AI Risk Report

Bias Detection in Loan Approval Dataset

Team: Srimukha Sarma, Project Name: Bias Detection in Loan Approval Dataset

1. Problem Overview

- **What was the task? (e.g., Predict loan approval)**

  The task was to detect bias in a Loan approval dataset.
- **Why does it matter in a real-world or ethical context?**
  This helps us identify which groups in our society are being unfairly favoured. It also helps us identify where the problem lies and come to an effective solution.
- **What dataset were you given, and what were the known sensitive attributes?**

  I was given a dataset which had loan approval reports, The sensitive attributes are:

  'Gender', 'Race', 'Employment Type', 'Education Level', 'Citizenship Status', 'Disability Status', 'Criminal Record', 'Age', 'Age Group', 'Income', 'Credit Score', 'Zip Code Group', 'Language Proficiency', '

2. Model Summary

- **What model(s) did you use and why?**

  After training and evaluating multiple models (Logistic Regression, XGBClassfier, Random Forest Classifier and Custom neural network) I've decided to use the logistic regression model as it had the highest accuracy.
- **Key preprocessing, feature engineering, or hyperparameter choices**
  FEATURE ENGINEERING: I added 3 more input features (Income Group, Credit score group, loan amount group) to better evaluate the fairness metrics and I label encoded all the columns for training.
  Hyperparameters: I trained the logistic regression model for a maximum iterations of 1000.
- **Performance on internal validation data (accuracy, precision, recall, etc.)**

  The accuracy, precision, recall of the model are as follows: `Accuracy: 0.635, Precision: 0.6009104704097117, Recall: 0.45886442641946695`

3. Bias Detection Process

- **Methods used to detect bias (e.g., SHAP, Disparity Impact Ratio, False Positive Rate comparisons, Grouped AUC, Fairlearn audit, etc.)**

  I considered metrics like Equal Oppurtunities Difference, Average Odds Difference, Theil Index, Statistical Parity Difference and Disparate Impact to detect bias. I also used SHAP and Lime to check which of the features had the highest contributions to the models output. Then I checked the intersection fairness of two groups at a time to check if there was a certain combination which was being favoured.
- **Did you audit raw data, model output, or both?**
  I audited both raw data and model output.
- **Were audits performed at the individual or group level?**

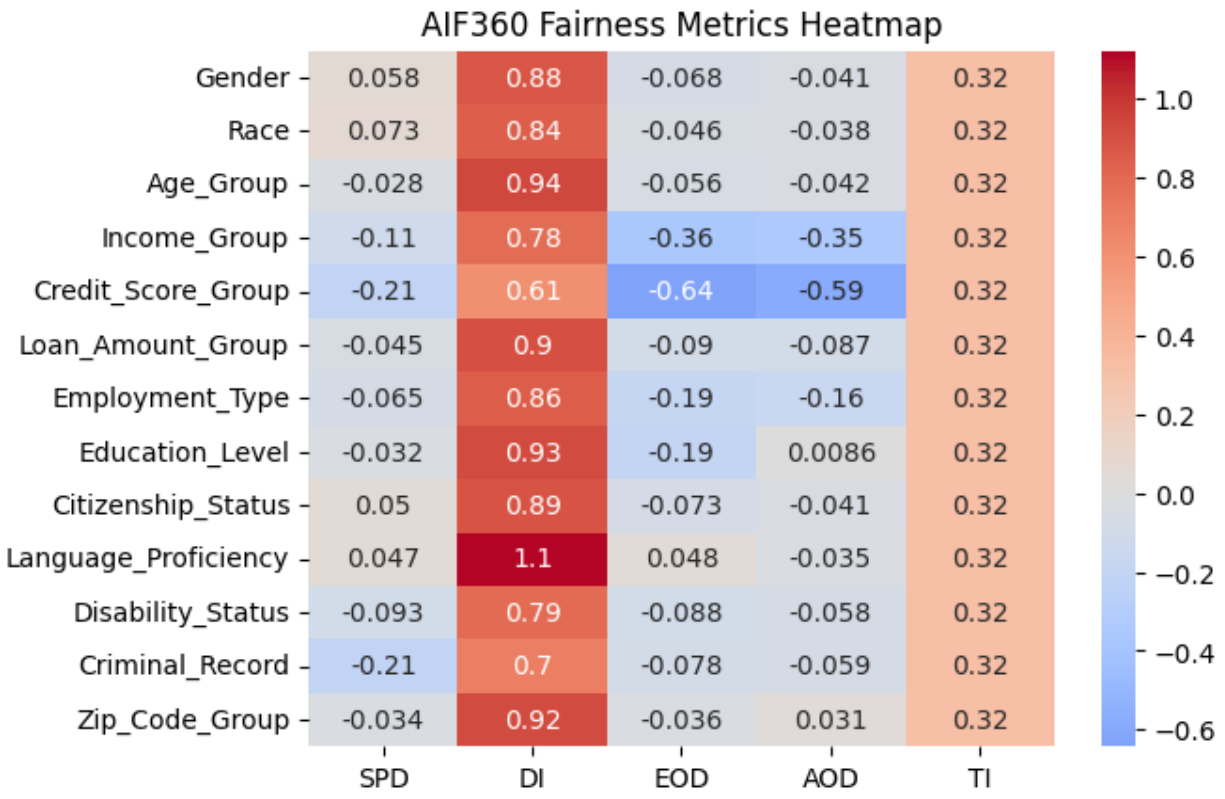
  Audits were performed at a group level.

