Master's Thesis

Master's Programme in Computer Science

Trustworthy Machine Learning: Fairness Project

Petteri Huvio, Luca Maahs

December 9, 2025

Faculty of Science

University of Helsinki

Contact information

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki, Finland

Email address: info@cs.helsinki.fi
URL: http://www.cs.helsinki.fi/

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty Faculty of Science	Department of Computer Science

Tekijä — Författare — Author

Petteri Huvio, Luca Maahs

Työn nimi — Arbetets titel — Title

Trustworthy Machine Learning: Fairness Project

Ohjaajat — Handledare — Supervisors

I don't know yet.

Työn laji — Arbetets art — Level Master's Thesis	Aika — Datum — Month and year December 9, 2025	Sivumäärä — Sidoantal — Number of pages 10 pages

Tiivistelmä — Referat — Abstract

ACM Computing Classification System (CCS)

General and reference → Document types → Surveys and overviews

Networks → Network algorithms → Control path algorithms → Network design and planning algorithms

Avainsanat — Nyckelord — Keywords

Trustworthy Machine Learning, Fairness, Bias, Mitigation

Säilytyspaikka — Förvaringsställe — Where deposited

Helsinki University Library

Muita tietoja — övriga uppgifter — Additional information

Course on Trustworthy Machine Learning

Contents

1 Introduction

2 Methods

2.1 Data

(Patel, 2025) provides a dataset of loan applications from the US and Canada, which we use to evaluate fairness in machine learning models. The dataset includes features such as applicant income, credit score, and loan amount, along with a binary target variable indicating whether the loan was approved.

2.2 Base Model Training

We trained two base models being a Random Forest and a Neural Network.

2.2.1 Random Forest

How we trained it and what results.

2.2.2 Neural Network

How we trained it and what results.

2.3 Equal Opportunity

As a Fairness Metric, we chose Equal Opportunity because..

