**Name**: **Halimat Olatunji**

**Project Name: Bias Detection and Audit in AI Loan Approvals**

**Problem Overview**

**Task**:

The task is to analyse a loan access dataset, develop AI models for bias, assess fairness and recommend suggestions to improve ethical and responsible AI development.

**Real-world Importance**

Loan approval algorithms can significantly impact people’s lives. It can determine people’s access to housing, education and business opportunities. However, the use of machine learning models trained on historical data has, on many occasions, maintained existing inequalities. This has disproportionately put some people based on gender, race, disability status, or other sensitive attributes at a disadvantage.

This project is of significant importance because it falls under the broader concerns of fairness in AI decision-making systems. It shows how biased training data or poor modelling can lead to unfair decisions for certain groups of individuals.

**Dataset overview and sensitive attributes**.

Train and test datasets were provided. The train contained a label – Loan\_Approved while the test had no label. The provided datasets included both predictive features (Age, Income, Credit Score, Employment Type, Loan Amount, Education Level) and sensitive attributes such as:

* Gender
* Race
* Disability Status
* Citizenship Status
* Language Proficiency
* Criminal Record
* Zip Code Group

**Model Summary**

**Models Used**

I applied the Random Forest Classifier for this task because:

* It is robust to overfitting on structured data
* It can handle categorical and numerical features without scaling
* It is also robust to outliers
* It has built-in feature importance insights which complements SHAP

I built two models:

1. **Sensitive Features Model**: Included all available features, including sensitive ones like Gender, Race and Disability Status.
2. **Non-Sensitive Model**: This was treated as the baseline, representing a fairer model that excludes demographic attributes. This allowed for a comparison with the sensitive model to evaluate how much sensitive data influences predictions and potentially introduce bias.

**Preprocessing and Feature Engineering**

* Numerical Features: They were passed through without applying any scaling since Random Forest handles numeric values well. Credit Score binned for visual analysis.
* Categorical Features: Encoded using OneHotEncoder to enable Random Forest to process them.
* Feature Selection: For the non-sensitive model, sensitive attributes were excluded.
* Features such as Age Group, a derived form of Age, was dropped to avoid redundant information. IDs and temporary features created for exploration were also dropped to ensure only meaningful predictors were used.

**Hyperparameters**

* n\_estimators=200
* max\_depth=10
* class\_weight=’balanced’
* random\_state=42

**Model Performance (on Validation Set)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **AUC** |
| Sensitive  Features  Model | 0.62 | 0.56 | 0.62 | 0.67 |
| Non-Sensitive Model (Baseline) | 0.61 | 0.55 | 0.64 | 0.65 |

**Bias Detection Process**

Bias was evaluated through:

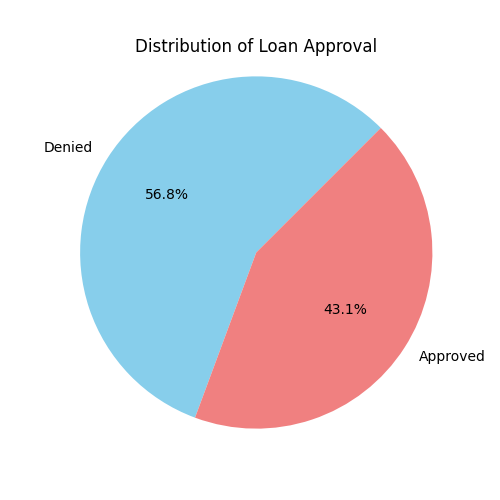
* Comparison on approval rate across groups
* Model interpretability using SHAP to evaluate feature importance
* False Positive Rate and False Negative Rate by group

