*Computers and Society Research Journal* (2022), **3**
https://doi.org/10.54822/BEWO3288

ARTICLE

# Practical bias correction in neural networks: a credit default prediction case study

Piet Snel and Sieuwert van Otterloo

Email: sieuwert@ictinstitute.nl

(First published online 11th May, 2022)

**Abstract**

Artificial intelligence (AI) is increasingly being used for decision-making. Technological developments have significantly increased the performance of AI models but have also increased their complexity. As a result, IT professionals are struggling to develop fair AI implementations as (1) measuring fairness in a practical case is difficult due to multiple definitions (2) literature on this topic is complex especially when multiple types of bias occur and (3) lack of practical cases in which corrections are made. Using a case study, we demonstrate how both gender and age bias can be addressed in practice. We do this by developing a credit default prediction model and detecting and mitigating both age and gender bias within this model. A neural network was trained using a real world credit data set from Taiwan. Existing 'bias' in the data set and bias introduced by this initial model was measured using a combination of previously published methods. A corrected model was created by training and evaluating a series of models to control bias along multiple dimensions. The final model eliminates the measured bias without sacrificing accuracy. It uses a top-down post-processing technique focusing on an equal increase of the default rate per group.

**Keywords:** Bias, Credit Default Prediction, Fairness, Taiwanese Credit Data, Fair AI

## 1. Introduction

Machine Learning (ML) is becoming an integral part of our daily lives. ML algorithms help us determine the route to drive to work, predict the weather and efficiently charge our phones. Recently, significant progress has been made in the performance of ML models and applications, driven by the development of new ML algorithms and theory, and the enormous growth in the availability of online data and low-cost computing [1]. As adoption increases, concerns increase about the transparency of these algorithms and if they cause or perpetuate bias against certain groups. For example, Amazon's hiring tool was found to disadvantage women [2]; facial recognition software can perform poorly for black women [3]; and the recent Dutch tax office "toeslagenaffaire" caused hardship to primarily people from immigrant communities. While ML models are getting better, their explainability is getting worse [4]. Because of the complex inner structure of ML, these algorithms are often referred to as "black-boxes". There is no comprehensive theoretical understanding of how neural networks arrive at their trained state or make predictions [5], particularly for deep neural networks [6; 7]. When these algorithms are used for decision making or screening, such as for loan-default prediction, risk management or applicant screening, these models

can have a significant impact on individuals. The black-box characteristics prevent the internal correction of the algorithms in the case of bias or other unjust decisions. As a result, several governmental bodies are working towards more strict regulation on AI. In April 2021, a draft proposal by the European Commission proposed a set of regulations on the use of AI [8]. Recent regulatory developments such as GDPR article 15 and 22, the US Algorithmic Accountability Act [9] and the OECD's AI principles [10] have also called for increased insight into ML interpretability and addressing bias.

Current research on bias fails to address the practical challenges that organisations using AI face: measuring bias along multiple axes; eliminating multiple types of bias simultaneously; and making sure overall performance does not deteriorate. Existing research is too complex and highly theoretical for practical applications, with many different methods and best practices. Research is not always focused on applying findings to real-world cases, meaning it is not practicable. Overall, in both the scientific and practical communities, there are few cases where a real-world dataset isused to test bias mitigation in a transparent and understandable manner.

In this paper, we use Yeh and Lien's (2009) [11] data set on Taiwanese credit card clients to demonstrate a algorithm-independent bias correct method to improve a machine learning prediction for the likelihood of default. We find that we cannot eliminate the bias in a single step, and must apply the bias-correction multiple times consecutively to mitigate the impact on different, overlapping groups. We use several different existing metrics to measure bias, to ensure that the algorithm is not tuned to one to the disadvantage of the others.

