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	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master’s Thesis	December 9, 2025	7 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 Project Idea

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

2.4 Implementation

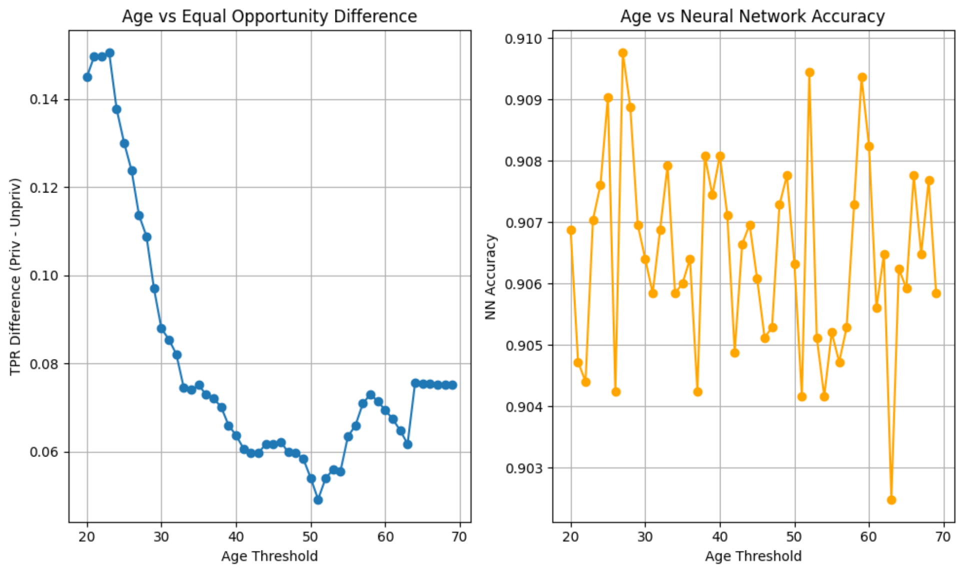

Figure 2.1: Fairness Results

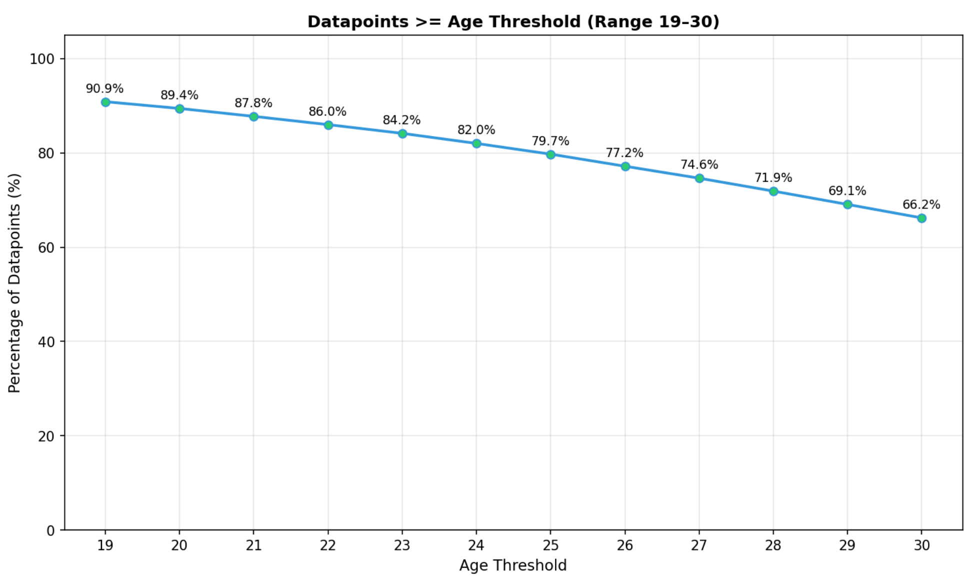

Figure 2.2: Age Threshold Distribution

```
 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

After having implemented our own Equal Opportunity loss function as described in Chapter 2, we started *Learning with Fairness Constraints*. We experimented with different hyperparameters and then decided that our λ coefficient and the number of training epochs e were the most interesting to analyze. For this we then set up two different Grid Searches, one for $\lambda \in \{0 \dots 1\}$ and one for $e \in \{10, 20, 30\}$.

The results from these these two experiments can be seen in Figure 3.1. As we can see, the with introducing a higher λ, and therefore fairness, we loose accuracy as expected. However, the fairness is increasing exponentially faster than the accuracy is decreasing, until a λ of about 0.5. After which the accuracy starts to drop faster than the fairness increases. This means that a λ of about 0.5 is a good trade-off between fairness and accuracy.

