**AI Risk Report: Loan Approval Fairness Analysis**

**Project Name:** Loan Approval Prediction with Bias Detection and Fairness Auditing

**Team:** Sreeprad

**1. Problem Overview**

The task involved developing a machine learning model to predict mortgage loan approvals using a comprehensive dataset of loan applications. This problem carries significant real-world weight because loan approval decisions directly impact people's ability to purchase homes, build wealth, and achieve financial stability. Biased lending practices have historically perpetuated economic inequality and denied opportunities to marginalized communities.

The dataset contained loan application records with various demographic and financial attributes. Key sensitive attributes identified for bias analysis included Gender, Race, Age Group, Citizenship Status, Disability Status, and Criminal Record. The target variable was binary, indicating whether each loan application was approved or denied, with a baseline approval rate of 43.1 percent across all applications.

Understanding and addressing bias in loan approval systems is critical because automated decision-making tools are increasingly used in financial services. If these systems exhibit discriminatory patterns, they can systematically disadvantage certain groups while appearing objective and data-driven.

**2. Model Summary**

When I first opened this dataset, it felt clear that any model I built needed an honesty check. So I treated this as my own mini audit. I selected logistic regression as the primary model after evaluating multiple alternatives including LightGBM, Random Forest, XGBoost, and Neural Networks. The logistic regression approach achieved the best balance of performance and interpretability, with a test accuracy of 63.15 percent and an AUC of 0.6777. This model was chosen specifically because it provides transparent, interpretable results that are essential for fairness auditing.

The preprocessing pipeline included several key steps. I used KNN imputation to handle missing values, though the dataset had minimal missing data. Standard scaling was applied to normalize continuous features like Income, Credit Score, and Loan Amount. Categorical variables were encoded using one-hot encoding for multi-class features and binary encoding for sensitive attributes with two levels.

Feature selection was performed using SHAP values to identify the most influential predictors. This approach revealed that Credit Score and Income were the dominant factors, contributing 0.3991 and 0.3092 to mean absolute SHAP importance respectively. The final model used the top 20 features based on SHAP analysis, which included demographic variables like Gender and Race alongside financial metrics.

The model was configured with L1 regularization and balanced class weights to address the inherent class imbalance in loan approvals. Cross-validation showed consistent performance across different data splits, suggesting the model generalizes reasonably well within the given dataset constraints.

**3. Bias Detection Process**

The bias detection process employed multiple complementary approaches to ensure comprehensive coverage of potential fairness issues. I ran a SHAP LinearExplainer on my initial model to understand feature importance and identify which attributes most strongly influenced model predictions. This revealed that demographic characteristics like Gender and Race appeared among the top predictive features, raising immediate fairness concerns.

I conducted both univariate and intersectional fairness analyses. I started simple - just plotting each group’s approval rate on its own. That first bar chart already revealed a glaring gap between disabled and non‑disabled applicants. The univariate analysis examined approval rates across individual sensitive attributes, while intersectional analysis looked at combinations of attributes to identify compounding disadvantages. For intersectional analysis, I focused on groups with sufficient sample sizes (n greater than or equal to 30) to ensure statistical reliability.

The audit process included raw data analysis to understand baseline disparities in the dataset, followed by model output analysis to determine whether the trained model amplified, reduced, or maintained these disparities. I used Fairlearn's MetricFrame to calculate selection rates and false negative rates across different demographic groups, providing detailed metrics on model fairness.

Disparate Impact analysis using AIF360 was employed to quantify bias using the 80 percent rule, which considers disparate impact ratios between 0.8 and 1.25 as acceptable. This analysis covered all six sensitive attributes to provide a comprehensive fairness assessment.

**4. Identified Bias Patterns**

The fairness analysis revealed several concerning bias patterns that would make this model unsuitable for deployment without significant corrections.

**Criminal Record Discrimination:** The most severe bias appeared against applicants with criminal records, who faced approval rates of only 18.8 percent compared to 52.7 percent for those without criminal records. The disparate impact ratio of 2.812 far exceeded acceptable thresholds, indicating that individuals with criminal records were nearly three times less likely to receive loan approval.