### *1.1 Definitions of bias*

Corbett-Davies and Goel (2018) [12] and Hutchinson and Mitchell (2019) [13] review of fair machine learning and the development of fairness principles. An essential consideration in both measuring and addressing bias is what is unfairness. For example, if one group has a higher chance of default, a model that predicts this reduces the chance of this group receiving a loan. As group, the model is fair, but at an individual level, this is unfair [14; 15]. Corbett-Davies and Goel [12] give three formal definitions of fairness:

1. *"anti-classification, meaning that protected attributes—like race, gender, and their proxies—are not explicitly used to make decisions;"*
2. *"classification parity, meaning that common measures of predictive performance (e.g., false positive and false negative rates) are equal across groups defined by the protected attributes;"*
3. *"calibration, meaning that conditional on risk estimates, outcomes are independent of protected attributes."*

However, Corbett-Davies and Goel conclude that these definitions are insufficient and taking such approaches could even harm the groups they intend to protect. They acknowledge how different groups can have different risks, which should be reflected in the algorithm.

Corbett-Davies and Goel argue, as do we, that not every group should have similar predictions, but that an AI algorithm should predict a similar risk for each group as the actual risk in the dataset. Applying this strategy would violate anti-classification and classification parity, but would eventually result in fairer machine learning.

Bias in AI algorithms can also result from unfair bias in underlying data sets [16]. This problem is out of the scope of our case study: we assume that the data set was compiled in a fair manner. For readers interested in this problem, we refer to the following examples: Buolamwini and Gebru [3] have shown such bias caused by both under or oversampling a specific group in the case of facial recognition software; while Zhang et al. (2020) investigated the interaction between decision making algorithms and a population can create a feedback loop, changing fairness assessments [17].

### *1.1.1 Defining fairness*

Fairness is an important characteristic of any algorithm. The EU general data protection regulation requires any data processing of personal data be fair. Unfortunately there is no uniform definition of fairness, despite a long history of scientific and legislative study [13]. Huang and Fu (2020) discuss distinct types of fairness [18], including:

- **Individual Fairness:** Requires decisions to be fair for any pair of individuals based on Dwork et al.'s fairness definition which states that similar individuals should be treated similarly [19; 14].
- **Group Fairness:** or the statistical fairness definition, requires equality in statistics across different demographic groups.

Most studies focus on either individual fairness (fair treatment between comparable individuals), or on group fairness, making the model outcome equal across groups [20; 15]. Naraynan states that the number of definitions is "non-exhaustive" making it impossible to decide upon a single definition and therefore approach [21].

The four-fifths rule originates from the US Equal Employment Coordinating Council (EEOCC) and is a quantitative guideline to determine whether rules have a disproportionate impact on any racial, ethnic or sex group [22]. They give the following definition based on the EEOCC procedures:

> *"A selection rate for any racial, ethnic, or sex group which is less than four-fifths (4/5) (or 80 percent) of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded as evidence of adverse impact…"*

An alternative view on this topic is based on checking if the model is using discriminative features such as gender, ethnicity, or nationality as a significant predictor. The increased usage of black-box models makes it inherently complex to determine if the features treat certain groups unequally. Guidotti et al. (2018)[23] survey multiple methods to explain Black Box models and make some interesting conclusions.

Achieving fairness in a real-world application will require trade-offs between fairness and other objectives. For example, Kleinberg et al. show that a model cannot integrate multiple fairness metrics simultaneously, except in highly constrained special cases [24]. They formalised three fairness conditions and showed how these cannot all be satisfied. Multiple other studies emphasise this need for a trade-off between fairness and the utility of a model [21; 25].

For this case study, we take a practical approach. We pose that any algorithm, in order to be fair, must have similar performance for all relevant subgroups, and show bias only against a subgroup if this bias is present in the data set. I.e. the algorithm is not allowed to introduce any additional disadvantage. We believe this promotes fairness in practice and is thus in line with ethics and regulations such as GDPR.

This still leaves the ethical question about how the results should be used. The bias in the prediction itself is removed, because the algorithm produces the same error for all groups. We only seek to address the first problem, and leave the second ethical discussion to other researchers [13].

