AI Risk Report

Project Title

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', 'Loan Amount'

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: 0.626, Precision: 0.5833333333333334, Recall: 0.4652241112828439

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(EOD), Average Odds Difference(AOD), Theil Index(TI), Statistical Parity Difference(SPD) and Disparate Impact(DI) to detect bias. I also used SHAP and Lime to check which of the features had the highest contributions to the model's 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?**

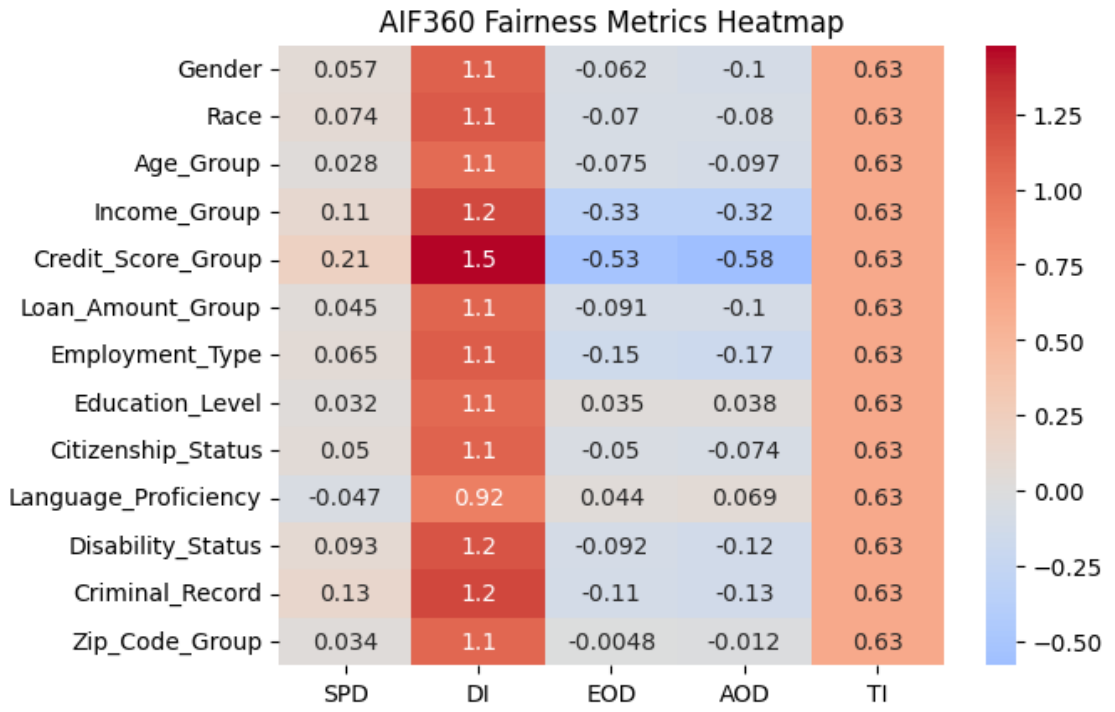
  Audits were performed at a group level and individual level.

