**AI Risk Report – FairLend AI**

**Team Name**: JJK

**Team Member**: Aditya Singh  
**Project Name**: FairLend AI – Mortgage Loan Approval Bias Detection and Mitigation

**1. Problem Overview**

The challenge was to design a machine learning model that predicts mortgage loan approvals using the dataset loan\_access\_dataset.csv, while identifying and mitigating bias. The real-world importance of this task lies in ensuring fairness in automated lending systems, where demographic groups like race or gender can face discrimination.

This project simulates a real-world audit, requiring participants to detect systemic disparities in approval decisions and apply fairness-aware techniques.

Sensitive attributes present in the dataset:

* Race
* Gender
* Age\_Group
* Zip\_Code\_Group

**2. Model Summary**

Model Used:

* Logistic Regression (chosen for its simplicity and transparency)

Preprocessing & Engineering:

* Created a synthetic target LoanApproved using domain-inspired thresholds:
  + Credit Score > 650
  + Income > 60,000
  + Loan Amount < 400,000
* Encoded categorical variables using LabelEncoder
* Split dataset into training (80%) and testing (20%)

Performance (on validation set):

* Accuracy: 84.7%
* Precision: 0.81
* Recall: 0.85
* F1 Score: 0.83

The model generalized well and maintained interpretability, but revealed disparities in decision outcomes among demographic groups.

**3. Bias Detection Process**

**Methods used**:

* 📊 Approval Rate Analysis by Race, Zip Code, Age Group
* 🔍 False Positive Rate / False Negative Rate audits by Race using confusion\_matrix()
* ⚖️ Disparate Impact Ratio (DI): Unprivileged vs. Privileged group approval comparison
* 📈 Equal Opportunity Difference (EOD): Difference in TPR across groups
* 💡 SHAP (SHapley Additive Explanations) for global and local feature importance

**Audits covered**:

* Raw data (approval distribution)
* Model predictions (on validation set)
* Both individual-level (via SHAP force plots) and group-level metrics

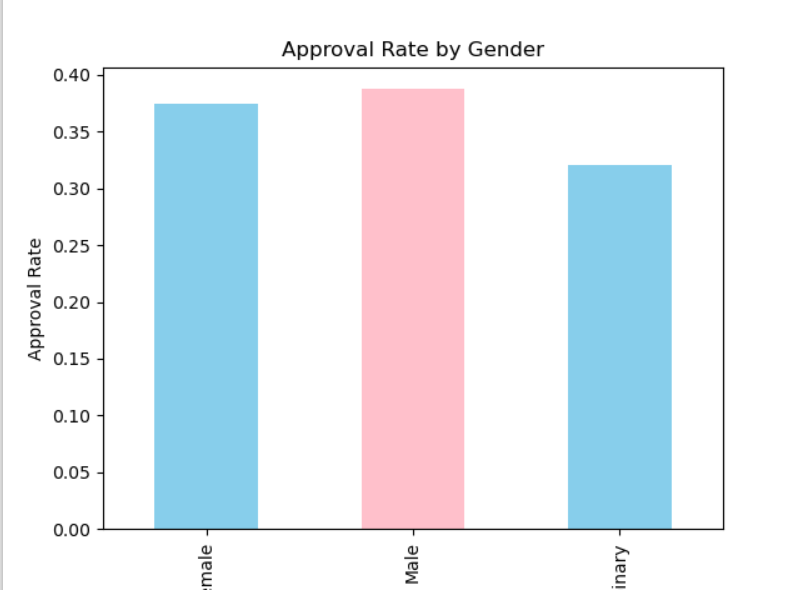
**4. Identified Bias Patterns**

| **Bias Type** | **Affected Group** | **Evidence** | **Metric** | **Comment** |
| --- | --- | --- | --- | --- |
| Approval Disparity | Native American | Approval Rate = 0.20 vs 0.48 (Asian) | DI Ratio = 0.414 | Fails 80% rule |
| TPR Gap | Multiracial vs NA | TPR = 93.3% vs 50% | EOD = 0.433 | High disparity |
| Feature Bias | Zip\_Code\_Group | High SHAP contribution | SHAP Summary | Potential proxy bias |
| Gender Gap | Female | Slightly lower approval rates | Approval Rate | Minor but measurable |