## *1.2 Addressing bias*

There is no agreement in the literature on a single best practice or approach for mitigating bias. Setting different decision thresholds for different groups can be very effective in achieving a fair balance, especially when bias originates from training variables [26; 24; 13]. We can classify methods to address bias into three distinct approaches which are discussed by the article of Silberberg and Manyika [26]:

1. Pre-processing of data to reduce the relationship between model predictions and protected characteristics such as gender, age and ethnicity. This also includes transforming the data such that it does not contain any features that contain sensitive attributes such as gender, age and ethnicity.
2. Post-processing techniques that transform the models' predictions to reduce bias for certain groups. Hardt, Price and Srebro provide a framework for adjusting the models' outcomes to be fair amongst groups and include guidelines on when mitigating measures are required and when not to [27].
3. Imposing fairness constraints in the optimisation process or using an adverse to minimise the system ability to predict the sensitive attribute. This method is proposed by Zhang, Lemoin and Mitchell [28]. A problem with this method is that although it might be able to adjust for bias it makes the model harder to explain.

We chose the second approach, applying post-processing. Post-processing is explainable, transparent and flexible, as is this does not consider the actual model, which is a complex black-box, and merely focuses on the outcomes. This approach is therefore algorithm independent.

## 2. Methods

### *2.1 Data set*

We use Yeh and Lien's (2009) [11] data set of Taiwanese credit card clients. The data set was collected through a Taiwanese credit institution to compare data mining techniques for the predictive accuracy of probability of default of credit card clients [11]. It is a large data set of 30,000 samples and includes characteristics that could potentially lead to discrimination, such as age, gender and marital status. The data set contains 24 attributes as listed in table 1.

We use this data set to predict the probability of default. Default has costs to the lender, to the borrower, to other customers of the lender and potentially to wider society. Lenders lose money when clients default, and pass this cost onto their other customers and investors. The borrower can face potential hardship and may struggle to borrow in the future, meaning a default has a long-term impact. If financial institutions regularly provide credit to customers who default, it can affect the credit worthiness of the institution or even the region.

Several studies have been conducted on predicting default of credit card clients [29; 11; 30; 31; 32; 33; 34]. These studies provide insights on (1) the importance of data mining methods and AI in effectively predicting credit default, (2) methods and techniques on predicting credit default, and (3) the impact of these predictions. Recent studies on predicting credit default regularly use the same public data sets, which include data sets from Taiwan [11], Belgium [35], Germany [35] and Tunisia [36].

Fig. 1 shows the number of customers that have defaulted by gender. Women are over-represented in the data set, and are less likely to default than men. This immediately raises an important point on bias: it is often the case that characteristics such as age and gender are relevant for predictions. We assume, based on our understanding of best practices in finance and insurance, that this is an acceptable bias and our algorithm may predict a higher default probability for men.

As in many real world cases, there are multiple variables that have a potential for bias. Fig. 2 shows the predictions per age group, where group 30–40 has a relatively lower probability of default, group 40–50 has a higher probability of default. Age groups 60–70 and 70+ have too few samples to draw conclusions. The lack of samples is highly likely to cause bias in our model because the data are easily skewed by outliers. We therefore assume that the model would be fundamentally unfair for these age groups.

**Table 1.** Taiwanese Credit Card Data Characteristics. There are 30,000 unique entries in the data set.

| Attribute | Description | Min | Mean | Max |
|---|---|---|---|---|
| ID | Unique identifier for each customer | | | |
| LIMIT_BAL | Amount of given credit in NT dollars (includes individual and family/supplementary credit | 10000.00 | 167484.32 | 1000000.00 |
| SEX | Gender (1 = male; 2 = female) | 1 | 1.60 | 2 |
| EDUCATION | Education (1 = graduate school; 2 = university; 3= high school; 4 = others) | 1 | 1.85 | 4 |
| MARRIAGE | Marital status (1 = married; 2 = single; 3 = others) | 1 | 1.55 | 3 |
| AGE | Age in years | 21.00 | 35.48 | 79 |
| PAY_0 – PAY_6 | Repayment status from September 2005 to April 2005 | | | |
| BILL_AMT1 – BILL_AMT6 | Amount of bill statement from September 2005 to April 2005 | | | |
| PAY_AMT1 – PAY_AMT6 | Amount of previous payment from September 2005 to April 2005 | | | |
| default.payment.next.month | Default payment | | | |

### *2.2 Neural network model*

We created an initial machine learning model to predict the probability of default. The machine learning package sci-kit learn was chosen for this research for a variety of reasons. First, this research focuses on addressing bias in real-world decision-making algorithm, as these often use Scikitlearn it would result in representative results. Secondly, Scikitlearn provides a range of algorithms together with data preprocessing tools such as scalers. Using a single package for most steps of this research improves explainability. The trained model should be able to assess a customer's risk of default and classify them as high or low chance of default, this makes it a classification problem.

