**AI Bias Bounty Hackathon Report**

Project Name: Loan Approval Bias Detection

Team Name : Anti-bias

## **Problem Overview**

The Objective of this project was to analyze potential bias in AI driven loan approval processes. The task was to build a model capable of predicting loan approval decisions and audit the model detecting and documenting the fairness risks.

This problem is critical in a real world context because biased AI systems in financial sectors can lead to systematic discrimination against underrepresented groups. Such biases decrease trust in automated systems and can result in ethical and legal violations.

The Provided dataset were: **loan\_access\_dataset.csv:** A labeled dataset simulating loan decisions **test.csv:** An unlabeled dataset for prediction **Sensitive attributes included:**

* Gender
* Race
* Age Group
* Employment Type
* Education level
* Citizenship status
* Zip Code Group

**Model Summary:**We used an XGBoost Classifier for its balance between interpretability and performance in structured data and its ability to capture non linear relationships which is crucial for this problem.

Key processing and Engineering Steps:

* Handled missing values(Mean for Numeric columns, Mode for Categorical)
* One hot encoded categorical variables
* Binary encoded (Yes/No fields)
* Removed ID and Age(an age-group feature already existed making this somewhat redundant)

Performance on Validation data:

* Accuracy: 0.61
* Precision: 0.57
* Recall: 0.45
* F1 Score: 0.50
* False Negative Rate: 0.55
* False Positive Rate: 0.26

## **Bias Detection Process**

* Group level approval rate comparisons and visualizations
* Disparate Impact Ratio
* Confusion matrix with FPR and FNR metrics
* SHAP Feature importance analysis

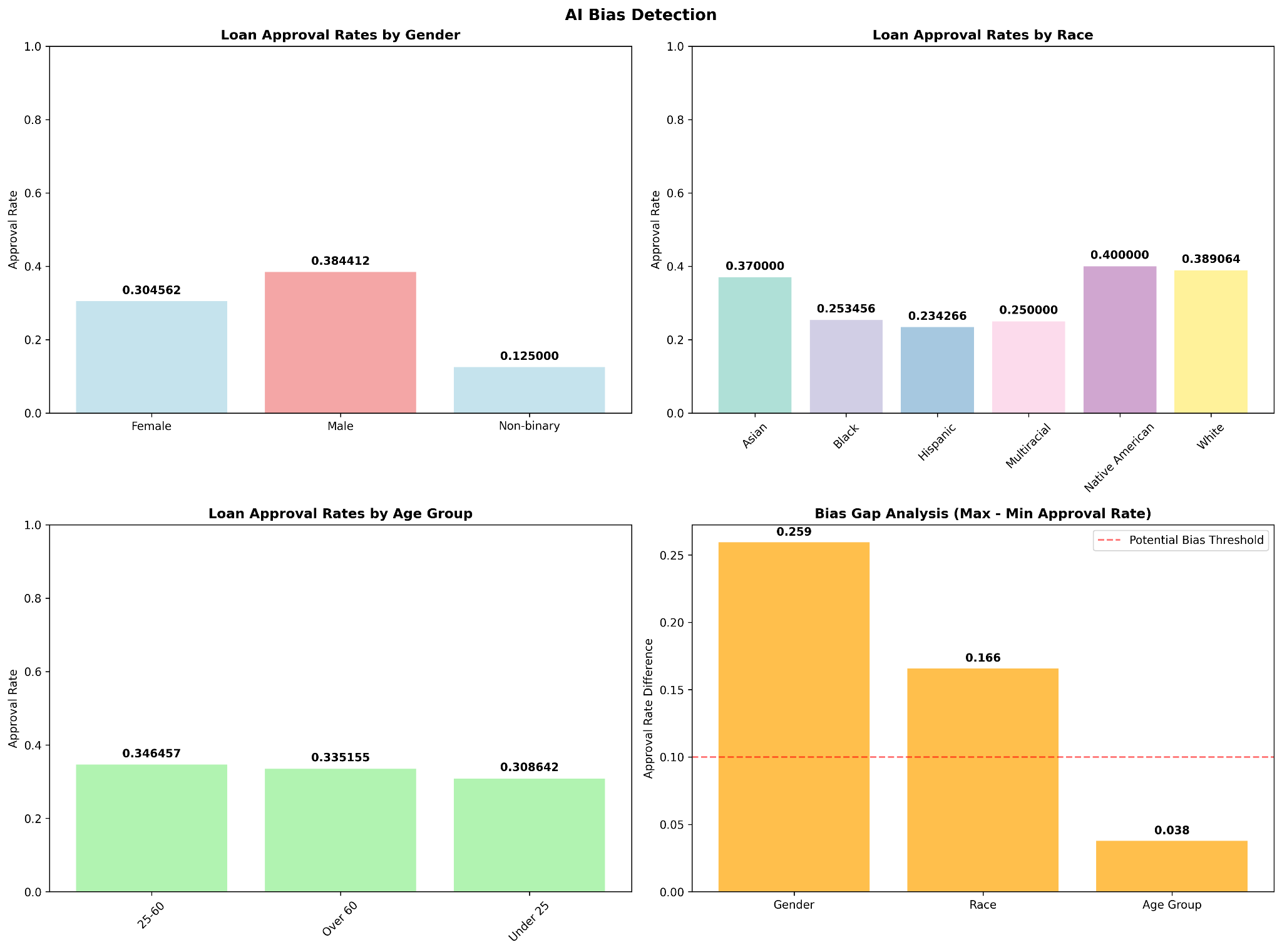
Audit Scope:

