Lesson 47: ML Ethics in Finance

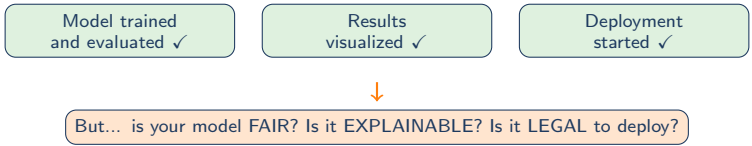
Data Science with Python – BSc Course

Data Science Program

BSc Course

45 Minutes

Previously in L46...



A model that works is not the same as a model that is safe to deploy

Learning Objectives

The Problem: ML models can perpetuate or amplify bias present in training data. A model that is accurate on average may be systematically unfair to specific groups.

After this lesson, you will be able to:

- Identify sources of bias in ML pipelines
- Measure fairness across demographic groups
- Use SHAP and LIME for model explainability
- Understand regulatory requirements (ECOA, GDPR, EU AI Act)
- Audit your own project for potential bias

Finance Application: Fair lending, credit scoring bias, algorithmic trading oversight

The Scenario That Changes Everything

Your model denies a loan.

The applicant is 22 years old, female, from a minority zip code. She has a steady income and no prior defaults.

Your model's prediction: **Denied. Default probability: 68%.**

Questions:

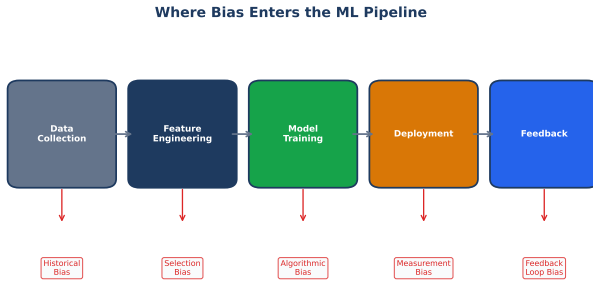
- Was this decision based on her creditworthiness – or her demographics?
- Can you PROVE which features drove the denial?
- Is your model using zip code as a proxy for race?
- Are you legally allowed to deploy this model?

This is not hypothetical. This happens every day with deployed credit models. The difference between a good data scientist and a dangerous one is whether they ask these questions.

Every prediction affects a real person. Act accordingly.

Bias in Data

Your Model Learns What Your Data Teaches



Historical data reflects historical discrimination – your model will learn to reproduce it

Algorithmic Fairness: The Core Concepts

What IS Algorithmic Bias?

- A model systematically treats some groups differently – without intent
- It learns patterns from data, including discriminatory patterns
- A model trained on biased lending decisions will deny loans the same way

Protected Attributes (Cannot Use for Decisions):

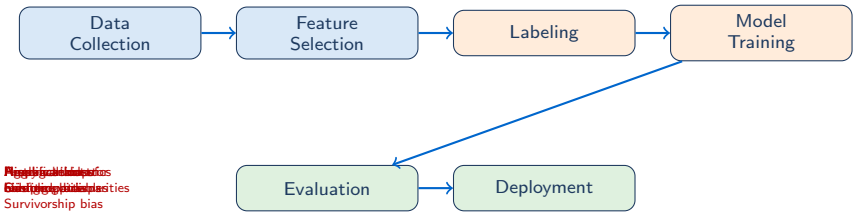
- Race, gender, age, religion, national origin, disability
- BUT: models can learn these indirectly through proxy variables

Proxy Variables (The Hidden Problem):

- ZIP code $\rightarrow$ correlates with race
- First name $\rightarrow$ correlates with gender and ethnicity
- University attended $\rightarrow$ correlates with socioeconomic status
- Removing the protected attribute is NOT enough

Bias is a data problem, not just an algorithm problem – garbage in, discrimination out

