FAIRNESS PROJECT

Petteri Huvio, Luca Maahs

Chosen Dataset

- Realistic Loan Approval Dataset of US & Canada from Kaggle
- Total Records: 50.000
- Binary Classification

Age Gridsearch

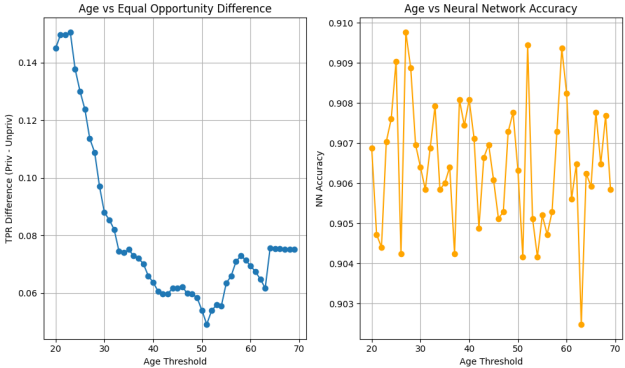

Figure: Age vs Equal Opportunity

Age Gridsearch

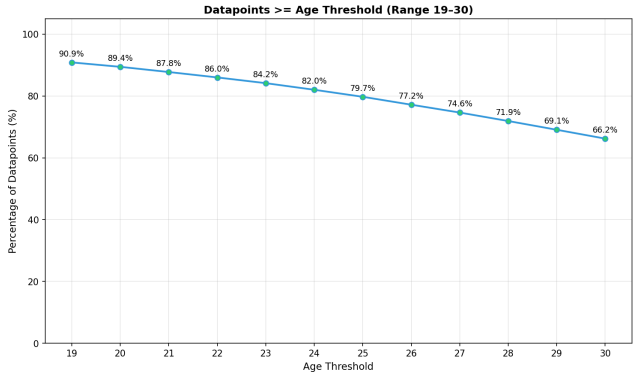

Figure: Age Threshold

Our Fairness Issue

- Equal Opportunity
  - Some binary classification features, should not make a difference in whether you be granted a loan.
  - Chosen Age threshold of 25 years
  - $\Delta = TPR_{priviledged} - TPR_{unpriviledged}$
  - $\delta = (1 - \frac{TPR_{unpriviledged}}{TPR_{priviledged}}) * 100$
- We chose now the feature Age with threshold X years
  - $\Delta = X$
  - $\delta = Y$

Baseline Models

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

Baseline Models

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

Baseline Models

- Neural Network:
  - Standard 3 Layer NN for Classification
  - Test Accuracy 90.69%
- Possible Fairness Regulation Options:
  - Pre-Processing
  - Post-Processing
  - Learning with Fairness-Constraints

Learning with Fairness-Constraints

- Idea was to penalize Loss-Function with $loss = loss + \Delta_{current} * \lambda$
- Where λ is a hyperparamter for the influence of Δ.
- For that we created our own EO-Loss-Function for.

Equal Opportunity Loss Function

```python
def EO_loss_fn(actual_loss, y_pred_probs, sensitive_attr, labels, lambda_coef=0.1, epsilon=1e-7):
    pos_mask = (labels == 1).squeeze()

    y_pred_pos = y_pred_probs[pos_mask]
    sens_attr_pos = sensitive_attr[pos_mask]

    priv_mask = (sens_attr_pos == 1)
    tpr_priv = (y_pred_pos[priv_mask].sum()) / (priv_mask.sum() + epsilon)
    unpriv_mask = (sens_attr_pos == 0)
    tpr_unpriv = (y_pred_pos[unpriv_mask].sum()) / (unpriv_mask.sum() + epsilon)
    eo_penalty = torch.abs(tpr_priv - tpr_unpriv)

    return actual_loss + (eo_penalty * lambda_coef)
```

Grid Search Results

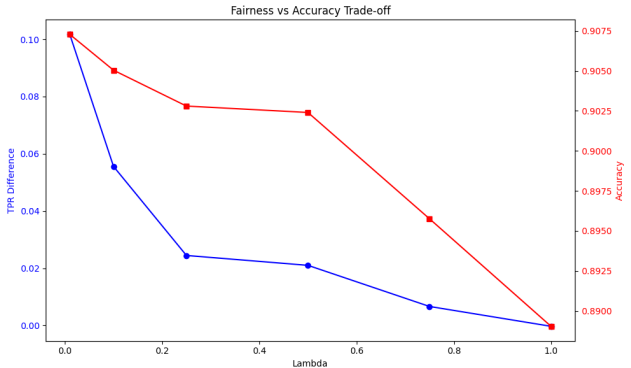

Figure: Fairness vs Accuracy Trade-off over λ

Grid Search Results

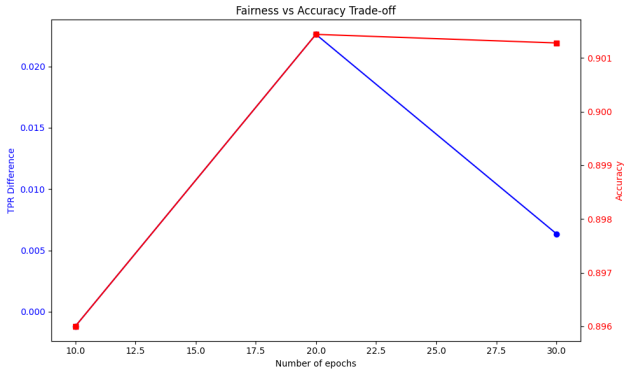

Figure: Fairness vs Accuracy Trade-off over epochs

Summary

	Unfair	Fair	Best $\lambda = 0.5$	Increased Epochs
Accuracy	90.69%	90.42%	90.24%	90.13%
Fairness (δ)	-	11.2%	1.2%	0.3%