The Alan Turing Institute

Data Study Group Final Report: Accenture

16-20 April 2018

Fairness in algorithmic decision-making

https://doi.org/10.5281/zenodo.2557795

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1

Contents

1 Executive Summary

1.1 Challenge

We were tasked with evaluating fairness in its myriad forms, and mapping the various expressions of fairness to the data science workflow. Accenture challenged us to aggregate and organise the elements of the fairness literature into a manageable structure, and to provide meaningful visualisations that facilitate productive discussions around fairness in an analytical project. In this study, we focus on financial services and, in particular, on credit allowance in retail banking, where there is a prevalence of algorithms impacting customers and the services they are able to receive.

1.2 Objectives

Our overall aim was to create a workflow to facilitate reasoning about potential issues of fairness at various stages of the analytical process. Accenture needed a means of quantifying fairness, and improving upon any unfairness present in a process. Imposing fairness constraints often comes at the cost of model accuracy. It was important for the tool to include evaluations of the fairness-accuracy and fairness-cost trade-offs. Most importantly, since Accenture deals with a diversity of analytical projects, a flexible tool that could work with many types of analysis was required.

1.3 Approach

Fairness comes in many varieties and different considerations apply to different stages of the analytics process. Furthermore, it is not always clear what the right approach to achieving fairness is. These considerations are sometimes the remit of the data scientist, but more often also require input from experts and consultant in the relevant field. With this in mind, we decided that the most useful tool we could develop

would lay the groundwork for a structured means of considering fairness at the various stages of data analytics, from data exploration through to model finalisation.

Our approach over the five days began with a review of the literature on the various aspects of fairness. We assessed the ways in which the data scientist's work is impacted by fairness considerations and then prototyped a workflow that could begin to facilitate these requirements. We then honed in on particular aspects of quantitative fairness that could form part of a *lightweight* analytics tool. Finally, experiments were carried out and options for visualisations explored. We believe we have laid the groundwork for this tool by providing a means for reasoning about, and visualising metrics of, fairness for analysts and consultants who may not be aware of hidden prejudices in their data set or model. We have mapped some qualitative definitions of fairness onto their quantitative counterparts.

1.4 Limitations

- The German Credit Data Set is relatively small. It contained the decision taken on whether to grant the loan or not, but did not include an indicator on whether individuals who were granted a loan repaid it or not.
- Sometimes unfairness cannot be truly remedied algorithmically but requires human intervention and possibly a change in for example the data-gathering process. This is an inherent limitation within which our week's work had to operate.
- We have focused on fairness issues mainly at the stages of data preparation and model building. While we brought attention to the wider context in which these stages belong, we did not have enough time to adequately expand upon the other parts of the workflow.

1.5 Sections overview

In Section 2 we provide further details on the workflow concepts. Section 3 introduces the German Credit Data [1]. Concepts of quantitative fairness are defined and discussed in Section 4, and experiments based on these definitions are in Section 5. Suggestions for further work are in Section 6. We were a team of fourteen, all from different fields of research, and all deeply engaged in the issues of fairness that loom large in a society that's increasingly regulated by algorithms. In Section 7 a short description of each member is provided.

2 Mapping the analytical journey

In order to get from the initial proposition of an analytical project to the finalised results, a lot of data processing occurs and a lot of decisions are taken. There are many points in the process where fairness concerns are relevant, so that it can be challenging to remain cognizant of every potential issue. We propose a lightweight fairness assessment tool that integrates the various modules of the data science workflow.

Our aim is to establish fairness safeguards as a crucial and standard element of model evaluation, and, more generally, to raise awareness amongst data scientists, consultants, and others involved in the decision making process. We spoke to a retail banking expert who underlined the need for integrated guidelines on the qualitative and quantitative forms of fairness.

Figure 1 compartmentalises the data journey, highlighting some fairness considerations that can be accounted for at each stage. Olteanu et al. [2] provide a taxonomy of biases that can occur. Here, we step through the analytics workflow, drawing attention to a few examples of biases that can arise. We recommend a close reading of Olteanu et al. to understand the complete picture.