2.3.1 Chosen Subsets

2.3.2 Implementation

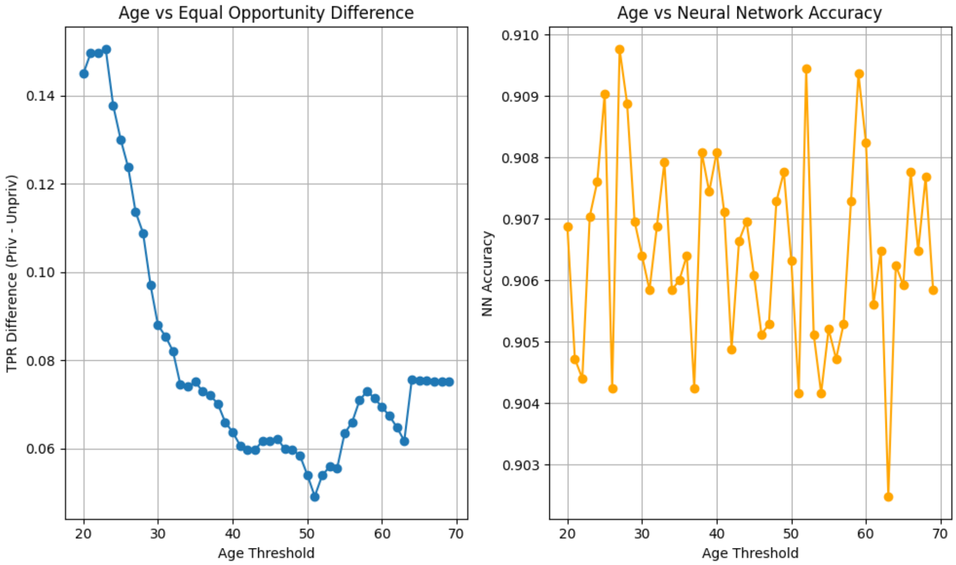

Figure 2.1: Fairness Results

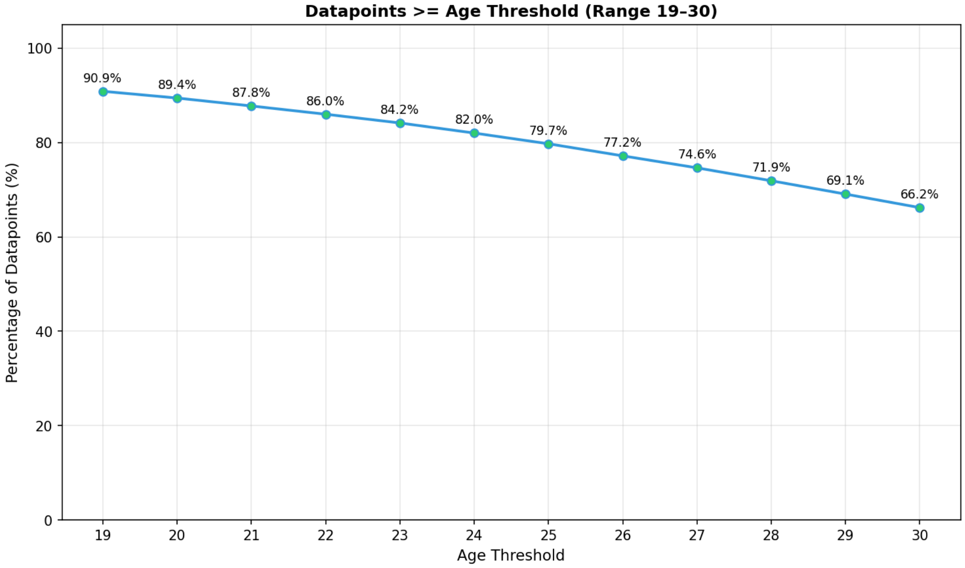

Figure 2.2: Age Threshold Distribution

```
 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)
```

