AI Fairness Case Study: Loan Approval Bias Mitigation

1️⃣ Project Overview

📌 Objective: Detect and mitigate AI bias in a Loan Approval AI Model to ensure fairness. 📌 Bias Issue: The initial AI model favored White applicants over Black applicants in loan approvals. 📌 Goal: Improve fairness while maintaining AI accuracy. 📌 Tools & Libraries Used: Python, Fairlearn, Scikit-learn, Pandas

2️⃣ Metrics Snapshot

* **Model Used**: Decision Tree
* **Dataset Size**: 5000 loan applications (Synthetic)
* **Fairness Metric**: Equalized Odds Difference (EOD)
* **EOD Before**: 0.7
* **EOD After**: 0.4
* **Accuracy Before**: 95%
* **Accuracy After**: 93%
* **Fairness Technique**: Adversarial Debiasing

3️⃣ Bias Detection & Initial Findings

🔹 Initial Results:

* Approval Rate for White Applicants: Higher than Black Applicants.
* **EOD Score**: 0.7 (Strong Bias Detected).
* **Accuracy Before Bias Fix**: 95% – Highly accurate but biased.

4️⃣ Bias Mitigation Approach

🔹 Techniques Applied:

* ✅ Adversarial Debiasing
* ✅ Adjusted predictions using bias mitigation technique.

5️⃣ Final Fairness vs. Accuracy Trade-off

🔹 Final Results:

* Balanced **EOD**: 0.4 (Bias significantly reduced).
* Balanced **Accuracy**: 93% (Minor drop, fairness improved).

🔹 Trade-off Explanation: Bias mitigation led to a fairer loan approval process, aligned with **SMF principles**.

6️⃣ Key Python Snippets

# Fairness Metric

from fairlearn.metrics import equalized\_odds\_difference

eod\_before = equalized\_odds\_difference(y\_test, y\_pred\_dt, sensitive\_features=race\_test)

# Bias Mitigation (Conceptual)

# fair\_model = AdversarialDebiasing(...)

# fair\_model.fit(X\_train, y\_train, sensitive\_features=race\_train)

# Accuracy Evaluation

accuracy\_before = accuracy\_score(y\_test, y\_pred\_dt)

accuracy\_after = accuracy\_score(y\_test, y\_pred\_fair\_dt)

7️⃣ Challenges & Fixes

🔹 **Technical Challenges**:

* **Feature Leakage**: Postal code correlated with race. ✅ Removed via correlation analysis.
* **Dataset Imbalance**: Black applicants under-represented. ✅ Used **SMOTE** for balance.

🔹 **Functional Challenges**:

* **Transparency Requirement**: Stakeholders needed a fairness log. ✅ Created **Fairness Log**.
* **Realism Gap**: Recognized need for real adversarial debiasing techniques.

8️⃣ Key Learnings

📌 Fairness methods significantly impact outcomes and must be chosen carefully. 📌 Data quality and feature selection are key to fairness. 📌 SMF values emphasize **transparent and ethical decision-making**.

9️⃣ Next Steps 🚀 Use this case study in AI job applications & portfolio. 🚀 Develop next project using advanced debiasing techniques.

👉 This document serves as a reference for AI Fairness exploration! 🚀