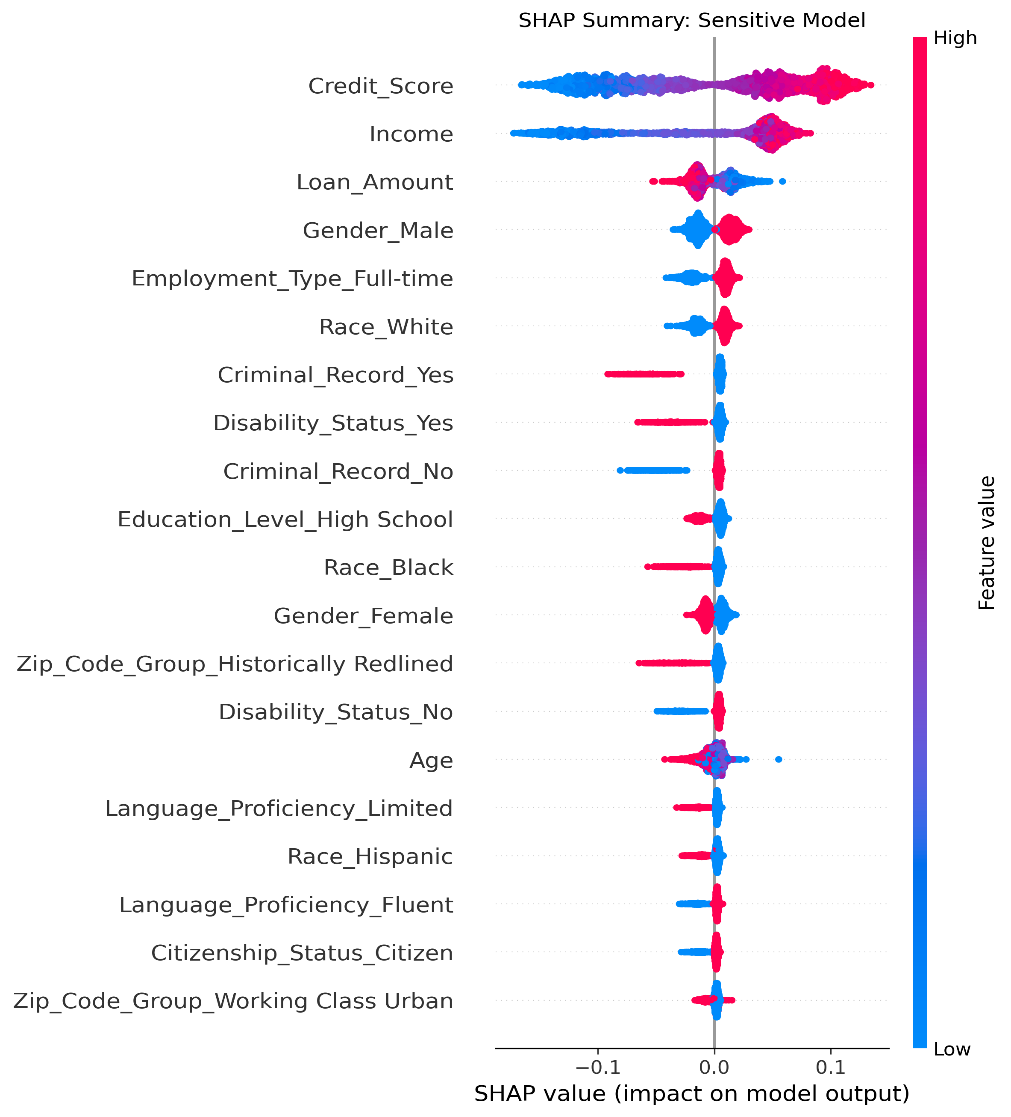
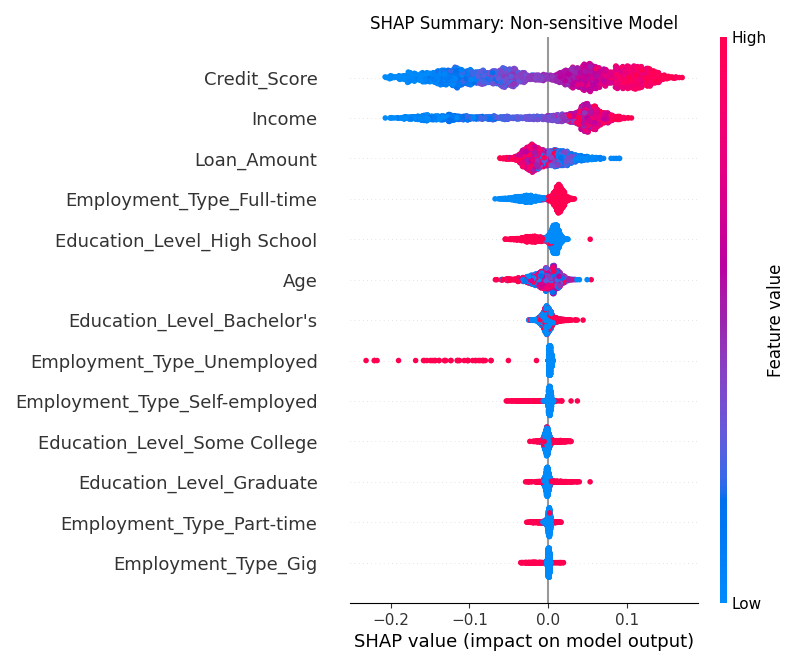
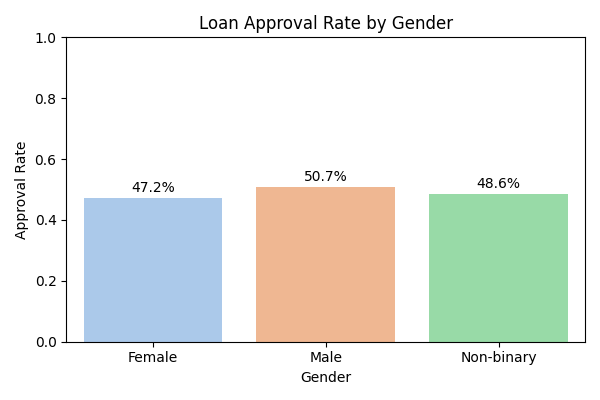
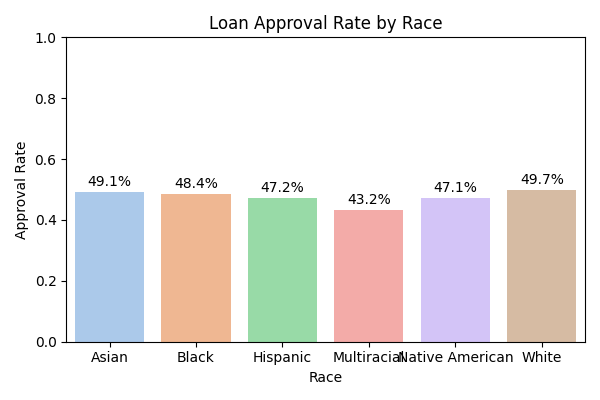
Audits were performed on raw data and model outputs.