Although this research focuses on black box model such as neural networks it is important to first review if such a complex algorithm is necessary. When a more transparent algorithm achieves similar or even better result than these are evidently preferred. A critical consideration for the algorithm was mentioned in the article that collected and published the Taiwanese credit default data set; Yeh and Lien (2009) mentioned, predicting credit default probability can only be achieved by an ANN [11]. Other previous credit card default prediction studies also determined how neural networks were top classifiers for this type of data and this particular data set [32; 11; 29; 31]]. Although some alternative algorithms such as extreme gradient boosting, and random forest were also applied to the data these did not have better performance.

LIME's submodular pick function and SHAP's KernelExplainer were initially used to explain the global model. Based on the combined top predictors of the LIME and SHAP analysis theses are:

- PAY_1
- LIMIT_BAL
- PAY_AMT2 _2

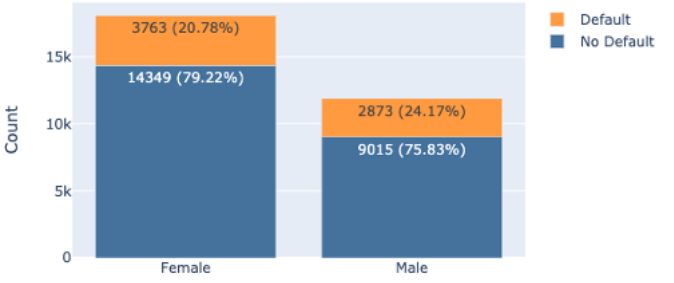


**Figure 1.** Customer Default by Gender

- PAY_4
- BILL_AMT2
- PAY_AMT5
- PAY_3

The values of these parameters were all normalised, such that each parameter has mean of zero and standard deviation 1 across the whole data set, prior to being presented to the network. The data were split random, with 20% of 30,000 samples set aside for the testing set.

We used a multi-layer perceptron network with 3 hidden layers of logistic functions, with architecture, with 12, 17 and 12 nodes respectively and the ADAM solver [37]. The learning rate was constant and we stopped the network after 10,000 iterations. This architecture was found to work best using Scikit-learn's GridSearchCV function, as is set up as below:

```
parameters = {'activation': 'logistic',
              'alpha': 0.0001,
              'hidden_layer_sizes': (12, 17, 12),
              'learning_rate': 'constant',
              'max_iter': 10000,
              'solver': 'adam'}

mlp = MLPClassifier(**parameters, random_state=0)
mlp.fit(X_train, y_train)
```

The data and code are available to download from https://ictinstitute.nl/bias-correction-credit-default/.

We use of multiple performance metrics in our subsequent analysis: accuracy, F1-score, precision, recall and false positive rate. Using only accuracy can hide unfairness due to differences in precision or recall.

## 3. Results

### *3.0.1 Adverse Impact Analysis*

We started with analysing bias based on age and gender separately. However this parallel approach leads to insights that are not easy to address simultaneously. A better approach is to consider age and gender simultaneously. Figure 2 shows the performance of our initial model for each subgroup, for five relevant metrics: accuracy, F1-score, precision, recall and false positive rate. The results are similar, but not exactly equal so there is an opportunity for improvement.