**Gender-Based Disparities:** Male applicants received approval at a rate of 59.3 percent, significantly higher than female applicants at 42.3 percent and non-binary applicants at 27.0 percent. The disparate impact ratio of 1.400 exceeded the fair lending threshold, with SHAP analysis showing that being male contributed positively to approval decisions with an importance score of 0.1391.

**Disability Status Bias:** Applicants with disabilities faced substantial discrimination, with approval rates of 26.1 percent compared to 53.1 percent for non-disabled applicants. The disparate impact ratio of 2.036 indicated that disability status created a significant barrier to loan approval, even when controlling for other factors.

**Racial Disparities:** While less extreme than other categories, racial bias was still evident. White applicants had approval rates of 55.8 percent, while Black applicants achieved only 38.7 percent approval and Hispanic applicants’ 39.6 percent. Asian applicants fell in between at 47.4 percent, while Native American applicants surprisingly showed higher approval rates at 70.6 percent, though this may reflect small sample sizes. I noticed that Native Americans had under 100 cases while mostly Whites were applicants; so I took their numbers with a grain of salt.

The intersectional analysis revealed that individuals facing multiple disadvantages experienced compounded discrimination. For example, applicants with both disabilities and criminal records had approval rates of only 24.5 percent, representing the most severely disadvantaged group in the dataset.

**5. Visual Evidence**

The bias visualization plots clearly demonstrate the approval rate disparities across demographic groups. Bar charts showing approval rates by sensitive attributes reveal stark differences, with some groups experiencing approval rates less than half of their more privileged counterparts.

SHAP feature importance plots illustrate how demographic characteristics rank among the most influential predictors, alongside financial metrics like Credit Score and Income. The summary plots show that being male, being white, and lacking a criminal record all contribute positively to approval predictions, while their opposites contribute negatively.

Distribution analysis of predicted probabilities shows that the model tends to assign lower approval probabilities to members of disadvantaged groups, even when their financial qualifications might be similar to approved applicants from privileged groups.

**6. Real-World Implications**

If deployed without modification, this model would perpetuate and potentially amplify existing inequalities in lending practices. Women, particularly non-binary individuals, would face systematic disadvantage in accessing home loans. People with disabilities would encounter significant barriers to homeownership, potentially violating the Americans with Disabilities Act and fair housing regulations.

The severe bias against individuals with criminal records raises questions about whether past legal issues should influence lending decisions, especially given efforts toward criminal justice reform and reintegration. The racial disparities would likely trigger regulatory scrutiny under the Fair Housing Act and Equal Credit Opportunity Act.

From an ethical perspective, the model's bias patterns could contribute to wealth inequality by denying homeownership opportunities to already marginalized communities. This creates a cycle where disadvantaged groups remain unable to build equity through property ownership, further widening socioeconomic gaps.

The model would almost certainly fail regulatory fairness audits in its current state. Financial institutions deploying such a system could face significant legal liability, regulatory penalties, and reputational damage. The disparate impact ratios well above acceptable thresholds would likely trigger investigations from banking regulators and civil rights enforcement agencies.

**7. Limitations and Reflections**

Several limitations became apparent during this analysis. The dataset itself may contain historical biases that reflect past discriminatory lending practices, making it challenging to train a truly fair model without additional interventions. The binary nature of many sensitive attributes may not capture the full complexity of identity and discrimination.

The feature selection process, while based on predictive importance, may have inadvertently retained variables that serve as proxies for protected characteristics. Future work should explore more sophisticated bias mitigation techniques applied during the training process rather than only post-hoc analysis.

Given more time and resources, I would implement bias mitigation techniques such as adversarial debiasing, fair representation learning, or constraint-based optimization to reduce disparities during model training. I would also explore whether alternative algorithms or ensemble methods could achieve better fairness-accuracy tradeoffs.

One key lesson learned is that achieving algorithmic fairness requires intentional design choices throughout the entire machine learning pipeline, not just careful analysis after model training. The tension between predictive accuracy and fairness demands careful consideration of what constitutes fair treatment in lending contexts.

Another important reflection is that technical solutions alone cannot address the underlying social and economic factors that create disparities in loan approval rates. Addressing bias in AI systems must be coupled with broader efforts to promote economic equity and opportunity for all communities.