FAIRNESS PROJECT

Petteri Huvio, Luca Maahs

Problematic Datasets

- Bangladeshi University Students Mental Health
  - No correlation
- Medical Insurance Cost Prediction
  - No fairness issues according to **AIF360**

Chosen Dataset

- Realistic Loan Approval Dataset of US & Canada from Kaggle
- Total Records: 50.000
- Features: 20 (customer_id + 18 predictors + 1 target)
- Target Distribution: 55% Approved, 45% Rejected
- Missing Values: 0
- Binary Classification

Chosen Dataset

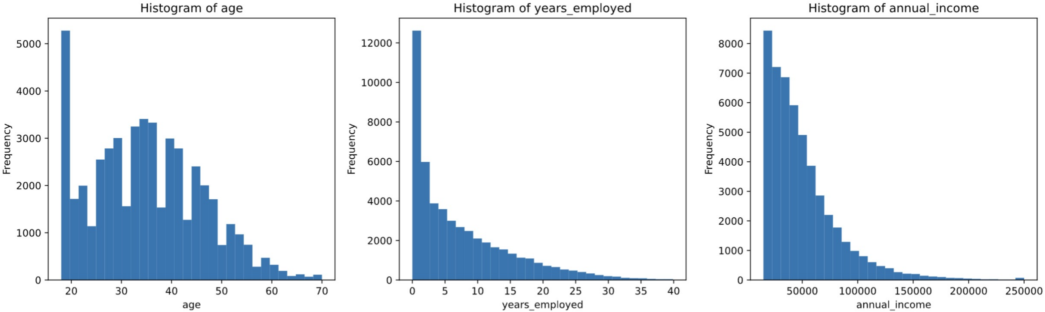

Figure: Priviledged Features

Trained Models

- Neural Network
  - Standard 3 Layer NN for Classification
  - Test Accuracy 90.69%
- Random Forest
  - Test Accuracy 91.26%

Fairness Issues

- Equal Opportunity
  - Some binary classification features, should not make a difference in whether you be granted a loan.
  - $TPR = \frac{TP}{TP+FN}$
  - $TPR_{priviledged} = P(Outcome = 1 \mid Qualified = 1, Group = A)$
    $TPR_{unpriviledged} = P(Outcome = 1 \mid Qualified = 1, Group = B)$
  - $\Delta = TPR_{priviledged} - TPR_{unpriviledged}$

Fairness Issues

- Studied Features
  - Age $> 40 \implies \Delta = 0.0638$
  - Top 20% years employed $\implies \Delta = 0.0416$
  - Top 20% yearly income $\implies \Delta = 0.0299$
  - Employement status $\implies \Delta = 0.0198$