At the point of ideation: Is the group of people tasked with brainstorming relatively homogeneous? If so, they may not have an adequately diverse set of viewpoints between them. This is particularly troubling if the analysis

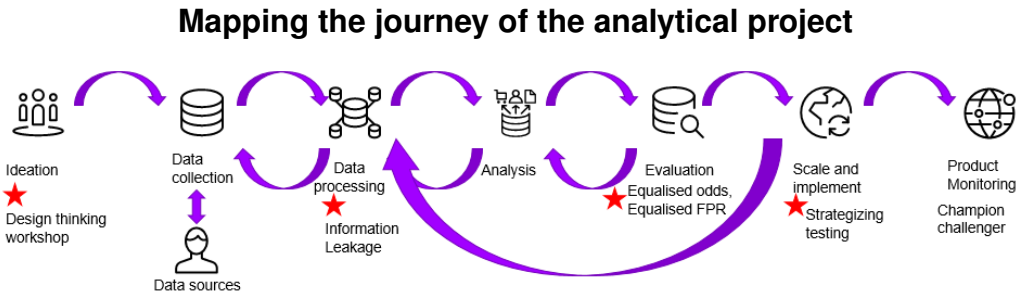


Figure 1: The stages of the analytical process.

concerns a population of people that include individuals very different to the brainstorming decision-makers.

Data collection: Is the collected sample representative of the population it is supposed to represent? If you're trying to build a model to predict whether a person will default on their loan, and the sample is made up of people who were granted a loan based on criteria such as credit score, then your model will not be predictive in the population at large, but merely in the subpopulation of people who had a credit score exceeding a certain threshold.

Data processing: There are many issues that can arise here. One example concerns missing values. If a choice is made to work with 'complete cases' only, then any observations with missing data will be excluded from the analysis. Yet this missingness could be systematic, so that a subpopulation is systematically underrepresented after the data cleaning. Data aggregation can also cause bias because reducing the granularity of the data may obscure crucial differences between subpopulations.

Analysis and evaluation: Depending on the model that is chosen to represent the data, very different results can be achieved. It is important to compare several models in terms of predictive performance in subgroups defined by protected characteristics such as gender. In Section 4 we highlight several methods evaluating and correcting for differences in model accuracy between these subgroups.

3 Data

The German Credit Data [1] was used to test and evaluate various fairness metrics. The data consists of 1000 loan applicants. For each observation, there are 20 variables, including demographic information (e.g. age, gender) and financial information (e.g. savings and credit history), and an decision outcome representing whether the applicant was classed as a good or bad credit risk. From now on we call it the credit data.

Very little processing of the data was required. The field 'personal status and gender' combined marriage status and gender in one. This was separated out into two separate variables.

A cost matrix was provided with this data set. A false positive cost five units, a false negative cost one unit, and there were no costs involved for a true positive or true negative.

In our experiments we suppose that gender and age are protected variables. Let $G \in (m, f)$ represent gender and $A \in [18, 65]$ represent the ages of the applicants, which fell within the range 18 to 65. Suppose a model has been built to predict whether an individual will be given a loan. The outcome $D \in (0, 1)$ represents the ground truth of whether the loan was granted ($d = 1$) or not, and the classification model outputs a predicted score for each individual, S, where a higher value indicates a stronger belief that the individual will be granted the loan. A classifier is created from the score by choosing a threshold, s^*, so that the predicted decision $\hat{D} = 1$ if $S > s^*$. The other predictors in this data set, $X = (X_1, ..., X_n)$, are not protected features, but *may* be associated with gender. Let p represent the prevalence of the decision outcome D. The prevalence is $p = 0.7$, meaning 70% of the applicants were granted a loan.

4 Concepts of Fairness in Analytics

There are many ways of enforcing fairness constraints in data analytics and not all of them can be achieved at once. Additionally, fairness often

involves a compromise on model accuracy. Thus, it is important to understand what the implications of each constraint are. In this section we go into greater detail about concepts of fairness that are most relevant to the realm of the data scientist's work: issues around data processing, modelling, and model evaluation. These concepts are illustrated with reference to the credit data. For simplicity, we suppose that gender is the only protected attribute when providing examples.

4.1 Disparate impact

The term disparate impact is used to describe the occurrence of unintended discrimination. Even if gender is excluded from a model, discrimination can still occur if there are variables associated with gender in the model. Examples are variables such as salary and profession, which have different distributions for each gender. It is also helpful to consider the less serious example of shoe size, because it is a good example of non-obvious associations. After a model is built (with protected variable G excluded from build), if

$$P(\hat{D} = 1|G = m) \neq P(\hat{D} = 1|G = f) \tag{1}$$

then a form of discrimination has occurred. Disparate impact can be rectified, either by enforcing demographic parity so that

$$P(\hat{D} = 1|G = m) = P(\hat{D} = 1|G = f), \tag{2}$$

or by requiring the predictions to be more alike, although not identical, by setting acceptable bounds on the ratio of disparity as proposed by Feldman et al., 2015 [3]:

$$\tau < P(\hat{D} = 1|G = m)/P(\hat{D} = 1|G = f) < 1/\tau \tag{3}$$

Feldman et al. suggest an 80% rule that sets a threshold $\tau = 0.8$.

