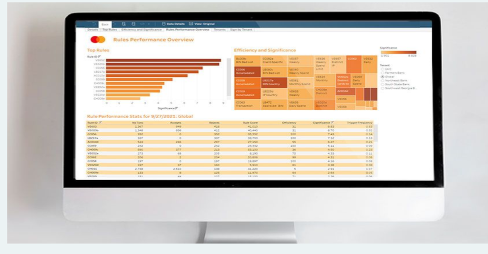
Enhancing Fairness in AI-driven Credit Approval Systems

UNIVERSITY OF TORONTO

Yuetong Jiang
Yiwen Ma
Chenyang Huan

Leveraging a Neural Network model, we exam the Australian credit card approval dataset for potential biases, employ Reweighting model for debiasing on Marital Status, and subsequently employ Permutation Feature Importance analysis to highlight the enhancements in fairness and decision-making accuracy.

I - Motivation



- The rise of AI-powered credit decision systems has revolutionized the risk assessment processes, optimizing the credit decision-making landscape.
- Still, there are concerns regarding the fairness, equity, and unbiased nature of these decision-making processes have surfaced, leading to a growing demand for clarity on their underlying logic.
- Provides a pivotal opportunity to ensure ethical compliance and societal responsibility, reinforcing the trust and confidence of consumers and stakeholders in digital finance platforms.

II - Dataset

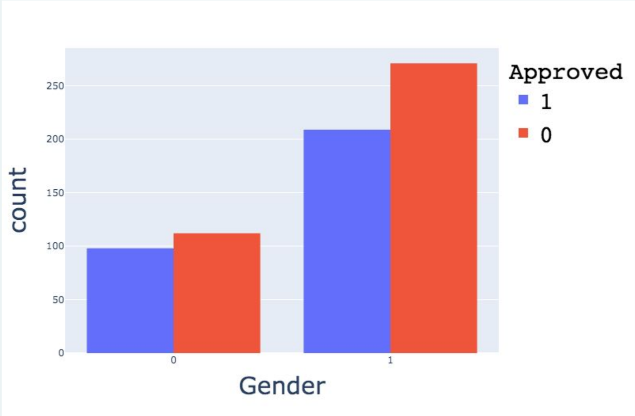
```
Data columns (total 16 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Gender          690 non-null    int64
 1   Age             690 non-null    float64
 2   Debt            690 non-null    float64
 3   Married         690 non-null    int64
 4   BankCustomer    690 non-null    int64
 5   Industry        690 non-null    object
 6   Ethnicity       690 non-null    object
 7   YearsEmployed   690 non-null    float64
 8   PriorDefault    690 non-null    int64
 9   Employed        690 non-null    int64
 10  CreditScore     690 non-null    int64
 11  DriversLicense  690 non-null    int64
 12  Citizen         690 non-null    object
 13  ZipCode         690 non-null    int64
 14  Income          690 non-null    int64
 15  Approved        690 non-null    int64
```

- **Binary Indicators**: Gender, Marital Status, Bank Customer Status, Employment, Prior Defaults, Driver's License Possession.
- **Numerical Values**: Age, Debt, Years Employed, Credit Score, Zip Code, Income.
- **Categorical Data**: Industry Sector, Ethnicity, Citizenship Status.
- **Target Variable**: 'Approved' (1 for approval, 0 for non-approval).

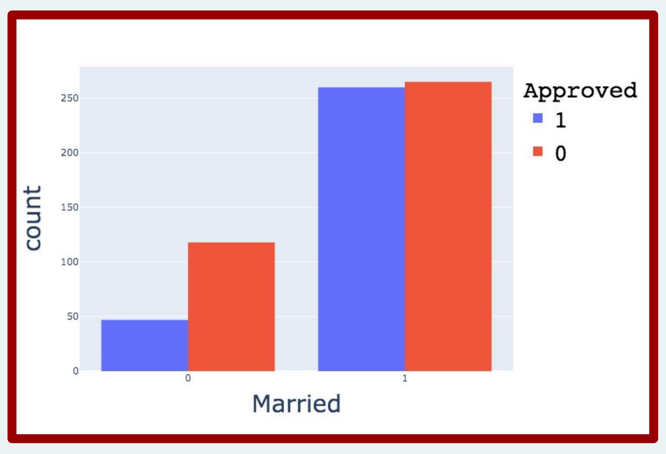
III - EDA

Distribution of favourable outcomes (Approval) among privilege and unprivileged groups:

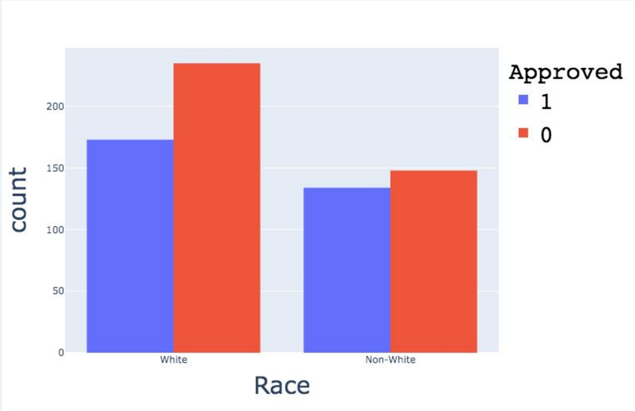
Gender



Marital Status



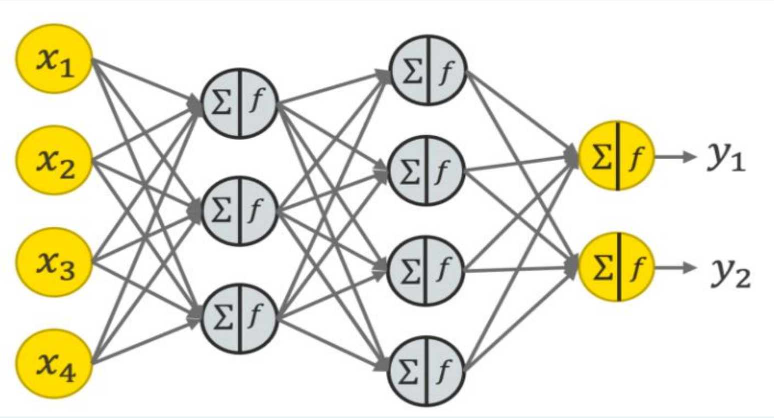
Race



It was observed that a most significant level of bias existed between married and unmarried individuals, with the unmarried group being more likely to face rejection for credit approval, than the other sensitive features.

IV - Neural Network

- Input Layer: 15 features
- Number of Hidden layers: 3
- Output Layer: Binary Label (Approved)
- Activate Function: 'logistic'
- Max-iter:10000
- Solver: adam
- Alpha: 0.01



V - Bias Detection

Statistical Parity Difference
Indicates the difference in probability of members from unprivileged group and privileged group being assigned the favorable label, Ideal: 0, indicating equal probability.

Disparate Impact
Estimates unintentional bias in a label assignment task which occurs when a group is assigned widely different outcomes to a protected class.
Ideal: 1, indicating a disadvantage for the unprivileged group.

Training Set Comparison

Metrics	Before Debiasing	After Debiasing
Statistical Parity Difference	-0.2326	0
Disparate Impact	0.5136	1
Accuracy	0.79	0.79

Testing Set Comparison

Metrics	Before Debiasing	After Debiasing
Statistical Parity Difference	-0.1955	0.09
Disparate Impact	0.646	1.19
Accuracy	0.83	0.77

After reweighting for variable 'Married', the metrics all reached ideal values on training set, and is much better on the testing set. The model accuracy is not much impacted.

VI - Permutation Feature Importance

Purpose
Aims to evaluate and compare the influence of various features on credit card approval predictions before and after applying a debiasing method.

Before & After

Results of Permutation Feature Importance

Feature	Before Debiasing	After Debiasing
Employed	0.087	0.034
Credit Score	0.077	0.101
age_group_> 40	0.029	0.010
Income	0.019	0.019
Years Employed	0.019	0.010
Married	**0.010**	**-0.014**
age_group_20-30	0.010	0.014
age_group_30-40	-0.019	-0.000
Citizen By Other Means	-0.019	-0.010
Debt	-0.010	0.010
Drivers License	-0.010	0.014
Race_White	-0.010	0.000
Gender	0.000	-0.014
Bank Customer	0.000	-0.014
Citizen Temporary	0.000	-0.005

PFI indicates the bias associated with "Married" has been successfully eliminated.