AI Risk Report

Project Title

Tanay AI Bias Report

1. Problem Overview

- The task that was given was to create a model that can predict loan approvals based upon a large dataset that was given.
- This is an increasingly important real world problem since AI has become such a prominent feature in our rapidly changing world. We have to ensure that the data and models we use are fair across demographics because if not this would lead to biased advantages towards specific groups.
- The dataset that was provided was a table of loan approval data which contained many applicants and their data and whether or not they had been approved or denied. This data had a lot of sensitive attributes that are commonly biased against including race, gender, and citizenship status just to name a few.

2. Model Summary

- The model that I used was Random Forest Classifier. I felt this was the most suitable choice as for starters this was a classification problem so using a classification machine learning method was important. Due to the limited data set, Random Forest's method of combining predictions from multiple trees helped reduce overfitting and improved the accuracy compared to when I tried to use TensorFlow which consistently overfitted the data. Along with this the model required minimal preprocessing since I did not have to normalize or scale the data.
- For preprocessing I used pandas function .get_dummies() as this allowed me to convert all the categorical data into Binary numerical data. Then using hyperparameter tuning I was able to find that the hyperparameters listed below yielded the best results:

  n_estimators=500, criterion="entropy", max_depth=20, min_samples_leaf=2, min_samples_split=10
- After setting all these hyperparameters and testing on the internal data this is the scores produced:
- **Accuracy:** 0.6370

- **Precision:** 0.6166
- **Recall:** 0.4427
- **F1 Score:** 0.5154

3. Bias Detection Process

- The methods that I used were SHAP interpretation and Failearn audits to calculate False Positive/Negative Rates.
- Both the raw data and the model output was audited. The raw data was checked to see how the data was already skewed against certain demographics. The model data was then audited and it was seen how this inputted skewed data resulted in a model that had a skewed response.
- The audits were formed both at the group level and the individual level. For the group level Fairlearn Audits and a SHAP beeswarm plot were used and for the individual level a SHAP waterfall plot was used.

4. 📉 Identified Bias Patterns

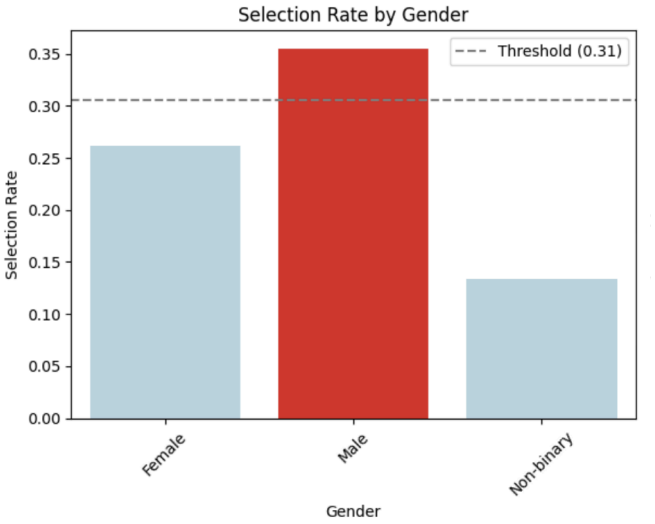
Summarize the biases your model exhibited. Aim for at least 4 clear findings. Include charts or tables if needed.

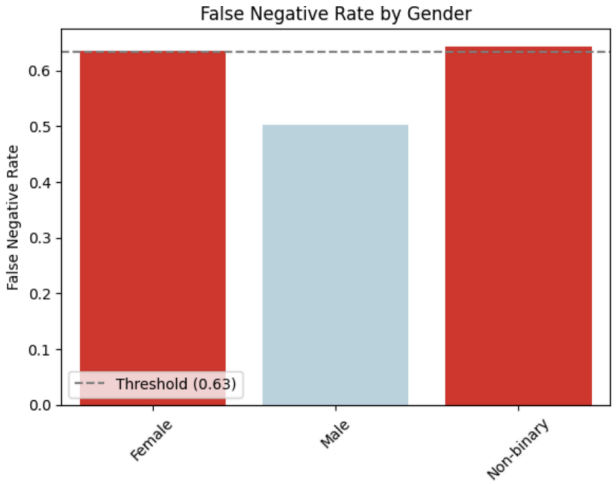
Bias Type	Affected Group	Evidence	Metric	Comment
Negative	Women	Lower loan approval rates	Selection Rate & False Negative Rate	Women saw a nearly 10% less selection rate for loans and a 10% higher false negative rate meaning they were incorrectly denied more often.
Negative	Black & Hispanic	Feature importance	SHAP plot	Black and Hispanic see negative impact value of SHAP model indicating that being of these races lowers chances of approval
Negative	Visa Holders	More denials	False Negative Rate	Visa Holders see a nearly 25% higher false negative rate meaning more incorrect denials than people who are citizens.