Figure 2.3: Equal Opportunity Loss Function Implementation

3 Results

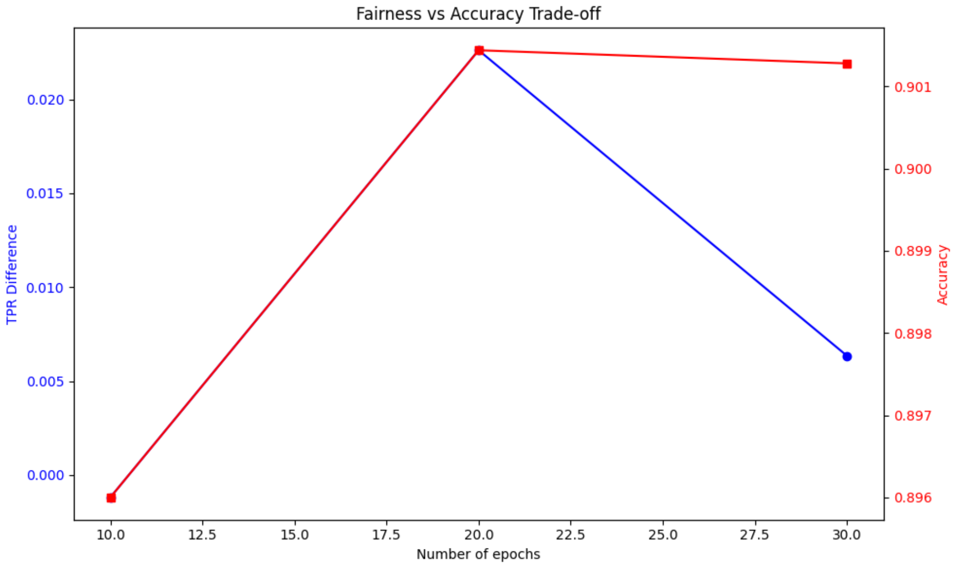

Figure 3.1: Fairness-Accuracy Tradeoff over Epochs

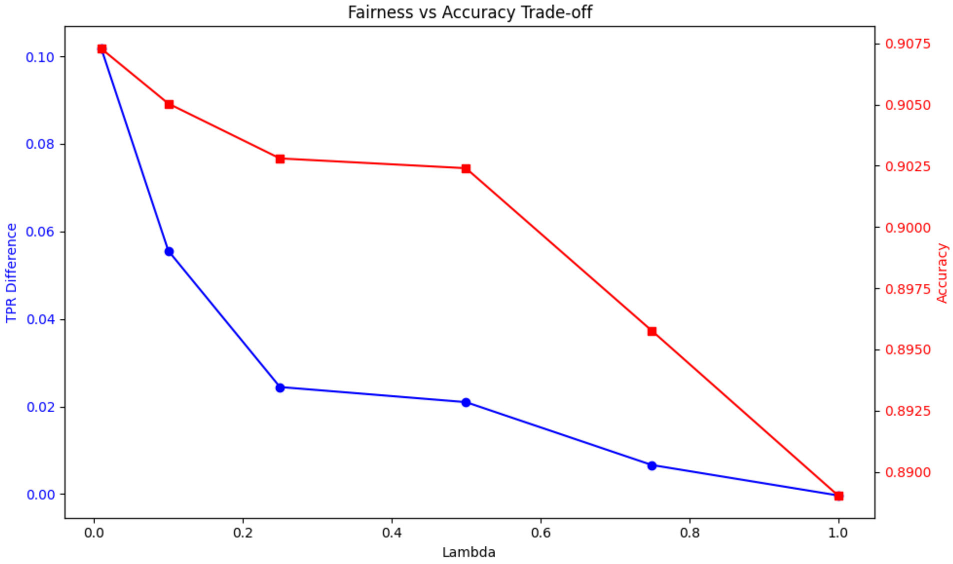

Figure 3.2: Fairness-Accuracy Tradeoff over Lambda

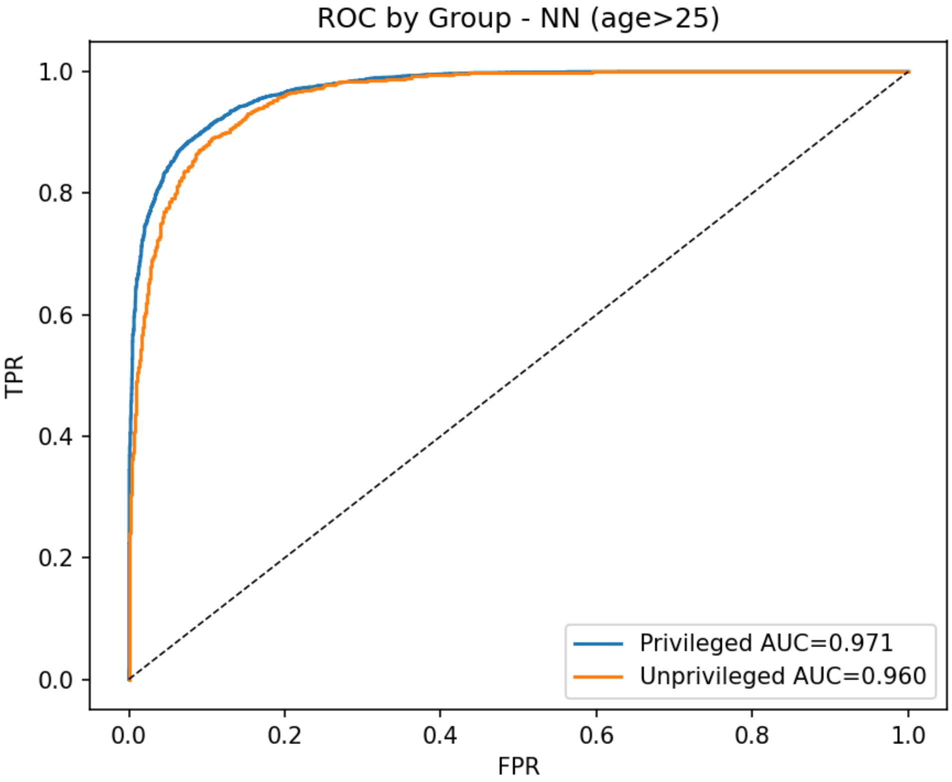

Figure 3.3: ROC by Group for Neural Network

	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%

Figure 3.4: Neural Network Results Summary

4 Discussion

5 Conclusions

Use of AI tools

NOT YET FILLED

Bibliography

Patel, P. (2025). *Realistic Loan Approval Dataset (US and Canada)*. https://www.kaggle.com/datasets/parthpatel2130/realistic-loan-approval-dataset-us-and-canada. Accessed: 2025-12-09.