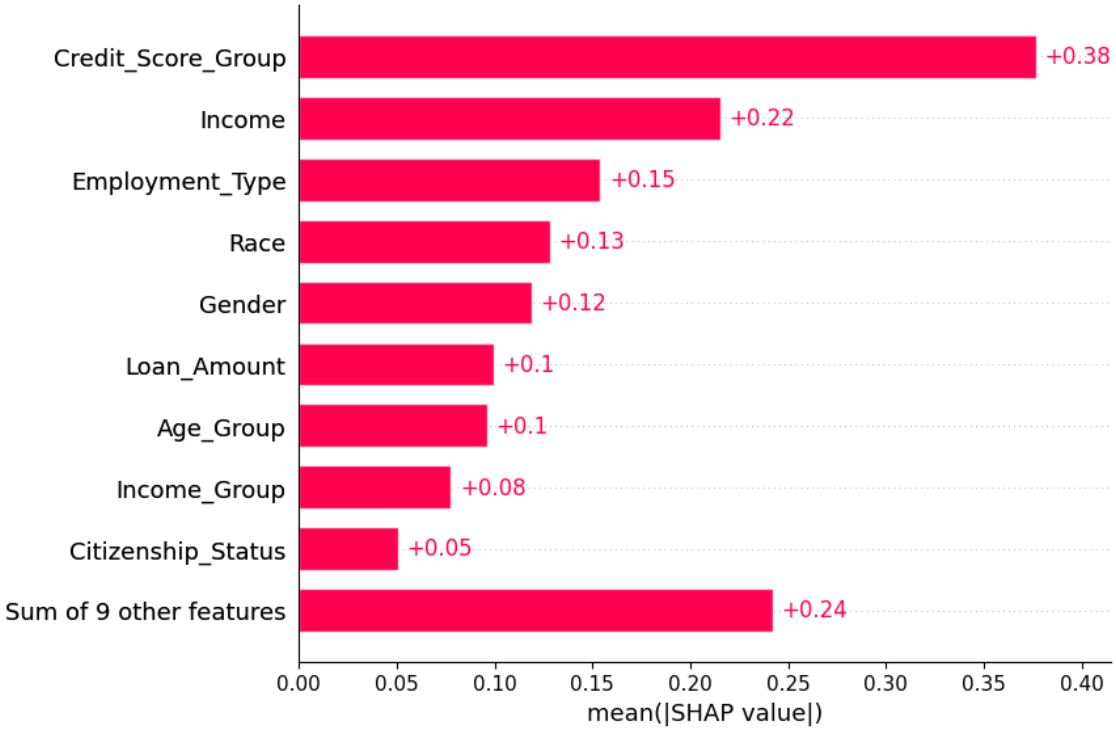
4. 📉 Identified Bias Patterns

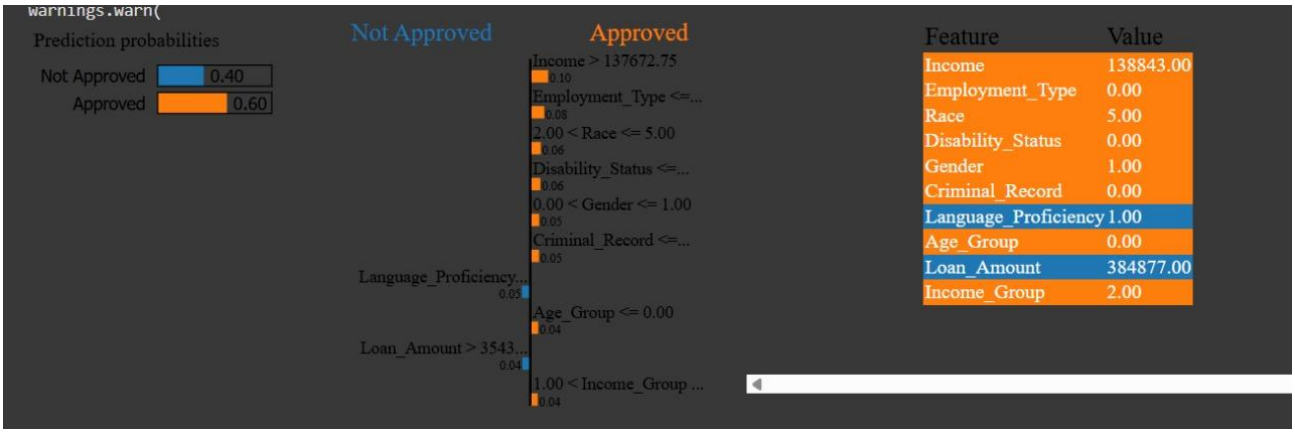
Bias Type	Affected Group	Evidence	Metric	Comment
Negative	Low Income Group	Aif360 heatmap	EOD,AOD,TI,SPD,DI	Low income groups have much lower approval rates
Negative	Low Credit Score Group	Aif360 heatmap	EOD,AOD,TI,SPD,DI	Theres a strong bias against people with a low credit score.

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 toward people taking a large loan.
Positive	Male	Aif360 heatmap	EOD,AOD,TI,SPD,DI	Theres a mild bias toward males.
Negative	Multi-Racial people with low income	Intersection Fairness heatmap	Loan Approval rate	Theres a moderate bias towards them
Positive	25-60 year olds with high credit score	Intersection Fairness heatmap	Loan Approval rate	Theres a strong bias towards them.

