

Lesson 2.4: Algorithmic Fairness and Bias – Exercises

Module 2: The Access Problem

Prof. Dr. Joerg Osterrieder

Exercise 1: Disparate Impact Calculation

A bank's auto-loan model produces the following approval rates by race:

Group	Applications	Approved	Approval Rate
White	5,000	3,500	70%
Black	2,000	1,080	54%
Hispanic	1,800	1,080	60%
Asian	1,200	900	75%

Tasks:

- 1 Calculate the disparate impact ratio for each group relative to the most-favored group
- 2 Which groups fail the four-fifths rule ($DI < 0.80$)?
- 3 The bank argues that the differences reflect legitimate credit risk. What additional analysis would you perform to test this claim?
- 4 If you applied a post-processing threshold adjustment to achieve demographic parity, what would be the potential legal and ethical concerns?

Exercise 2: Proxy Variable Audit

A mortgage lending model uses the following features. For each, assess whether it could act as a proxy for a protected attribute.

Feature	Proxy for...?
Annual income	
Zip code (5-digit)	
University attended	
Number of dependents	
Years at current employer	
Grocery store frequency	
Name of employer	
Web browser language setting	

Tasks:

- 1 For each feature, identify which protected attribute(s) it could proxy for (race, gender, age, national origin, religion, disability) and explain the mechanism
- 2 Rank the features from highest to lowest proxy risk
- 3 For the top-3 proxy risks, propose a mitigation strategy (remove, transform, or monitor)

Exercise 3: Fairness Metric Comparison

A credit scoring model produces the following outcomes for two groups:

Metric	Group A (n=3,000)	Group B (n=1,500)
Base default rate	8%	15%
Approval rate	72%	55%
True Positive Rate (TPR)	89%	74%
False Positive Rate (FPR)	6%	19%
Calibration (score=0.7 → actual default)	71%	69%

Tasks:

- 1 Does this model satisfy demographic parity? (threshold: $DI \geq 0.80$)
- 2 Does this model satisfy equalized odds? (threshold: TPR and FPR within 5 percentage points)
- 3 Does this model satisfy calibration? (threshold: within 3 percentage points per score bin)
- 4 Given the impossibility theorem, explain why all three metrics cannot be improved simultaneously
- 5 If you were the Chief Risk Officer, which metric would you prioritize and why?

Exercise 4: SHAP Interpretation for Bias Detection

A SHAP analysis of a denied loan application produces the following feature contributions:

Feature	Feature Value	SHAP Value
Debt-to-income ratio	0.45	+0.18 (toward denial)
Zip code	60619 (South Side Chicago)	+0.12 (toward denial)
Years of credit history	3 years	+0.09 (toward denial)
Annual income	\$52,000	+0.05 (toward denial)
Number of open accounts	4	-0.03 (toward approval)
Employment length	5 years	-0.06 (toward approval)

Tasks:

- 1 Which feature contributed most to the denial? Is this a legitimate risk factor?
- 2 Zip code 60619 is a predominantly Black neighborhood. What fairness concern does this raise?
- 3 Draft the adverse action notice for this applicant, listing the top 3 reasons for denial as required by ECOA
- 4 Propose a test to determine whether zip code is acting as a racial proxy in this model

Exercise 5: Counterfactual Fairness Analysis

Consider a loan applicant with the following characteristics:

- Gender: Female, Age: 34, Income: \$58,000, Occupation: Nurse
- Education: Bachelor's degree, Credit score: 710, Debt-to-income: 0.32
- Model prediction: **Denied** (probability of default: 0.42)

Tasks:

- 1 To test counterfactual fairness, you flip gender to Male while holding all other features constant. The model now predicts: Approved (probability of default: 0.28). Is this model counterfactually fair? Why or why not?
- 2 A colleague argues: "Income should also change when you flip gender, because men earn more on average. The counterfactual should use \$68,000." Do you agree? What does this depend on?
- 3 Using a causal graph, distinguish between features that are (a) caused by gender, (b) correlated with but not caused by gender, and (c) independent of gender
- 4 Should the model be allowed to use "occupation" if occupations are gender-segregated?

Exercise 6: EU AI Act Compliance

Scenario: A European digital bank uses an XGBoost model with 200 features for automated credit decisions. It processes 50,000 applications per month.

Tasks:

- 1 Under the EU AI Act, what risk category does this system fall into? What are the consequences?
- 2 List 5 specific technical requirements the bank must implement for compliance
- 3 The model uses alternative data (social media activity, smartphone usage patterns). What additional data governance concerns does this raise under the AI Act?
- 4 Design a “human oversight” mechanism that satisfies the EU AI Act without slowing down the application process (50,000/month is approximately 1,700/day)
- 5 The bank operates in Germany, France, and Spain. Are there additional national requirements beyond the EU AI Act?

Exercise 7: Bias Mitigation Strategy Selection

A fintech lender discovers the following fairness issues in its credit model:

- Disparate impact ratio: 0.68 (below 0.80 threshold)
- False positive rate for minority applicants is 2.4x higher than for majority applicants
- The model heavily relies on “years at current address” (proxy for race via housing discrimination)
- Training data is 80% majority group, 20% minority group

Tasks:

- 1 For each of the three mitigation stages (pre-processing, in-processing, post-processing), propose one specific intervention and explain what it would address
- 2 Which combination of interventions would you recommend? Justify your choice
- 3 Estimate the likely impact on model accuracy. Is a 2–3% accuracy drop acceptable?
- 4 After implementing your mitigation strategy, what monitoring would you set up to ensure ongoing fairness?

Exercise 8: Case Analysis – Building a Fair Credit Model

You are building a new credit scoring model for a bank that operates in both the US and EU. You have access to 500,000 historical loan applications with outcomes.

Constraints:

- US: Must comply with ECOA, provide adverse action notices, pass four-fifths rule
- EU: Must comply with EU AI Act high-risk requirements
- Business: Model must maintain $AUC > 0.82$ (current model achieves 0.87)
- The training data reflects historical lending from 2015–2024

Tasks:

- 1 Design the complete fairness audit pipeline (pre-training, during training, post-deployment)
- 2 Which fairness metric would you optimize for, given that you operate in both jurisdictions?
- 3 How would you handle the historical bias in 2015–2024 data? (Consider: the data includes the COVID-19 period with government-backed forbearance programs)
- 4 Draft the model card for this system, including sections on intended use, fairness evaluation, known limitations, and monitoring plan