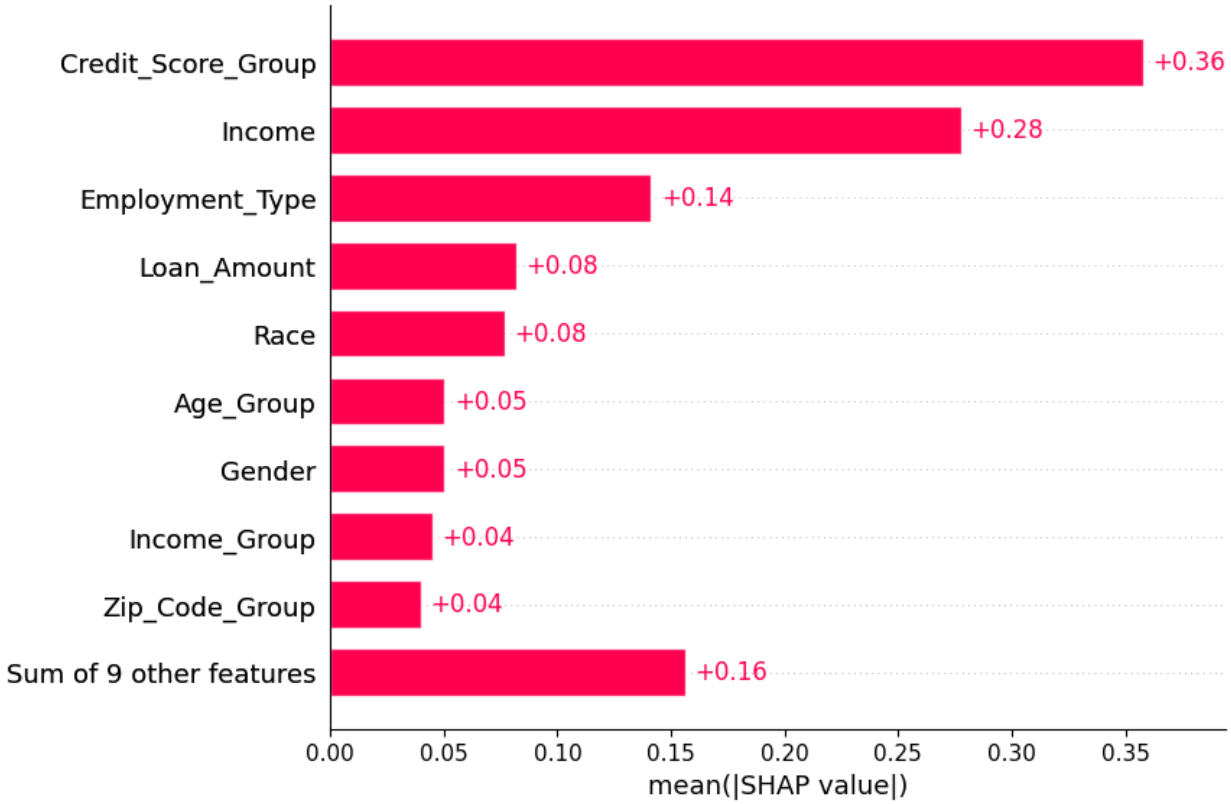
4. 📉 Identified Bias Patterns

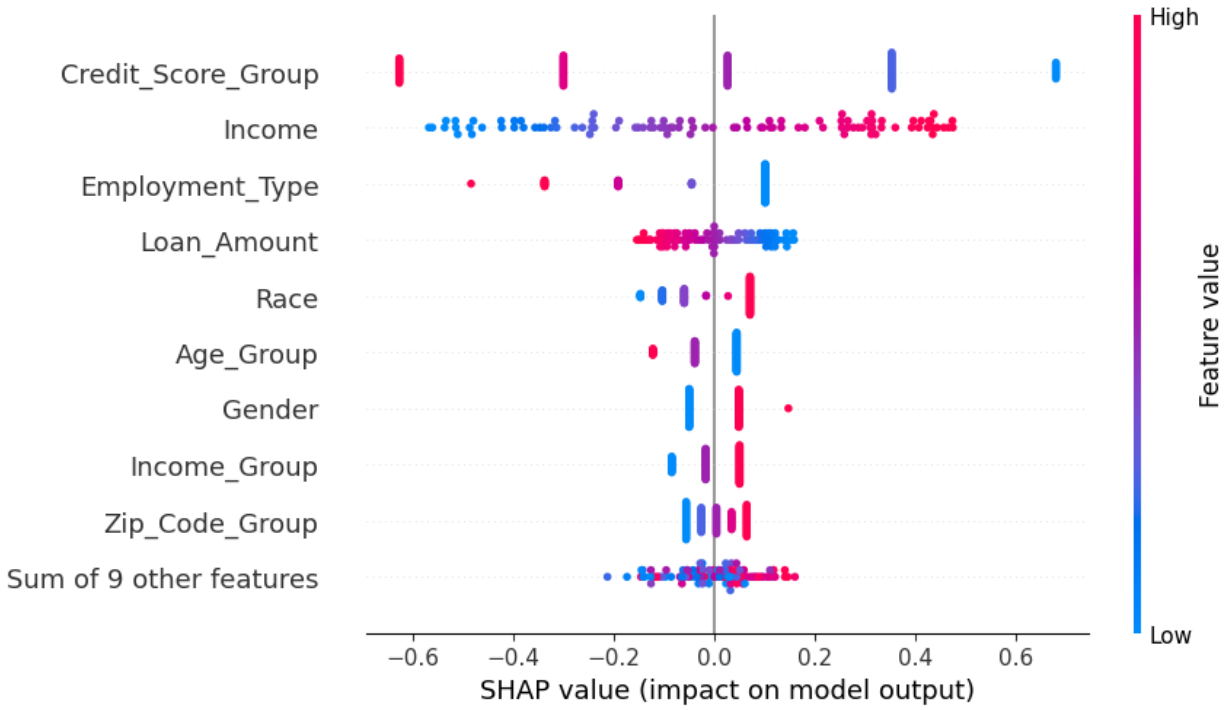
Bias Type	Affected Group	Evidence	Metric	Comment
Negative	Low Income Group	Aif360 heatmap	EOD,AOD,TI,SPD,DI	Theres a strong bias against low income groups.
Negative	Low Credit Score Group	Aif360 heatmap	EOD,AOD,TI,SPD,DI	Theres a strong bias against people with a low credit score.