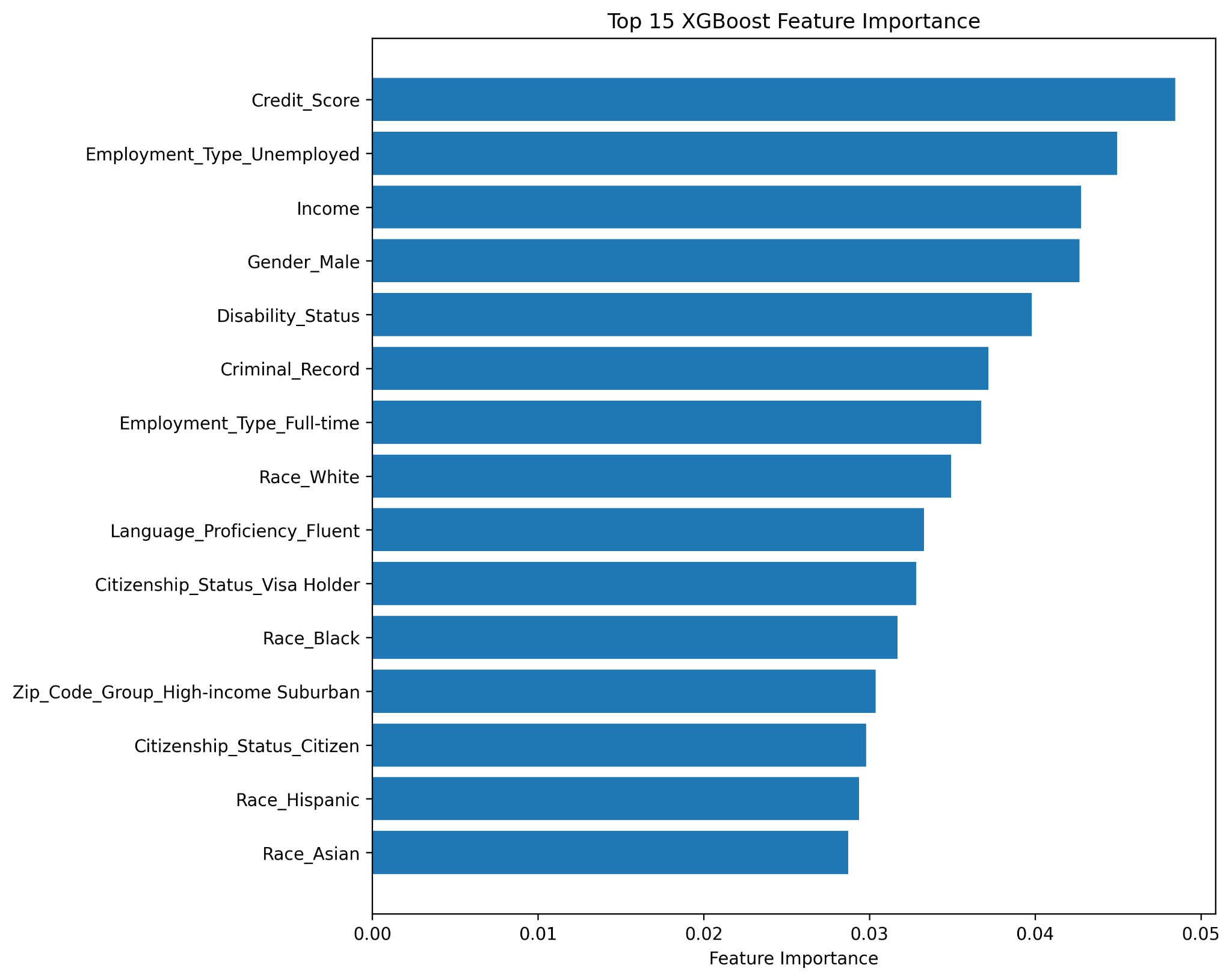
* Audited model Output on the test data
* Focused on group level bias patterns