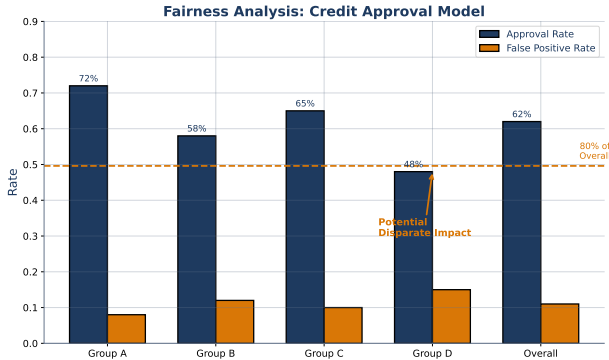
Types of Bias in ML Pipelines



Bias can enter at EVERY stage – checking only the model is not enough

Fairness Metrics

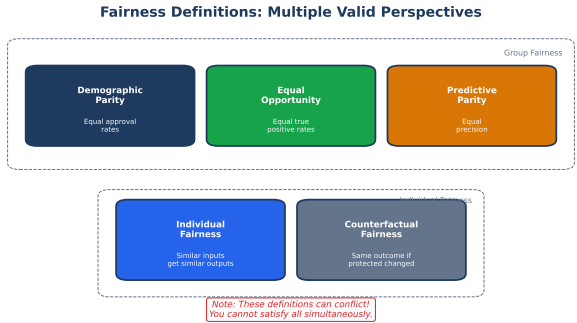
Measuring Fairness: The 80% Rule and Beyond



The four-fifths rule: no group below 80% of the highest group's approval rate

Fairness Definitions: They Conflict

You Cannot Have Everything



Impossibility theorem: you cannot satisfy all fairness criteria simultaneously (Chouldechova 2017)

Checkpoint: Equal Accuracy vs Equal Opportunity

Which Is Fairer?

Scenario: Your credit model has 85% accuracy overall. But:

- Group A (majority): 90% accuracy
- Group B (minority): 70% accuracy

Equal Accuracy says: Make both groups have 85% accuracy

- May require different thresholds per group
- Treats groups equally in outcomes

Equal Opportunity says: Equal true positive rates across groups

- A qualified applicant should have the same chance regardless of group
- Focuses on those who DESERVE approval

There is no single right answer. The choice depends on context, stakeholders, and regulation.

Fairness is not a formula – it is a value judgment that must be made explicitly

GDPR and EU AI Act Requirements

The Law Is Already Here

Regulation	Key Requirement
EU GDPR Art. 22	Right not to be subject to purely automated decisions
EU AI Act (2024)	High-risk AI needs transparency, human oversight
US ECOA	Must explain credit denial reasons
US Fair Housing Act	Cannot discriminate in mortgage lending
Basel III/IV	Model risk management for banks

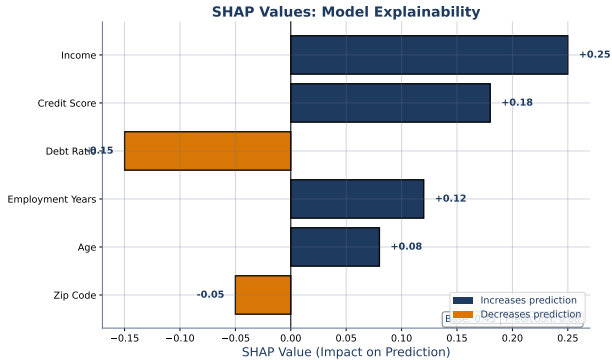
What This Means for Your Model:

- You MUST be able to explain individual predictions
- You MUST document your model's limitations
- You MUST maintain human-in-the-loop for high-stakes decisions
- You MUST keep audit trails of model versions and training data

"I did not know" is not a legal defense. Know the regulations before you deploy.