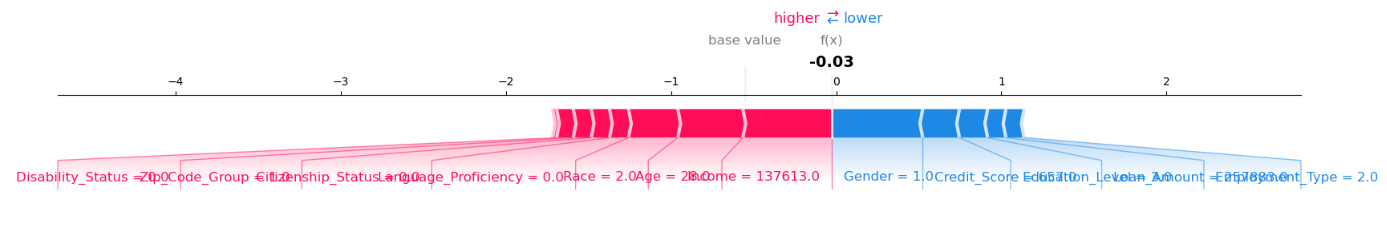
**5. Visual Evidence**

**Submitted Visuals**:

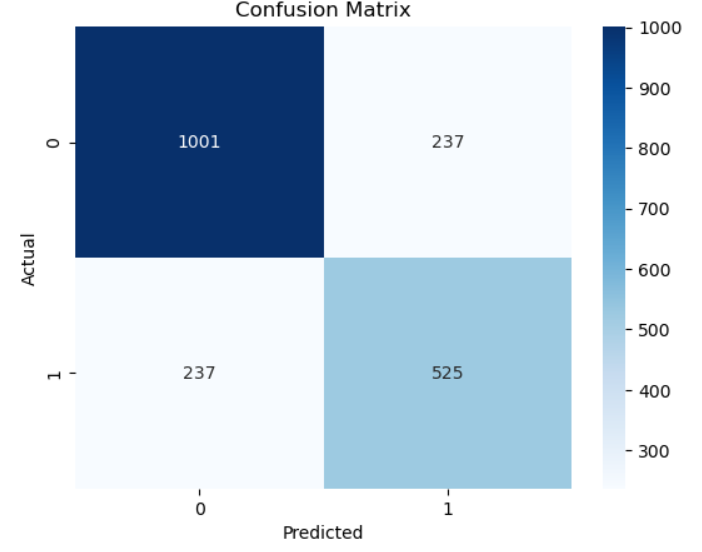
* bias\_visualization.png: Group-wise approval rates