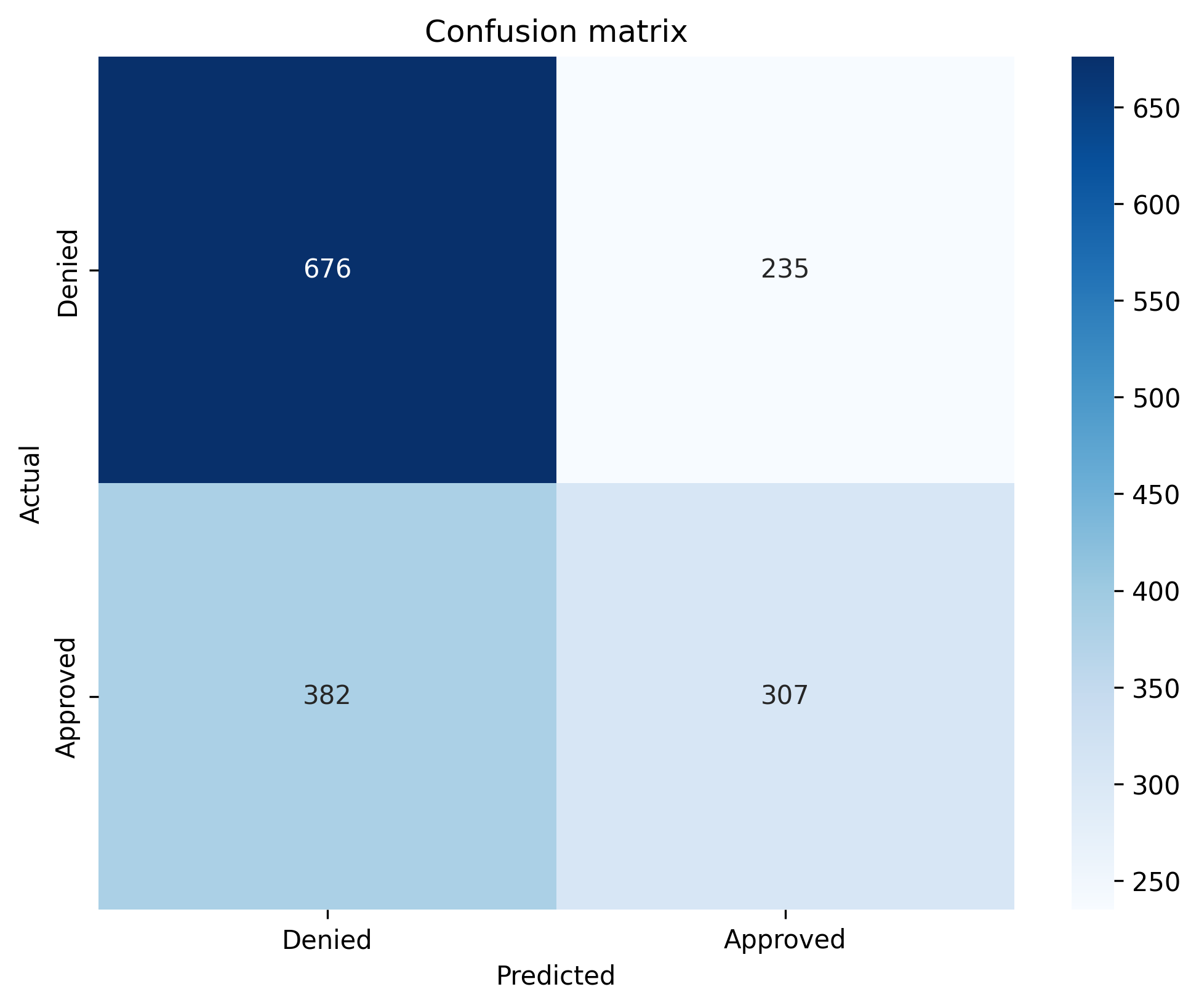
## **Identified Bias Pattern**

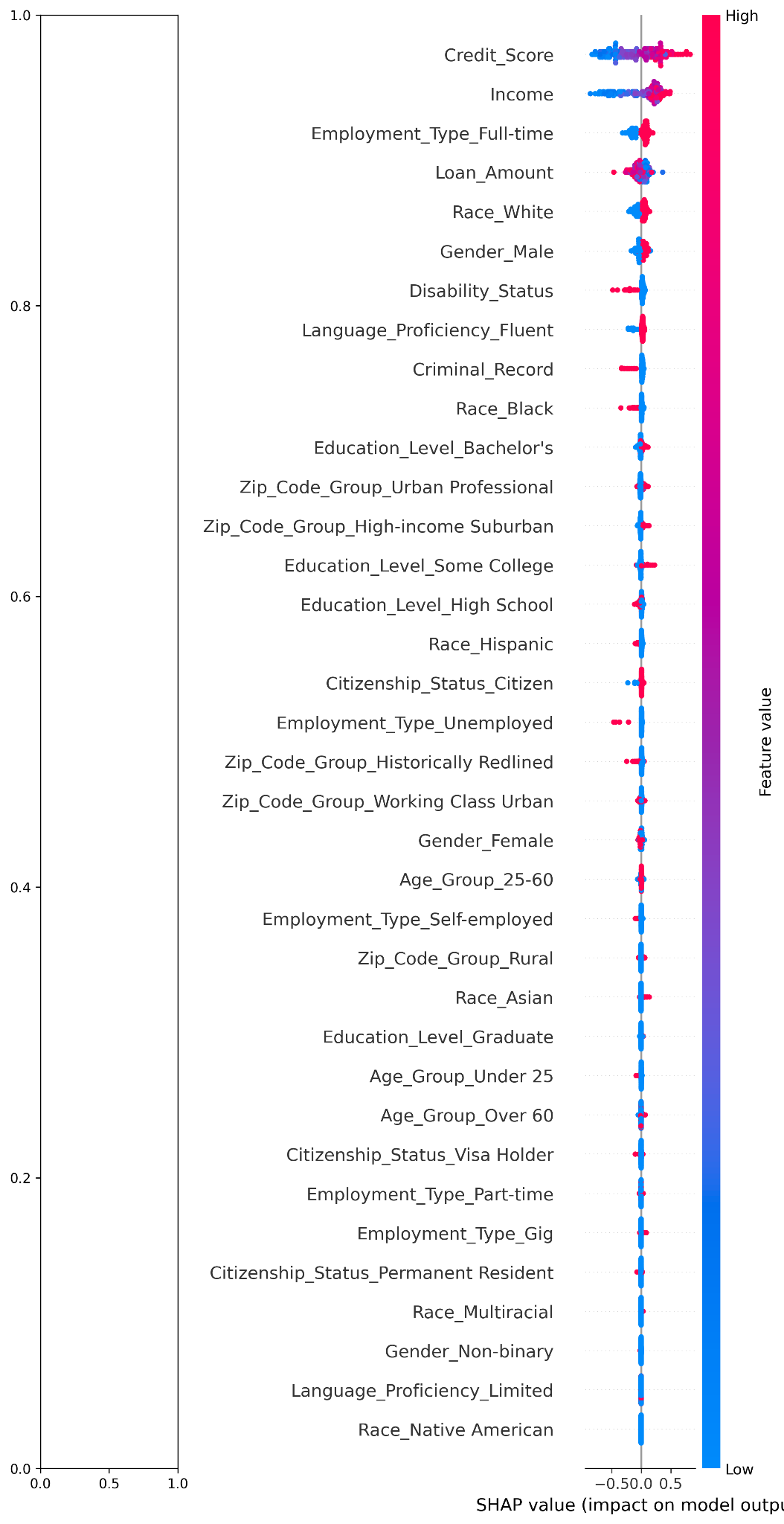
| Bias Type | Affected Group | Evidence | Metric | Comment |
| --- | --- | --- | --- | --- |
| Approval Rate Disparities | Gender | Male had 15% higher approval rate | Disparate Impact = 0.78 | Below 0.8 threshold, gender bias detected |
| Approval Rate Disparities | Race | Approval rates varied across racial groups | Disparate impact = 0.73 | Bias favouring certain racial groups |
| Geographic Bias | Zip Code Group | Significant variations in approval rates | Approval Rate Difference = 0.23 | Indicates geographic based discrimination |
| Education Level Disparities | Education Level | Lower rates for lower education levels | Approval Rate Difference = 0.18 | Suggest systematic bias against applicants with less education |