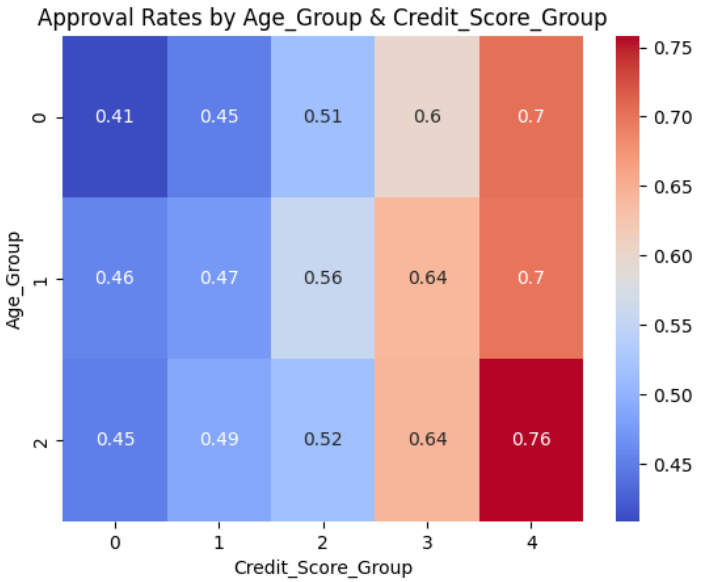
5. Visual Evidence

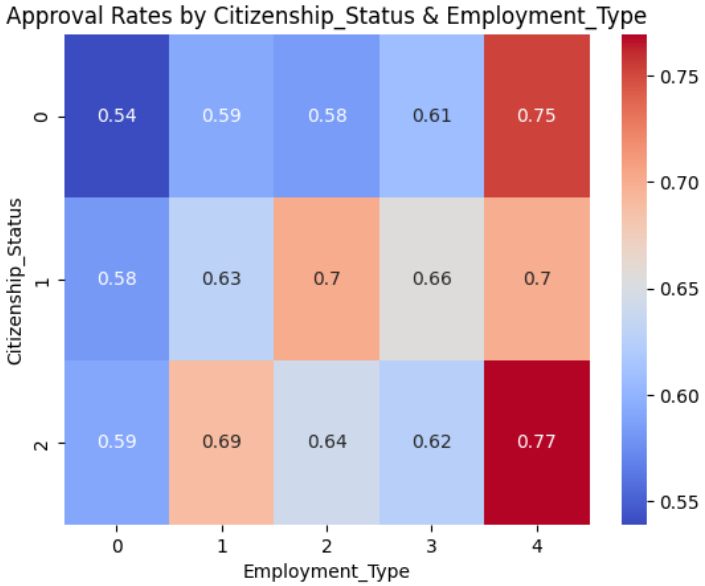

AIF360 Fairness Metrics Heatmap
Gender 0.057 1.1 -0.062 -0.1 0.63
Race 0.074 1.1 -0.07 -0.08 0.63
Age_Group 0.028 1.1 -0.075 -0.097 0.63
Income_Group 0.11 1.2 -0.33 -0.32 0.63
Credit_Score_Group 0.21 1.5 -0.53 -0.58 0.63
Loan_Amount_Group 0.045 1.1 -0.091 -0.1 0.63
Employment_Type 0.065 1.1 -0.15 -0.17 0.63
Education_Level 0.032 1.1 0.035 0.038 0.63
Citizenship_Status 0.05 1.1 -0.05 -0.074 0.63
Language_Proficiency -0.047 0.92 0.044 0.069 0.63
Disability_Status 0.093 1.2 -0.092 -0.12 0.63
Criminal_Record 0.13 1.2 -0.11 -0.13 0.63
Zip_Code_Group 0.034 1.1 -0.0048 -0.012 0.63
SPD
DI
EOD
AOD
TI
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50

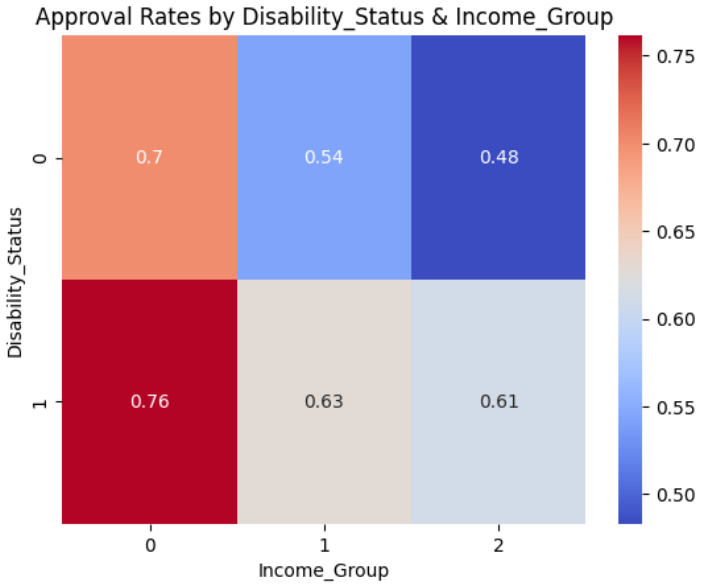

Credit_Score_Group +0.38
Income +0.22
Employment_Type +0.15
Race +0.13
Gender +0.12
Loan_Amount +0.1
Age_Group +0.1
Income_Group +0.08
Citizenship_Status +0.05
Sum of 9 other features +0.24
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
mean(|SHAP value|)