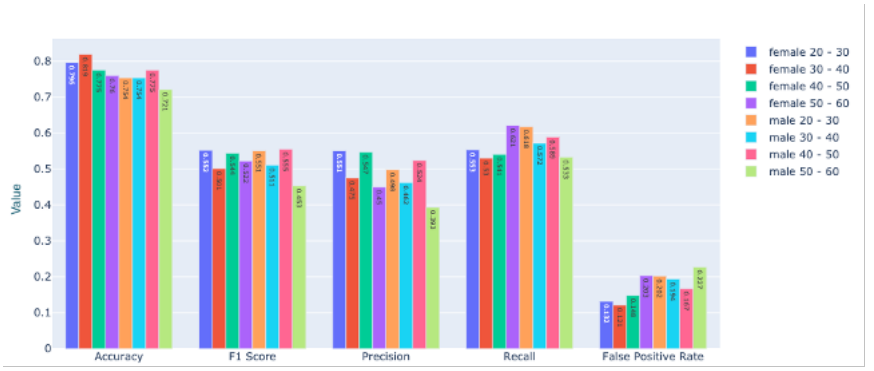


**Figure 2.** Performance by age group and gender initial model

We defined fairness as having similar performance across all metrics and values close to those of the data set statistics. In fig. 3, we use the network to predict the default rate for each single-sex age group. The universal rate of default in the data is 0.22, and the model predicts a somewhat higher universal rate of 0.25.

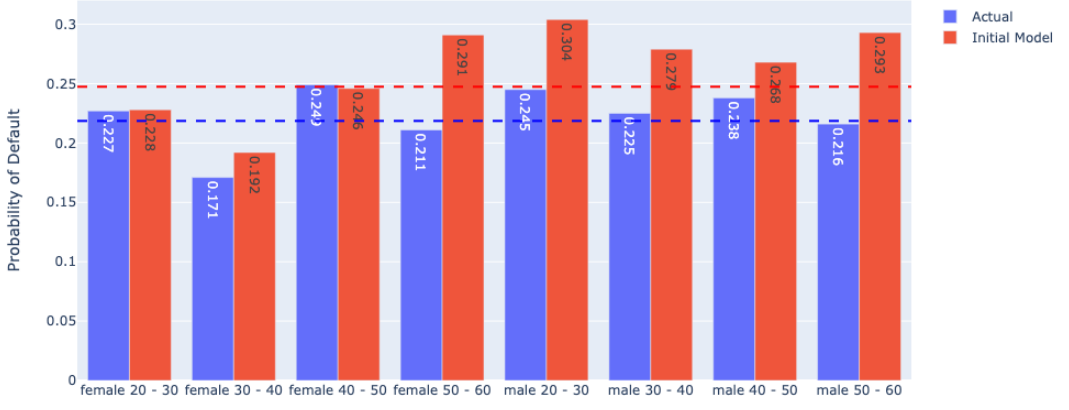


**Figure 3.** Initial model: actual and predicted probability of default age group and gender

As fig. 3 shows, there is significant bias against men aged 20–30 and 30–40, relative to the universal average. Negative bias for males is measurable amongst all age groups. Based on the analysis above we define two groups in which to mitigate bias.

- Males of all age (over-predicted default rate);
- Females aged 30–40 (under-predicted default rate).

### *3.1 Correcting gender bias by adjusting model for male subgroups*

To reduce bias, a new model was defined by adjusted the output of the neural network. The output was adjusted according to the following formula. This formula adjusts all predicted probabilities slightly and impacts all performance metrics. The formula was first applied to men as group.

$$\text{Target Probability} = \frac{\text{Mean Predicted Probability} \notin \text{Group}}{\text{Mean Actual Probability} \notin \text{Group}} \cdot \text{Mean Actual Probability} \in \text{Group} \tag{1}$$

To the best of our knowledge, this method has not previously been used for bias correction.

Fig. 4 shows how the second model corrected bias in all age groups.