**Identified Bias Patterns**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bias Type | Affected Group | Evidence | Metric | Comment |
| Approval Rate Disparity | Gender (Female) | Approval rate for Female (47.2%) lower than Male (50.7%) | Approval Rate | This suggests that females have lower approval rate |
| Approval Rate Disparity | Race | Multiracial -43.2%, Native American – 47.1%, lower than White (49.7%) | Approval Rate | There are disparities in race-based approval. |
| False Positive Disparity | Race (Black) | Highest FPR: 0.319 | FPR | The model is approving a high proportion of black applicants who should have been denied |
| False Negative Disparity | Multiracial, Native American | Highest FNR;  0.300, 0.250 | FNR | More Multiracial and Native American applicants who deserved approval were rejected. |
| False Negative Disparity | Gender (Female) | Highest FNR;0.225 | FNR | More qualified females are wrongly denied. |

**Visual Evidence**





**Real-World Implications**

* If the model were deployed as is, the groups at most risk will be Females, Black, Multiracial, and Native American Individuals.

**Ethical and social consequences**

* These biases could lead to denial in financial access for historically marginalized groups, increasing poverty and exclusion.
* These biases could also lead to people losing trust in AI
* Models learning bias from historical data can also perpetuate stereotypes.
* Lastly, it can cause emotional distress for individuals and their families.

In its current state, my model would likely fail a fairness audit in a regulated setting.

**Limitations and Reflections**

**What didn’t work?**

* I had an assumption that removing sensitive attributes will totally eliminate bias but I learned other correlated features may also be contributing to bias.
* Time didn’t permit me to try adversarial debiasing.

**Next time, with more data, I would try**

* Bias mitigation algorithms, and
* Explore bias mitigation techniques during every stage, from preprocessing to modelling.

**Lessons Learned**

* Bias audits must not be limited to removing sensitive attributes. Other methods should be explored.
* False positive rates and false negative rates provide more nuance than approval rate.
* It is essential to build transparent AI systems in order to implement responsible AI.