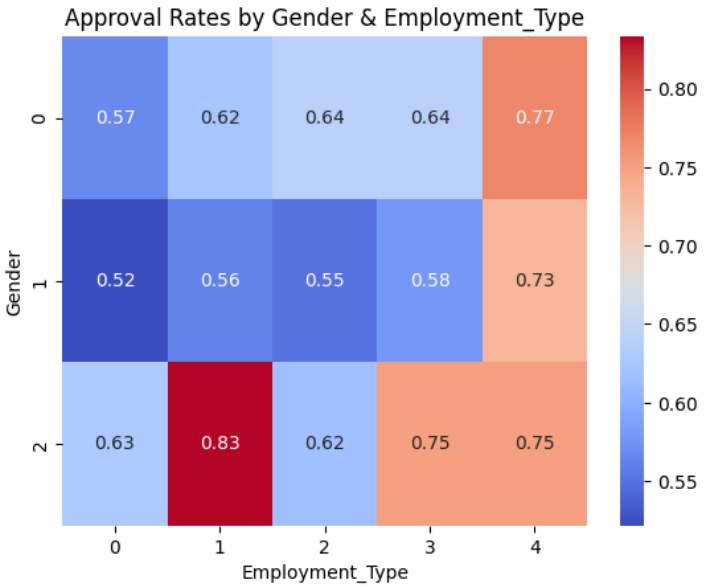

warnings.warn(
Prediction probabilities
Not Approved 0.40
Approved 0.60
Not Approved
Approved
Income > 137672.75
0.10
Employment_Type <=...
0.08
2.00 < Race <= 5.00
0.06
Disability_Status <=...
0.06
0.00 < Gender <= 1.00
0.05
Criminal_Record <=...
0.05
Language_Proficiency...
0.05
Age_Group <= 0.00
0.04
Loan_Amount > 3543...
0.04
1.00 < Income_Group ...
0.04
Feature
Value
Income 138843.00
Employment_Type 0.00
Race 5.00
Disability_Status 0.00
Gender 1.00
Criminal_Record 0.00
Language_Proficiency 1.00
Age_Group 0.00
Loan_Amount 384877.00
Income_Group 2.00