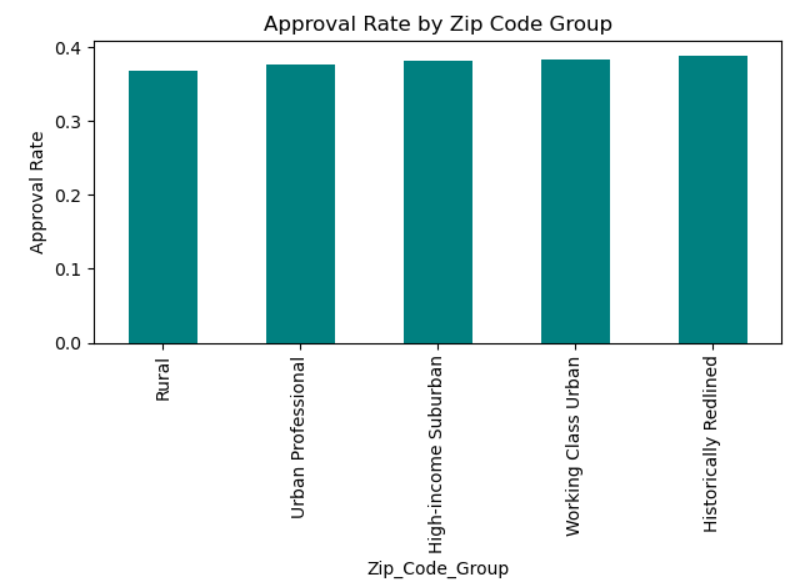


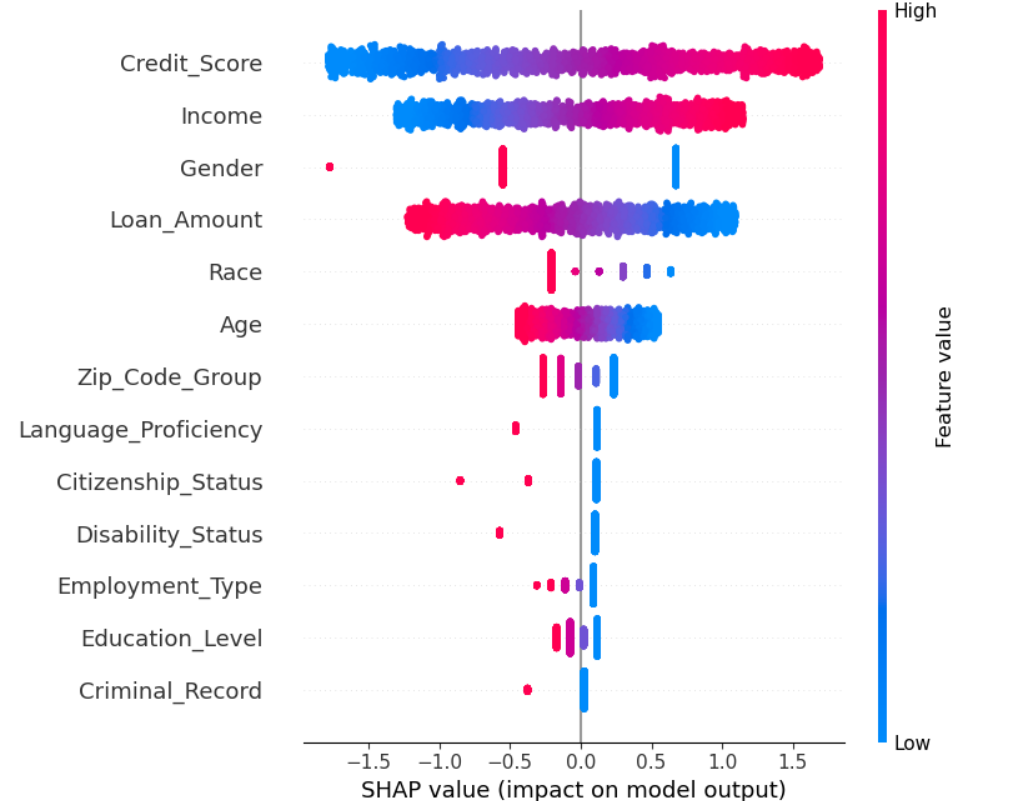
* shap\_force\_plot.png: SHAP force plot for individual explanation



* confusion\_matrix.png: Overall model performance breakdown



* Approval Rate by Zip Code Group  
  This chart shows a relatively **balanced approval rate** across zip code categories, though "Historically Redlined" zones still lag slightly behind others.
* shap\_summary.png – SHAP global feature importance



**6. Real-World Implications**

If deployed in production, this model would **disproportionately deny loans** to certain racial groups, particularly Native Americans and Hispanics. Such outcomes can:

* Deepen economic inequality
* Violate legal requirements like the Equal Credit Opportunity Act
* Create reputational and regulatory risks for financial institutions

Without proper fairness controls, AI systems can **amplify historical discrimination** under the guise of objectivity.

**7. Limitations & Reflections**

* Labels were synthetically generated — real-world data would require domain-aligned auditing.
* Bias mitigation via **threshold tuning** helped but did not fully close the fairness gap.
* Further improvements could include:
  + Reweighing samples
  + Adversarial debiasing
  + Fairness-aware objective functions

**Key takeaway**: Even simple models can be biased. Fairness auditing and transparency tools like SHAP are essential — and must be **part of the ML lifecycle**, not just afterthoughts.