Negative	People with Criminal Records	Aif360 heatmap	EOD,AOD,TI,SPD,DI	Theres a strong bias against people with criminal records.
Negative	People with Disabilites	Aif360 heatmap	EOD,AOD,TI,SPD,DI	Theres a moderate bias against people with no disabilites
Positive	Full-Time worker	Aif360 heatmap	EOD,AOD,TI,SPD,DI	Theres a moderate bias toward people with full time jobs.
Negative	High loan amount	Aif360 heatmap	EOD,AOD,TI,SPD,DI	Theres a mild bias against people taking a large loan.
Positive	People living in High Income Suburban and Urban Professional Areas	Aif360 heatmap	EOD,AOD,TI,SPD,DI	Theres a mild bias toward males.
Positive	Graduates and Bachelors	Aif360 heatmap	EOD,AOD,TI,SPD,DI	Theres a mild bias toward them
Negative	Multi-Racial people with low income	Intersection Fairness heatmap	Loan Approval rate	Theres a moderate bias against them
Negative	Multi-Racial people	Intersection Fairness heatmap	Loan Approval rate	Theres a moderate bias against them
Positive	25-60 year olds with high credit score	Intersection Fairness heatmap	Loan Approval rate	Theres a strong bias towards them.

5. Visual Evidence


AIF360 Fairness Metrics Heatmap
Gender 0.058 0.88 -0.068 -0.041 0.32
Race 0.073 0.84 -0.046 -0.038 0.32
Age_Group -0.028 0.94 -0.056 -0.042 0.32
Income_Group -0.11 0.78 -0.36 -0.35 0.32
Credit_Score_Group -0.21 0.61 -0.64 -0.59 0.32
Loan_Amount_Group -0.045 0.9 -0.09 -0.087 0.32
Employment_Type -0.065 0.86 -0.19 -0.16 0.32
Education_Level -0.032 0.93 -0.19 0.0086 0.32
Citizenship_Status 0.05 0.89 -0.073 -0.041 0.32
Language_Proficiency 0.047 1.1 0.048 -0.035 0.32
Disability_Status -0.093 0.79 -0.088 -0.058 0.32
Criminal_Record -0.21 0.7 -0.078 -0.059 0.32
Zip_Code_Group -0.034 0.92 -0.036 0.031 0.32
SPD
DI
EOD
AOD
TI
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Credit_Score_Group +0.36
Income +0.28
Employment_Type +0.14
Loan_Amount +0.08
Race +0.08
Age_Group +0.05
Gender +0.05
Income_Group +0.04
Zip_Code_Group +0.04
Sum of 9 other features +0.16
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
mean(|SHAP value|)



LIME RESULT

6. Real-World Implications

- **Who is most at risk if your model were deployed as-is?**

  All the groups which have been under a negative bias would suffer as the model is trained on a biased dataset.
- What are the ethical or social consequences?
- **Would your model pass a fairness audit in a regulated setting?**

No, the current model would likely not pass a fairness audit.

To pass an audit, the model would need:

- Bias mitigation applied in training or decision thresholds.
- Regular audits with fairness metrics.
- Transparent documentation explaining design decisions and their social implications.

7. Limitations & Reflections

- **What didn't work?**

  Purely basing the detection on approval rates and confusion matrices was very ineffective due to the data imbalances. Metrics like EOD,AOD,TI,SPD,DI worked better in detecting bias.
- **What would you try next time with more time or data?**
  I would try intersection fairness with many more groups and try fine-tuning my model more. I would also try out more metrics available in the IBM aif360 tool.
- **Any lessons learned on fairness or auditing?**

  I've learnt the different parameters which need to be considered for bias detection and I've gotten proficient with fairness tools like IBM aif360 tool.