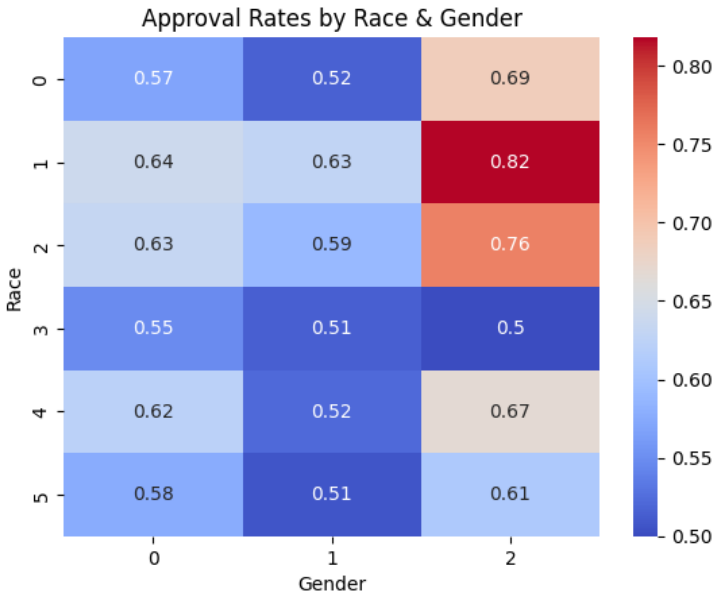
LIME RESULT

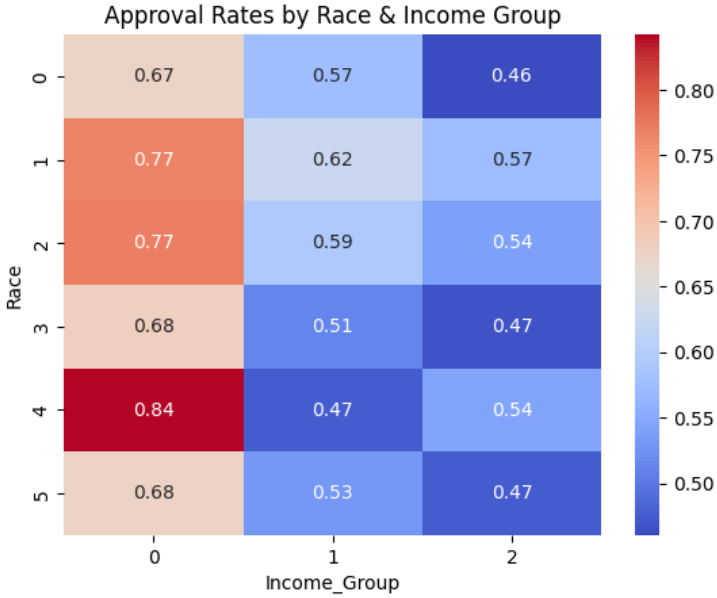

Approval Rates by Age_Group & Credit_Score_Group
Age_Group
Credit_Score_Group
0
1
2
0.41
0.45
0.51
0.6
0.7
0.46
0.47
0.56
0.64
0.7
0.45
0.49
0.52
0.64
0.76
0.75
0.70
0.65
0.60
0.55
0.50
0.45

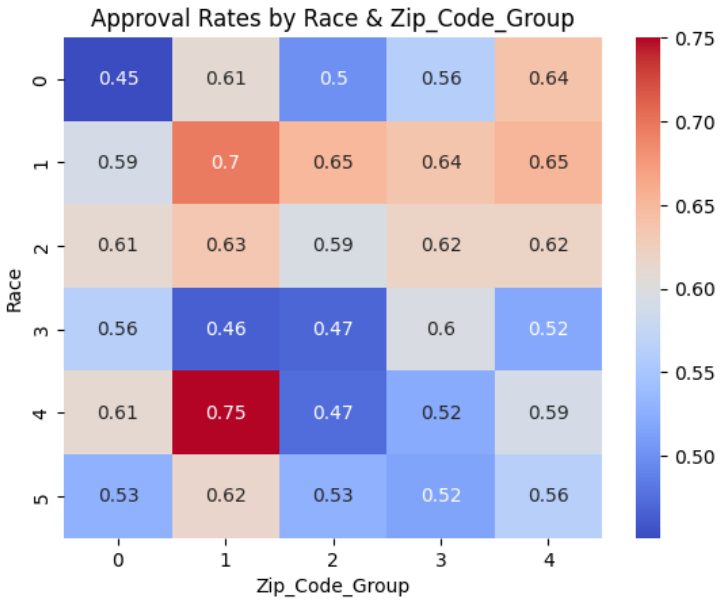

Approval Rates by Citizenship_Status & Employment_Type
Citizenship_Status
Employment_Type
0
1
2
0.54
0.59
0.58
0.61
0.75
0.58
0.63
0.7
0.66
0.7
0.59
0.69
0.64
0.62
0.77
0.75
0.70
0.65
0.60
0.55


Approval Rates by Disability_Status & Income_Group
Disability_Status
Income_Group
0
1
2
0.7
0.54
0.48
0.76
0.63
0.61
0.75
0.70
0.65
0.60
0.55
0.50


Approval Rates by Gender & Employment_Type
Gender
Employment_Type
0
1
2
3
4
0.57
0.62
0.64
0.64
0.77
0.52
0.56
0.55
0.58
0.73
0.63
0.83
0.62
0.75
0.75
0.80
0.75
0.70
0.65
0.60
0.55


Approval Rates by Race & Gender
Race
Gender
0
1
2
3
4
5
0.57
0.52
0.69
0.64
0.63
0.82
0.63
0.59
0.76
0.55
0.51
0.5
0.62
0.52
0.67
0.58
0.51
0.61
0.80
0.75
0.70
0.65
0.60
0.55
0.50


Approval Rates by Race & Income Group
Race
Income_Group
0
1
2
3
4
5
0.67
0.57
0.46
0.77
0.62
0.57
0.77
0.59
0.54
0.68
0.51
0.47
0.84
0.47
0.54
0.68
0.53
0.47
0.80
0.75
0.70
0.65
0.60
0.55
0.50



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 and all the groups which have a positive bias will continue to have an unfair advantage.
- **What are the ethical or social consequences?**
  Biased loan models can unfairly deny access to credit for certain groups, deepening social and economic inequalities.
  They risk violating anti-discrimination laws and AI principles, leading to legal consequences.
  Such bias also destroys public trust and causes injustice.
- **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 combinations to detect specific biases. I would also try out more metrics available in the aif360 tool.
- **Any lessons learned on fairness or auditing?**

  Ive learnt the different parameters which need to considered for bias detection and have gotten very comfortable with fairness detection tools like IDM aif360.