Positive	White	More acceptances	Selection Rate	White people saw a nearly 15% higher selection rate than any other demographic indicating that the model approves them more than anyone else.
Negative	Historically Redlined	Feature Importance	SHAP Beeswarm plot	People living in historically redlined neighborhoods saw were denied more often as the model saw this as an increased reason to deny loan.

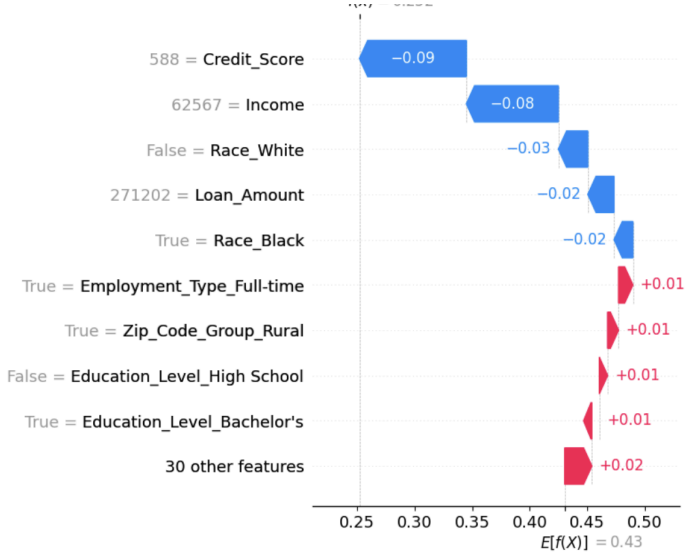
5. Visual Evidence

1.

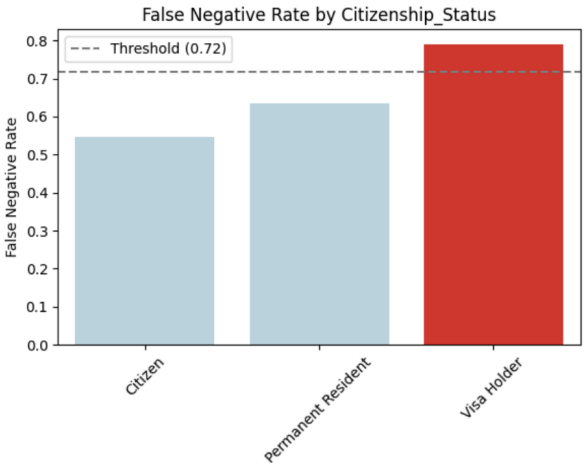




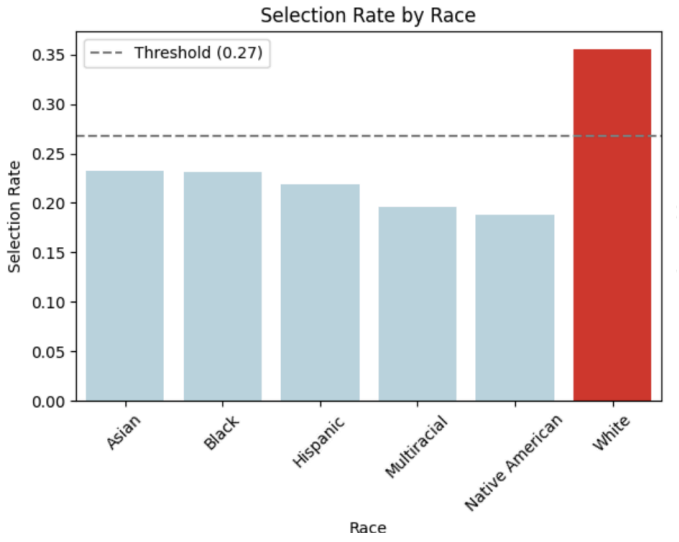
2.