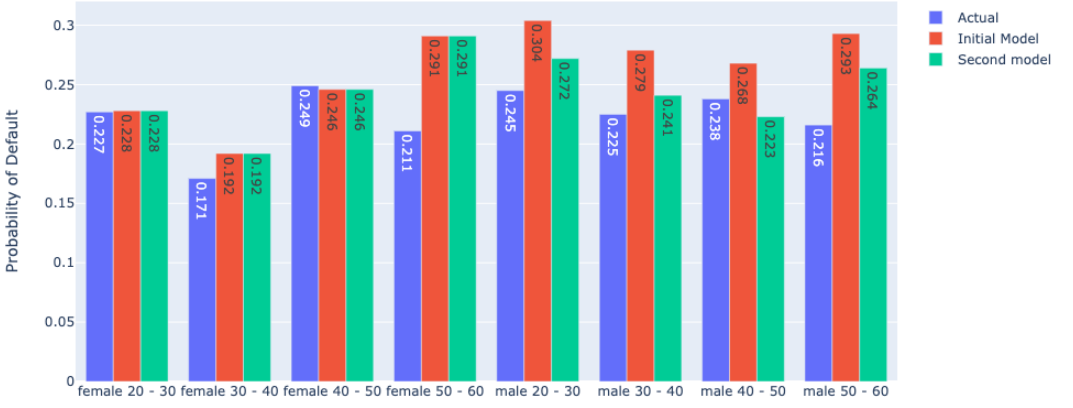


**Figure 4.** Actual, initial model and second corrected model probability of default by gender age groups

The second model is fairer that the initial model. Based on figure 4 we still notice two adversely impacted groups: female 50–60 and male 40–50. Additionally, one can make small compensations for male 50–60, female 40–50 and female 20–30. Although these last three groups show limited bias it is still be worthwhile to address, as the threshold correction function will only make limited adjustment when the adverse impact is small.

### *3.2 Correcting age bias by adjusting model for selected subgroups*

We then use equation 1 to reduce the bias for each age group. The data has already been corrected for sex-based bias. Fig. 5 shows the outcome of the third model which contains several corrections to some gender age groups. The bias mitigation method adjusted the increase in default from the initial model to the third model. While values ranged from –1.2% to a 37.9% increase in the first model this is reduced to a mere range of 6.7% to 11.7% for the improved third model.

Overall, each probability is now close to the actual probability in our dataset. A critical consideration to assess the effectiveness of this method is how it impacts model performance. Fig. 6 shows the performance of the initial model against the performance of the bias corrected third model. Surprisingly, the models performances are very similar. The corrections for bias that we introduced are sufficiently small to have no negative impact on algorithm performance.

## 4. Discussion and Conclusion

Organisations that use AI solutions to make predictions about individuals, such as credit default predictions, must address fairness. Despite the fact that a lot of research is available on ethical AI, this is challenging: the literature uses different definitions and does not describe how to solve practical challenges we encountered in a case example. We also believe that a criterium for acceptable bias is required, since we have shown that bias can be reduced but could not completely