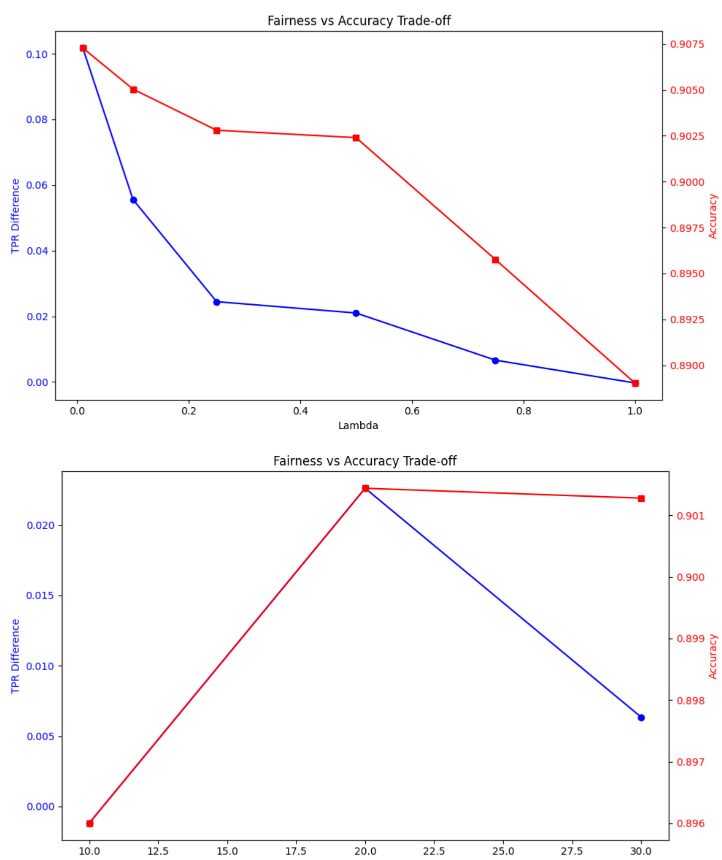

Figure 3.1: Fairness-Accuracy Tradeoff from Grid Searches

After running the experiment then also with the joined $\lambda = 0.5$ and $e = 30$ we had all our results to be summarized in Figure 3.2. Here we can see that the accuracy only drops slightly in each step of more fairness, while the relative fairness δ improves drastically from no fairness with $\delta = WRONG\%$ to 0.3% with $\lambda = 0.5$ and $e = 30$.

	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 3.2: Neural Network Results Summary

Finally after concluding our experiments were a success, we plotted the ROC curves by sensitive attribute groups for the Fair Neural Network in Figure 3.3. That has been done to be sure not to only rely on the fairness metric and accuracy, but also see that the smaller class is not being ignored by the model to accieve this success. As we can see the two ROC curves are quite close to each other, which indicates that the model is treating both groups fairly equally. This backes up the conclusion of our model now being much fairer than in the beginning, while only loosing a small amount of accuracy.

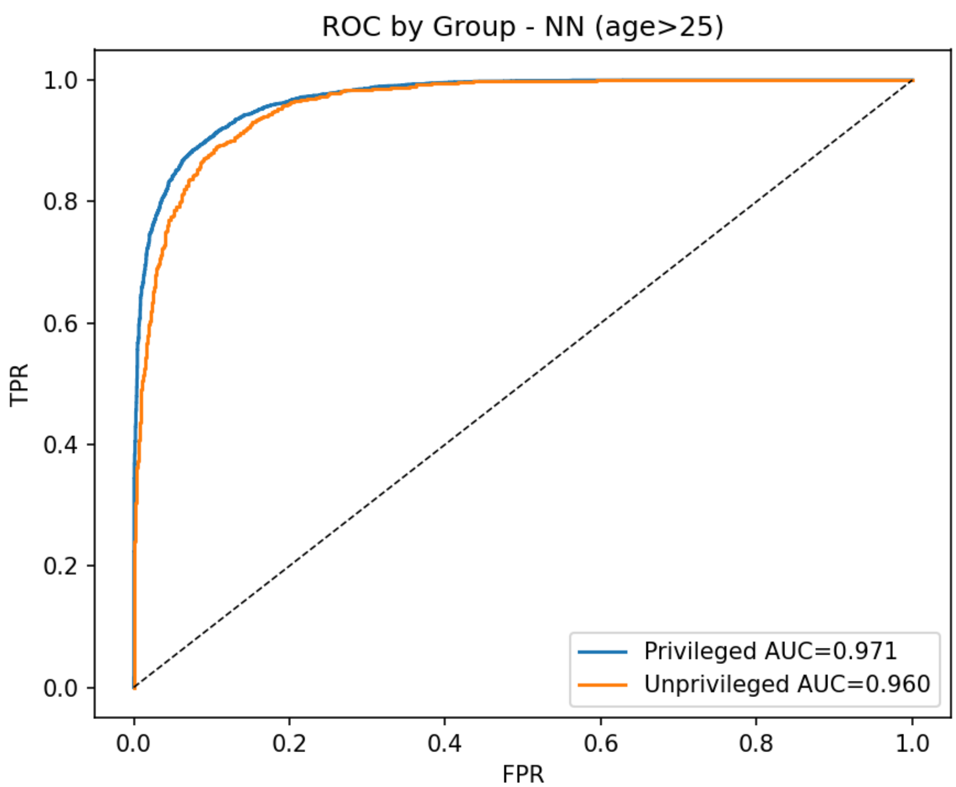

Figure 3.3: ROC by Group for Neural Network

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.