## **Visual Evidence**

BIAS VISUALIZATION   


FEATURE IMPORTANCE FROM THE XGBOOST MODEL  






SHAP ANALYSIS SUMMARY

**REAL WORLD IMPLICATIONS**  
  
If deployed as-is, the model could unfairly disadvantage: **Female applicants**, **Certain racial groups**, **Individuals from specific ZIP code group** and **Less educated applicants**

**Ethical & Social Consequences**Reinforces existing systematic inequalities, Decrease of trust in financial institutions and potential violations of anti-discrimination laws.

**Would this model pass a fairness audit in a regulated setting ?**This model would likely fail a regulatory fairness audit without proper mitigation steps and the use of synthetic data to train the models makes the model liable to see false patterns.

LIMITATIONS:

* No Bias Mitigation techniques (e.g reweighting, post-processing) were involved.
* Only approval metrics with the test data(16 percent of the initial data )were used to access bias, limiting scope of detection

WHAT WOULD WE TRY NEXT TIME WITH MORE TIME OR DATA:

* We would train multiple models architecture for comparison and perform cross validation tests
* Explore Bias mitigation techniques
* We could use a neural network with more hidden layers to capture more complex relationships that exists.

LESSONS LEARNED:

* Bias detection requires more than just model accuracy.
* Fairness audits should be a standardized part of AI model Validation