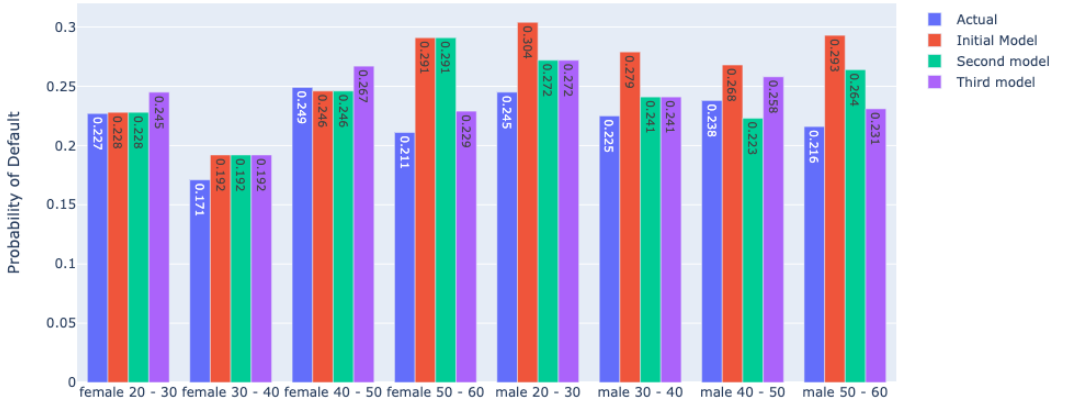


**Figure 5.** Actual, predicted and third model probability of default by gender age groups

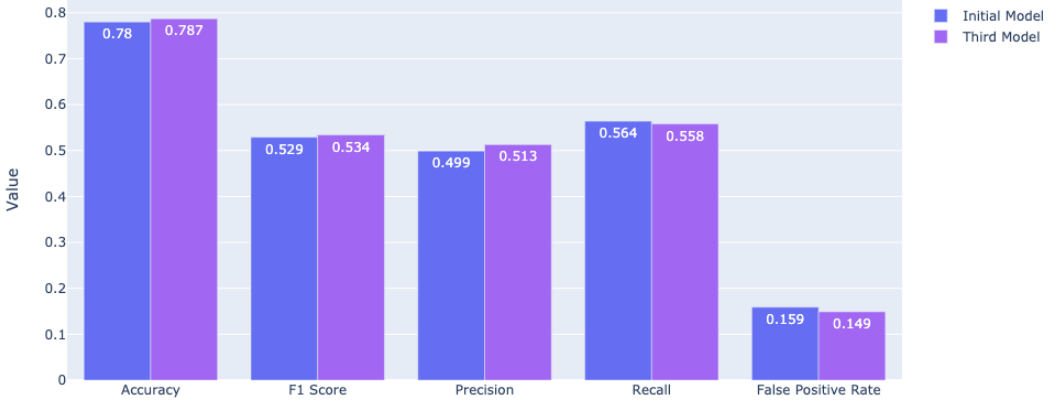


**Figure 6.** Initial model versus third model performance

eliminate it. Criteria are needed, perhaps different for different domains, whether a small amount of bias would be acceptable in real world applications.

Our conclusion is that resolving bias in real world AI solutions is feasible if professionals invest time in the detailed bias measurement and correction. For the example data set, unwanted bias can be eliminate without significant loss in performance. The main practical challenge is the time required for to measure and correct for bias.

In our example, we have found that it is possible to correct model prediction bias so that all groups have default probability predicted fairly. Our definition of fairness is that all groups have prediction rates inline with the data statistics.

However, our method required detailed assessment. We measured bias across multiple sub-group variables, and decided how to bin subgroups (e.g. age brackets). Several of these categories were sensitive (e.g. age, gender and marital status). Before using this method, we would need to consult domain experts about how best to address bias, so that we take into account cultural

factors, local laws and norms, and other ethical considerations.

We used a combination of performance metrics to define performance, in order to make sure all types of fairness can be addressed. We conjecture that our set of performance metrics (accuracy, fairness, precision, F1-score and false positive rate) is redundant, and that only using F1-score and precision or recall would also work. It would be interesting to validate what the most efficient definition of fairness is that works well on real world data sets.

We chose to mitigate the neural network model bias by applying a post-processing correction. There are other methods we could use to reduce the neural network bias. For example, we could pre-process the training and test input data by subgroup. However, this method is not algorithm independent. We could also tune the neural network, using extra bias nodes or a more complex error function. The disadvantage is that this method lacks transparency. A key feature of fairness is being able to describe how and why decisions have been made, and therefore training the network to combat unfairness defeats the objective.

An key advantage of our method is that it also corrects for model drift, where the underlying statistics change with time. In this case, an example of model drift could be caused by a particular generation moving from one age bracket to another.

## Acknowledgements

This work formed the basis of P. Snel's master's thesis during study at the VU Amsterdam under supervision of S. v. Otterloo. We would like to thank Yiannis Kanellopoulos and an anonymous reviewer for their feedback.

The code and data used in this study are available to download from the article page: https://ictinstitute.nl/bias-correction-credit-default/

## References

[1] M.I. Jordan and T.M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015, doi:10.1126/science.aaa8415.

[2] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 2018.

[3] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[4] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*, 1:506–519, 2019, doi:10.1038/s42256-019-0048-x.

[5] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv prpreprint*, 2017, doi:arXiv:1708.08296.

[6] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2016.

[7] Ravid Schwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv prpreprint*, 2017, doi:arXiv:1703.00810.

[8] Natalia Drozdiak. Notes from the AI frontier: Tackling bias in AI (and in humans). *Bloomberg*, 2021.