Explainability: SHAP Values

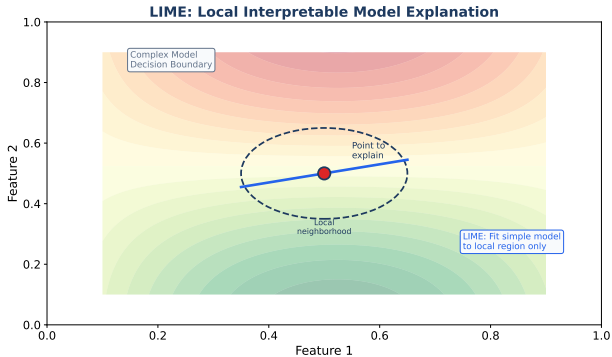
WHY Did the Model Say That?



SHAP assigns each feature a contribution score – essential for regulatory compliance

Explainability: LIME

Local Explanations for Any Model



LIME perturbs the input and fits a simple model locally – works with any black box

Model Documentation: Model Cards

The "Nutrition Label" for ML Models

A Model Card Should Include:

1. **Model Details:** Algorithm, version, training date, owner
2. **Intended Use:** What it was designed for (and what it was NOT)
3. **Training Data:** Source, size, demographics, known limitations
4. **Evaluation:** Metrics broken down by subgroup
5. **Ethical Considerations:** Known biases, potential harms
6. **Limitations:** When the model should NOT be trusted

Why Model Cards Matter:

- Prevents misuse (using a model outside its intended domain)
- Required by EU AI Act for high-risk applications
- Shows professionalism and accountability

If you cannot fill out a model card, you do not understand your model well enough

Finance: Credit Scoring Bias

The Most Common Real-World Example

How Bias Enters Credit Scoring:

- Historical data: Past loan officers denied minorities at higher rates
- Model learns: minority zip codes = higher default risk
- Reality: The historical denials were biased, not the applicants

The COMPAS Case (Criminal Justice):

- Recidivism prediction algorithm used in US courts
- ProPublica (2016): Black defendants scored higher risk than equivalent white defendants
- Model developers argued: "We optimize for accuracy"
- Critics: "Accuracy for whom?"

Lesson: Optimizing for overall accuracy can INCREASE group-level unfairness

---

In credit scoring, historical bias in data becomes automated discrimination in production

Finance: Algorithmic Trading Ethics

When Speed Meets Responsibility

Ethical Issues in Algorithmic Trading:

- **Market manipulation:** Spoofing, layering, wash trading
- **Flash crashes:** May 2010 – Dow dropped 1000 points in minutes
- **Information asymmetry:** HFT firms profit at retail investors' expense
- **Systemic risk:** Correlated algorithms amplify market crashes

Regulatory Response:

- MiFID II (EU): Algorithm testing, kill switches, audit trails
- SEC Rule 15c3-5: Pre-trade risk controls required
- Circuit breakers: Automatic trading halts on extreme moves

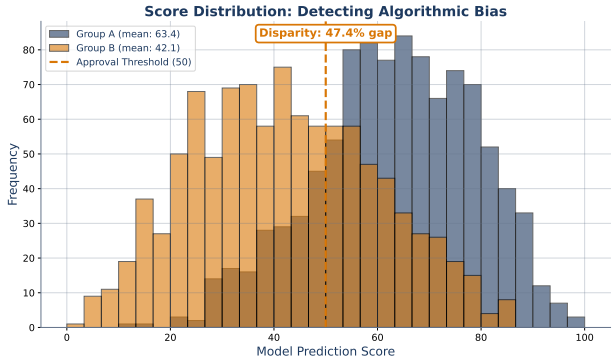
Your Responsibility: Even a student project must consider market impact

---

"The algorithm did it" is never an acceptable excuse – humans are responsible

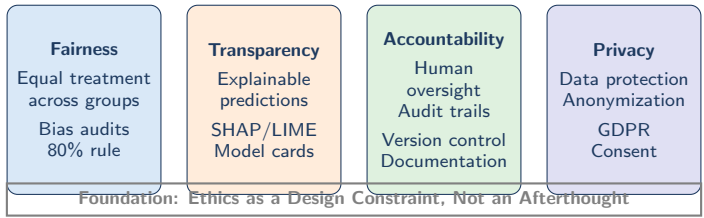
Environmental Cost of ML

Training Models Has a Carbon Footprint



GPT-3 training: 500 tonnes CO2. Do you need a billion parameters, or is logistic regression enough?

