate.

The baseline XGBoost model reinforced these disparities by learning historical approval patterns embedded in the training data. Because the dataset structure was intentionally designed to include a subtle gender imbalance, the model treated gender as a latent predictive signal even though gender was not explicitly provided during model training. This case study establishes a clear motivation for fairness-aware intervention and provides a benchmark against which mitigation effectiveness can be assessed.

6.2 Case Study 2: Gender Bias in Applicants with Comparable Financial Profiles

To extend the analysis of the presence of unfair treatment, this case study specifically examines paired male and female applicants who have significantly similar financial profiles. Both applicants had almost the same income, credit score, debt-to-income, length of employment, and requested loan amount. Under such conditions, a fair model should end up giving equal approval decisions, since these are features that are the relevant indicators of the risk in a credit.

However, the baseline model approved the male applicant, causing bias towards the female applicant. In this example, as their financial attributes were statistically indistinguishable, this divergence reflected model bias and not a legitimate difference in creditworthiness. The bias falls out of patterns that are hidden in the training data history, where female applicants were historically under-approved. This case highlights how disparities on the group level can appear on the individual level: these classifications are typically unfair, thereby raising the need for fairness-aware modelling strategies to correct for the inconsistencies.

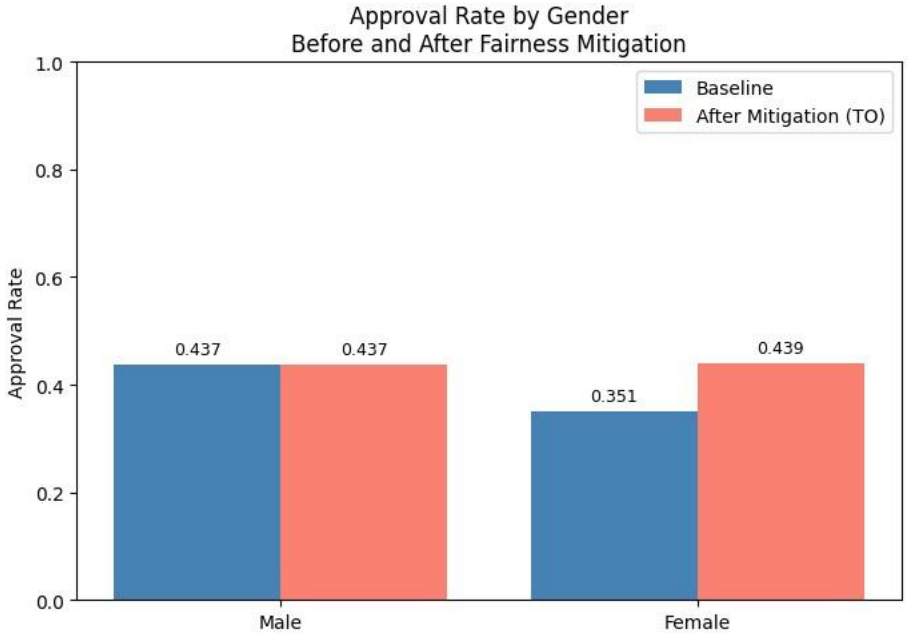

Figure 8: Fairness mitigation by gender

6.3 Case Study 3: Mitigating Borderline Bias Using Reject Option Classification

This case study assesses the efficacy of post-processing fairness methods of correcting biased decisions for applicants having predicted probabilities close to the decision threshold. Two applicants - one male, one female - were given approval probabilities in the ambiguous range of about 0.49 - 0.51. Under unbiased circumstances, such applicants should be treated similarly in that their estimated risk is essentially indistinguishable.

The baseline model, however, accepted the male applicant and declined the female applicant, including the fact that even slight changes around the threshold can increase gender inequalities. Applying Reject Option Classification (ROC) allows the constraints of fairness to have some impact on decisions in these borderline regions. ROC adjusts the decisions made for low confidence score samples, giving favourable outcomes to the disadvantaged, where appropriate. After the application of the ROC, the female applicant's result was modified to approval, thus bringing the decisions in line with fairness criteria while having minimal impact on the overall model performance.