[9] 116th Congress (2019-2020). Algorithmic accountability act of 2019. *H.R.2231, 116th Congress*, 2019.

[10] OECD. OECD principles on AI, 2019.

[11] I. C. Yeh and C. H. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009, doi:10.1016/j.eswa.2007.12.020.

[12] Sam Corbett-Davies and Goel Sharad. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint*, 2018, doi:arXiv:1808.00023.

[13] Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58, 2019.

[14] Robert L Thorndike. Concepts of culture-fairness. *Journal of Educational Measurement*, 8(2):63–70, 1971.

[15] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness, 2016.

[16] Jake Silberg and James Manyika. EU set to ban surveillance, start fines under new AI rules. *McKinsey Global Insitute*, 2019.

[17] Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. Long-term impacts of fair machine learning. *ergonomics in design*, 28(3):7–11, 2020.

[18] Runshan Fu, Yan Huang, and Param Vir Singh. Artificial intelligence and algorithmic bias: Source, detection, mitigation, and implications. In *Pushing the Boundaries: Frontiers in Impactful OR/OM Research*, pages 39–63. INFORMS, 2020.

[19] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, pages 119–133. PMLR, 2018.

[20] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[21] Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, volume 1170, 2018.

[22] Irwin Greenberg. An analysis of the EEOCC 'four-fifths' rule. *Management Science*, 25(8):762–769, 1979, doi:10.1287/mnsc.25.8.762.

[23] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018, doi:10.1145/3236009.

[24] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv prepreprint*, 2016, doi:arXiv:1609.05807.

[25] Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pages 1389–1398, 2018.

[26] James Manyika and Brittany Silberg, Jake. Presten. What do we do about the biases in AI? *Harvard Business Review*, 2019.

[27] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016, doi:arXiv:1610.02413.

[28] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[29] Ahmad Ghodselahi and Ashkan Amirmadhi. Application of artificial intelligence techniques for credit risk evaluation. *International Journal of Modeling and Optimization*, 1(3), 2011.

[30] Sheikh Rabiul Islam, William Eberle, and Sheikh Khaled Ghafoor. Credit default mining using combined machine learning and heuristic approach. *arXiv prpreprint*, 2018, doi:arXiv:1807.01176.

[31] Shenghui Yang and Haomin Zhang. Comparison of several data mining methods in credit card default prediction. *Intelligent Information Management*, 10(5):115–122, 2018.

[32] Nikolaos Sariannidis, Stelios Papadakis, Alexandros Garefalakis, Christos Lemonakis, and Tsioptsia Kyriaki-Argyro. Default avoidance on credit card portfolios using accounting, demographical and exploratory factors: decision making based on machine learning (ml) techniques. *Annals of Operations Research*, 294(1):715–739, 2020, doi:10.1007/s10479-019-03188-0.

[33] Abbas Keramati and Niloofar Yousefi. A proposed classification of data mining techniques in credit scoring. In *The Proceeding of 2011 International Conference of Industrial Engineering and Operations Management, Kuala Lumpur, Malaysia, Jurnal*, pages 22–4, 2011.

[34] Sunil Kumar Vishwakarma, Akhtar Rasool, and Gaurav Hajela. Machine learning algorithms for prediction of credit card defaulters—a comparative study. In *Proceedings of International Conference on Sustainable Expert Systems*, pages 141–149. Springer, 2021.

[35] Talha Mahboob Alam, Kamran Shaukat, Ibrahim A Hameed, Suhuai Luo, Muhammad Umer Sarwar, Shakir Shabbir, Jiaming Li, and Matloob Khushi. An investigation of credit card default prediction in the imbalanced datasets. *IEEE Access*, 8:201173–201198, 2020, doi:10.1109/ACCESS.2020.3033784.

[36] Sihem Khemakhem, Fatma Ben Said, and Younes Boujelben. Credit risk assessment for unbalanced datasets based on data mining, artificial neural network and support vector machines. *Journal of Modelling in Management*, 13(4):932–951, 2018, doi:10.1108/JM2-01-2017-0002.

[37] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.