📄 AI Risk Report

Team Name: Dheepak AI Ethics Team
Project Title: Fair Loan Approval Model

---

1. Problem Overview

In this project, we were tasked with building a machine learning model to predict loan approval using a synthetic dataset. The goal extended beyond accuracy—we needed to audit the model for potential bias across sensitive demographic attributes like gender, race, disability status, and more.

This problem has real-world significance: biased loan approval systems can worsen economic inequalities. For instance, an unfair model could systematically deny loans to specific communities, reinforcing financial exclusion.

We were provided with a dataset containing features such as Gender, Race, Age, Income, Credit Score, Employment Type, and more. Among these, sensitive attributes like **Gender**, **Race**, **Disability_Status**, **Language_Proficiency**, and **Zip_Code_Group** were flagged for bias analysis.

---

2. Model Summary

We used a **Logistic Regression** classifier due to its interpretability and simplicity, making it easier to trace the impact of features on predictions.

Preprocessing Steps:

- Dropped missing values
- Encoded categorical variables using one-hot encoding
- Standardized numerical features

Performance Metrics:

- **Accuracy**: 0.653
- **Precision (Approved)**: 0.60
- **Recall (Approved)**: 0.51
- **F1-Score (Approved)**: 0.55

These metrics reflect moderate model performance, but more importantly, enabled us to analyze prediction patterns across demographic groups.

---

3. Bias Detection Process

We conducted group-level bias audits by comparing approval rates across demographic categories. We focused on detecting disparate approval rates using bar plots and statistical comparison.

Bias was evaluated based on **approval rate differences** between subgroups (e.g., male vs. female, fluent vs. limited language proficiency).

- Audits were done on **model outputs**.
- Group-level fairness audits were conducted using **approval rate averages**.

---

4. 📉Identified Bias Patterns

Bias Type	Affected Group	Evidence	Metric	Comment
Gender Bias	Female	Lower approval rate (27%)	Approval Rate	Males had 43% approval rate
Race Bias	Multiracial, Black groups	Notably lower approval rates	Approval Rate	Disparity evident in grouped plots
Disability Bias	Disabled applicants	Slightly lower approval rate	Approval Rate	Indicates potential indirect bias
Language Proficiency	Limited Proficiency	Lower approval than fluent speakers	Approval Rate	May reflect language-based disparity
Zip Code Group	Rural regions	Lower approvals than urban areas	Approval Rate	May correlate with economic factors

---

5. Visual Evidence

The following visuals were generated: - Bar charts for approval rates across: - Gender - Race - Disability_Status - Language_Proficiency - Zip_Code_Group

These plots clearly illustrated disparities in approval likelihood across subgroups. All are stored in `outputs/bias_plots/` and included in the submission.

---

6. Real-World Implications

If deployed without adjustments, our model may: - Disadvantage women and racial minorities in loan access - Penalize individuals with limited language fluency - Create unfair outcomes for rural applicants

This could lead to regulatory scrutiny, public distrust, and systemic harm. The model in its current form may **not pass a fairness audit** required under financial regulations.

---

7. Limitations & Reflections

• **What didn't work**: The model was sensitive to missing values and imbalance in certain categories.
• **What's next**:
• Add bias mitigation techniques (reweighing, adversarial debiasing)
• Evaluate with additional fairness metrics like Disparate Impact Ratio
• Incorporate SHAP for feature-level bias analysis

Lessons Learned:

• Fairness must be integrated throughout the ML pipeline, not just at the end.
• Simple models can still exhibit complex biases.
• Visual tools are powerful for bias communication.

Final Note: This project deepened our understanding of AI ethics and responsible model deployment. We hope our submission contributes to the goal of building fairer, more inclusive systems.