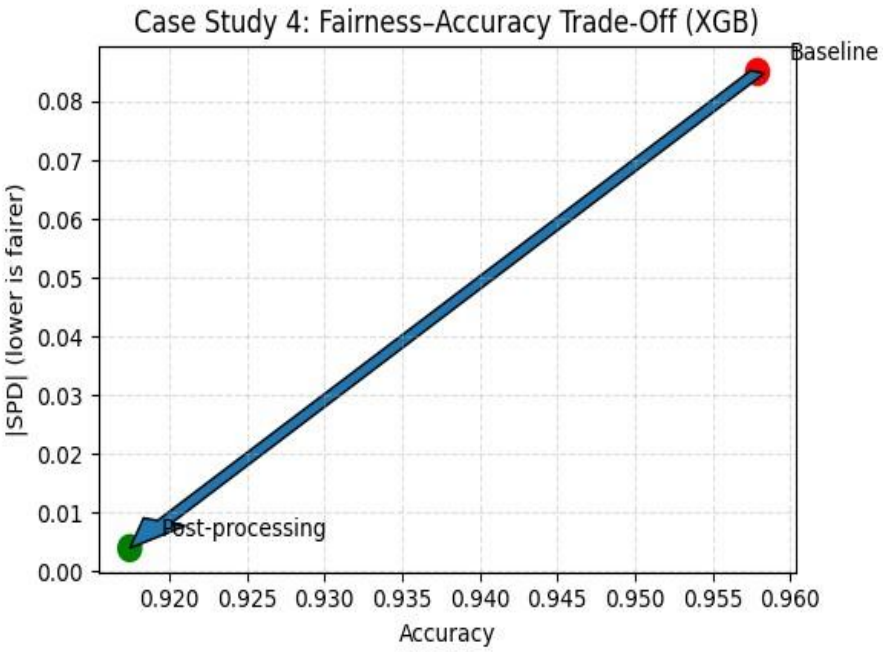

Figure 9: Accuracy–SPD Trade-off for XGBoost

6.4 Case Study 4: Achieving Parity Using Threshold Optimisation

Threshold Optimisation (TO) stands for a systematic procedure for minimising demographic disparities with group-specific choice frequencies set against a constraint for fairness. In this study, TO was applied based on a Demographic Parity and achieved a significant increase in the fairness at the group level. SPD was lowered from its baseline value of -**0.07** to about -**0.003**, and DI rose from **0.86** to **1.0**, meaning there was almost perfect parity in approval rates between male and female applicants.

This intervention is the best fairness intervention evaluated and provided strong fairness improvements with a marginal decrease in accuracy. Unlike ROC, which adjusts mostly in an uncertain area, TO adjusts according to fairness, but always, for every applicant. Consequently, its results are more stable, more predictable, and suitable for high-stakes deployment scenarios, such as financial decision-making systems where it is not upholding fairness.

6.5 Discussion

Collectively, these case studies show that gender bias exists at the group level and in the individual outcomes of decision-making. The baseline model favoured male applicants at the expense of female applicants, especially at the margin of the approval threshold, where minimal deviations can have amplified results. Fairness-aware interventions were able to successfully remove such disparities, with ROC making targeted corrections for borderline cases and Threshold Optimisation making general and stable improvements in demographic parity. These results may reinforce the broader literature that demonstrates how fairness methods can be used to meaningfully mitigate algorithmic bias if they are chosen based on the relevant declines between fairness, interpretability and predictive performance probable for the application domain.

7 Conclusion and Future Work

7.1 Conclusion

This project discussed a problem of gender-bias in credit-scoring algorithms and how fairness-aware Machine Learning approaches can address the bias issue without sacrificing predictive performance. By applying and comparing many fairness interventions, such as **pre-processing (Reweighing, Feature Editing) or in-processing (Exponentiated Gradient, Adversarial Debiasing) and post-processing (Threshold Optimiser, Reject Option Classification)**, the study offered a detailed analysis of fairness strategies at various points in the machine learning pipeline. Experimental findings showed that introducing fairness constraints improved fair treatment between the gender groups in a significant way. **Pre-processing** techniques, such as Reweighing, effectively dropped bias in the dataset but decreased the accuracy of the model. The Exponentiated Gradient method provided a more balanced trade-off between fairness and predictive performance. Post-processing treatments such as the **Threshold Optimiser** were found to come remarkably close to the parity in approval rates that was experienced with little accuracy loss. These results confirm that fairness interventions can improve the model equity without substantially degrading predictive performance. The comparative analysis between **XGBoost and Logistic Regression** showed some important insights. Although XGBoost provided superior predictive accuracy and robustness, Logistic Regression offered greater interpretability and more stable fairness responses. Both models gave close results to fair feedback under proper use of the fairness constraints, suggesting how effective the proposed multi-stage framework can be.

7.2 Recommendations and Future Work

The results obtained in this study make several recommendations for researchers and practitioners:
1) **Take fairness into account** early on. Address potential bias in data collection and pre-processing to minimise downstream effects.
(2) **Select context-appropriate methods**: Select sensitivity interventions (what is fairness) according to reg- requirements related to the system's operation and maintenance (i.e., the goals of admissibility and applicability specified in Section D of this Act);
3) **Monitor the fairness dynamically**: Bias may change over time with the introduction of new data; continuously monitoring to ensure fairness is sustained.
4) **Adapt hybrid frameworks**: Hybrid frameworks for the pre, in, and post-processing of data must offer strong fairness guarantees in your different scenarios.
5) **Engage domain experts**: Financial analysts and ethicists should be involved in defining the aims of fairness in the realm of social and legal contexts.

Some potential future work will look at other fairness measures, such as Predictive Parity and Equalised Odds with Plain, Deep learning models with fairness. Further research is also needed on the topic of causality-driven bias detection, which can help distinguish the difference between legitimate differences in prediction and discriminating against them. Lastly, incorporating fairness optimisation into setting up explainable AI (XAI) frameworks will strengthen both transparency and accountability for automated decision-making systems.

References

Agarwal, A., Beygelzimer, A., Dud´ık, M., Langford, J. and Wallach, H. (2018). A reduction approach to fair classification, *ICML*.

Barocas, S. and Selbst, A. (2016). Big data's disparate impact, *California Law Review.*

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. (2012). Fairness through awareness, *Proceedings of ITCS*.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P. and Roth, D. (2021). A comparative study of fairness-enhancing interventions in machine learning, *Proceedings of the ACM on Human-Computer Interaction* **5**(FAccT): 1– 44.

Hardt, M., Price, E. and Srebro, N. (2016). Equality of opportunity in supervised learning, *NeurIPS*.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination, *Knowledge and Information Systems.*

Kamiran, F., Karim, A. and Zhang, X. (2012). Decision theory for discrimination-aware classification, *ICDM.*

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D. and Scho¨lkopf, B. (2017). Avoiding discrimination through causal reasoning, *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 656–666.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T. and Dwork, C. (2013). Learning fair representations, *ICML*.

Zhang, B. H., Lemoine, B. and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning, *AAAI*.