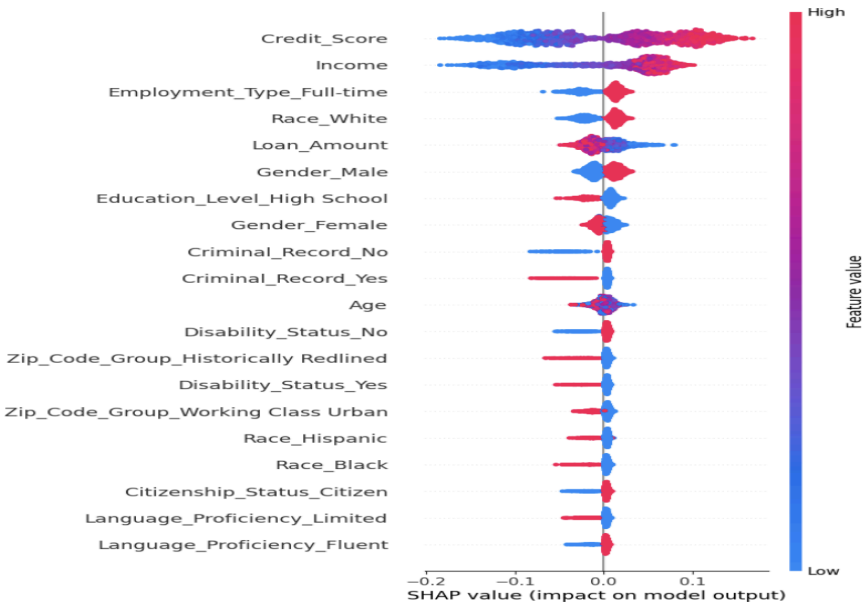
3.


False Negative Rate by Citizenship_Status
Threshold (0.72)
False Negative Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Citizen
Permanent Resident
Visa Holder

4.


Selection Rate by Race
Threshold (0.27)
Selection Rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Asian
Black
Hispanic
Multiracial
Native American
White
Race

5.


Credit_Score
Income
Employment_Type_Full-time
Race_White
Loan_Amount
Gender_Male
Education_Level_High School
Gender_Female
Criminal_Record_No
Criminal_Record_Yes
Age
Disability_Status_No
Zip_Code_Group_Historically Redlined
Disability_Status_Yes
Zip_Code_Group_Working Class Urban
Race_Hispanic
Race_Black
Citizenship_Status_Citizen
Language_Proficiency_Limited
Language_Proficiency_Fluent
−0.2
−0.1
0.0
0.1
SHAP value (impact on model output)
High
Feature value
Low

6. Real-World Implications

- The people who would be at most risk would be low-income, low-education, immigrants, women, or people of color. These are the groups that saw the most false-negatives and least false positives in the model. This means that they were denied incorrectly more and less likely to be incorrectly approved.
- The ethical consequences of this is that it leads to limited opportunities for these groups of people and demographics. If these certain groups are unfairly denied loans it can inhibit their ability to grow financially leading to them permanently being stuck in their current situation.
- This model would not likely pass a fairness audit. This is because through various methods like SHAP and false positive/negative comparisons we can see there is significant bias in the model. This would likely limit the usage of this model.

7. Limitations & Reflections

- One thing that did not work for me was getting the model much more accurate. I am unsure of what I did wrong but I was unable to get the model to have an accuracy higher than .64-.65.
- With more data and tieI would maybe try using TensorFlow. I tried using it but I ran into issues with the model overfitting. With it being my first time using tensor flow I decided to pivot away from that approach as it seemed very complex, but with more time I might retry that as it seemed promising initially.
- With this being my first time having worked on machine learning this project has taught me a lot about fairness in the world of artificial intelligence. It is incredibly important to have fair data to train models on otherwise it can lead to biases in the model very easily. Especially if newer models are trained on the decisions of other models it can lead to rampant inbred growth of bias. It is also incredibly important to constantly audit models to ensure fairness and continue equality.