Responsible AI Framework



These four pillars should be part of every ML project from day one

Case Studies: When AI Goes Wrong

Real Failures You Should Know

Amazon Hiring Tool (2018):

- Trained on 10 years of hiring data (mostly male employees)
- Penalized resumes containing "women's" (e.g., "women's chess club")
- Amazon scrapped the tool entirely

Apple Card Gender Bias (2019):

- Gave men higher credit limits than women with same finances
- Goldman Sachs investigated by NY regulators
- "The algorithm is not biased" – but the outcomes were

COMPAS Recidivism (2016):

- Black defendants falsely labeled high-risk at 2x the rate of white defendants
- Used in actual sentencing decisions in US courts

These are not edge cases – they are predictable outcomes of ignoring bias

Bias Mitigation Strategies

Three Points of Intervention

1. Pre-processing (Fix the Data)

- Rebalance training data across groups
- Remove or transform proxy variables
- Synthetic oversampling for underrepresented groups

2. In-processing (Fix the Model)

- Add fairness constraints to the loss function
- Adversarial debiasing (model cannot predict protected attributes)
- Use `fairlearn` library for fairness-aware training

3. Post-processing (Fix the Output)

- Adjust thresholds per group to equalize outcomes
- Calibrate probabilities across groups
- Trade-off: may reduce overall accuracy slightly

No single approach is perfect – combine multiple strategies for robust fairness

Hands-On: Bias Audit (25 min)

Task: Audit a Credit Scoring Model for Bias

1. Load a credit dataset (e.g., German Credit from UCI)
2. Train a Random Forest classifier for credit approval
3. Compute approval rates by age group and gender
4. Check the 80% rule: Any group below 80% of highest rate?
5. Use SHAP to explain 3 individual predictions

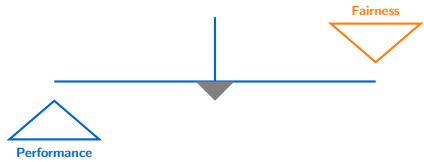
Deliverable:

- Fairness metric table (approval rates by group)
- SHAP waterfall plot for one denied applicant
- 2-sentence recommendation: Is this model safe to deploy?

Extension: Try `fairlearn` to add demographic parity constraints

Apply this same audit to YOUR project before final presentation

With Great Power...



The goal is not maximum accuracy.

The goal is responsible prediction.

A biased model with 95% accuracy is worse than
a fair model with 85% accuracy.

With great predictive power comes great responsibility

Lesson Summary

Problem Solved: ML models can discriminate. Measuring and mitigating bias is essential before deployment.

Key Takeaways:

- Bias enters at data, feature, label, and evaluation stages
- Fairness metrics often conflict – choose based on context
- SHAP and LIME provide individual prediction explanations
- GDPR and EU AI Act REQUIRE explainability for high-risk AI
- Document your model with model cards

For Your Project:

- Add a fairness check to your evaluation
- Include SHAP or feature importance in your presentation
- Acknowledge potential biases in your limitations section

Next Session: Final Presentations (L48) – showtime

Memory: Fair $\neq$ equal outcomes. Fair = equal opportunity. Always explain WHY.

Looking Ahead: L48

Final Presentations – Showtime

- 5 minutes to present your project to the class
- Live demo of your deployed application
- Q&A: defend your model choices and results
- Course retrospective and next steps

Come Prepared With:

- 3–5 slides: Problem, Data, Model, Results, Demo
- Working demo (test it 5 minutes before)
- Backup screenshots in case demo fails
- Answers to: "Why this model?" and "Is it fair?"

This is your moment. You built something real. Be proud of it.