AI Risk Report

Project Title

FairLens: Exposing & Reducing Bias in Loan Approval Models
Team: Willy McClanathan (Solo)

1. Problem Overview

The challenge was to audit a machine learning model that predicts loan approvals and uncover any potential bias, particularly involving sensitive groups. I chose to focus on gender because it's a key factor often embedded with bias in financial systems. I used the provided dataset that simulates real world loan approvals. It includes attributes like gender, income, age, race, and employment type, some of which are sensitive or potentially discriminatory. The goal was not only to build a predictive model but also to analyze whether it treats different groups fairly and, if necessary, make improvements.

2. Model Summary

I chose Logistic Regression as my base model. It's interpretable, straightforward, and useful for observing how individual features affect the outcome. After encoding all categorical variables like Gender and Employment Type, I split the dataset into training and testing using an 80/20 ratio. The initial model had an accuracy of about 62.1%, with precision at 64.3% and recall at 61.3%.

To ensure fairness, I used SHAP to explain which features most influenced predictions. Then, I applied Fairlearn's Exponentiated Gradient algorithm to mitigate bias. After mitigation, accuracy stayed about the same (62.3%) but the fairness metrics improved.

3. Bias Detection Process

I primarily focused on gender bias. To detect it, I used SHAP plots to understand feature influence and then used Fairlearn's MetricFrame to evaluate selection rate, demographic parity difference, and equalized odds. Audits were performed before and after mitigation to compare how fairly the model treated different gender groups. The evaluations were done at the group level — for example, comparing how often males and females were approved.

4. Identified Bias Patterns

Before mitigation, the model had a Demographic Parity Difference (DPD) of 0.0726. That meant one gender group was being approved noticeably more often. There was also an Equalized Odds

Difference of 0.1095, meaning that the model's accuracy and false positive/negative rates were not consistent across gender. SHAP analysis showed that gender was being used as a predictive feature, which is problematic. After mitigation, DPD improved to 0.0298 and Equalized Odds improved to 0.0549, which is a clear improvement.

5. Visual Evidence

I included the following visuals in my submission:

- A SHAP beeswarm plot showing how features like income and gender affected predictions.
- A SHAP bar plot showing which features were most influential overall.
- A bar chart comparing approval rates across gender groups before and after bias mitigation.
- A summary table of all the fairness metrics including DPD and Equalized Odds.

6. Real World Implications

If used in the real world, the unmitigated model could cause serious harm by unfairly denying loans to certain gender groups. This not only affects individual lives but could also violate regulations and damage the trust people have in financial institutions. By mitigating bias and keeping the model interpretable with tools like SHAP and Fairlearn, I believe this version of the model would perform much better under ethical and regulatory review.

7. Limitations & Reflections

This project taught me a lot about fairness in machine learning. I mainly focused on gender due to time, but other biases likely exist in the dataset, like race or age, and those could be explored further. I would have liked to add confusion matrix breakdowns for each group and explore intersectional bias, like how gender and race interact. SHAP and Fairlearn were both super helpful but also tricky at first — using them made me realize how much more responsibility comes with building real world AI models.