4.2 PPV, TPR, and FPR

4.2.1 Predictive parity

If a model is well-calibrated, then

$$P(D = 1|S > s^*, G = m) = P(D = 1|S > s^*, G = f) \tag{4}$$

The COMPAS algorithm for predicting recidivism was well-calibrated so that it satisfied this equality [4]. However, this was thought to be an inadequate safeguard against discriminatory algorithms because it underpredicted recidivism for white defendants and overpredicted it for black defendants [5]. In other words, the true positive rate (TPR) and false positive rate (FPR) rates in the black and white subgroups were different and this was a source of unfairness.

4.2.2 Equal opportunity

Equal opportunity means that the TPR are equal across the protected groups:

$$P(\hat{D} = 1|G = m, D = 1) = P(\hat{D} = 1|G = f, D = 1), \tag{5}$$

where $\hat{D}$ is the predicted outcome. Thus, this is a requirement that the model is performing equally well in terms of TPR, but the proportions of women and men that are predicted to be loan-worthy won't necessarily be the same.

In our example, a true positive is a person who paid back their loan and for whom it was predicted that they would do so. In this context, a difference in TPR is unfair because it means that the rate at which the model predicts $\hat{D} = 1$ for the individuals *who were decidedly loan-worthy* is different between subgroups.

4.2.3 Equalized false positive rate

This is sometimes referred to as predictive equality. It is similar to equal opportunity, except it instead concerns the FPR:

$$P(\hat{D} = 1|G = m, D = 0) = P(\hat{D} = 1|G = f, D = 0). \tag{6}$$

In banking, a false positive can represent a loss of business. Corbett-Davies et al. [6] suggest it is more sensible to equalise FPR to achieve a measure of fairness in certain contexts. For example, in deciding whether or not to grant a loan to individuals we can only measure FPR, not FNR (nor TPR), in future model validations, because where the loan application was rejected we cannot know whether that individual would have returned the loan or not.

4.2.4 Equalized odds

If we enforce both the constraints of equal opportunity and equalized FPR we achieve equalized odds,

$$P(\hat{D} = 1|G = m, D = d) = P(\hat{D} = 1|G = f, D = d) \tag{7}$$

for $d \in (0, 1)$. Enforcing both constraints simultaneously has the potential to significantly detrimentally affect accuracy and should be implemented with caution. We suggest the various visualisations outlined in Section 11 that show the trade-off involved in any of the above-outlined fairness adjustments.

4.3 Counterfactual fairness

Kusner et al. [7] define counterfactual fairness in the context of decision making as "a decision that is fair towards an individual if it [is] the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group." The authors point out that enforcing equal opportunity (Section 4.2.2) at the modelling stage will not meaningfully protect equal opportunity if the individuals in the training data have not experienced equal opportunity in already.

The formulation of counterfactual fairness is similar to that of demographic parity (Section 4.1) but the constraint is more context dependent in that it also accounts for non-protected features $X \in (X_1, ..., X_n)$. A predictor is counterfactually fair if, for all $X = x$ and $G \in (m, f)$

$$P(\hat{D} = 1|G = m, X = x) = P(\hat{D} = 1|G = f, X = x). \tag{8}$$

This approach is required if the set of variables X is not a complete representation of the world. Kusner et al. propose a counterfactual fairness model that involves modelling the latent features that have a bearing on the data set $D = (G, X, Y)$. Let there be a latent variable U that represents all socioeconomic, demographic, etc variables that influence the life of each individual and the mechanism for deciding whether or not to grant them a loan. The authors draw on causal inference to extract the latent variable U and then build a predictive model. Bantilan [8] provides a python tutorial that implements the work of Kusner et al. It uses the German Credit Data [1].

Where the true causal model is unknown, Russell et al. [9] provide an extended definition of counterfactual fairness that holds in every plausible causal 'world'. Cihappa et al. [10] propose a definition of *path-specific* counterfactual fairness which states that "a decision is fair towards an individual if it coincides with the one that would have been taken in a counterfactual world in which the sensitive attribute *along the unfair pathways* were different."

Due to the complexity of implementing conterfactual fairness methods, they don't form part of a lightweight workflow, so they are not included experiments. Nevertheless, we wanted to draw attention to them because we find they have the potential to meaningfully redress unfairness in data science.

5 Experiments

For each concept of fairness outlined in Section 4 (apart from counterfactual fairness), we have implemented methods for quantifying and correcting for unfairness in data and the modelling process. This

includes evaluating the fairness-accuracy trade-off inherent in each process. We propose a platform for integrating the considerations of model accuracy and its cost implications with those of fairness. We show what this could look like for each type of fairness constraint.

5.1 Initial data exploration

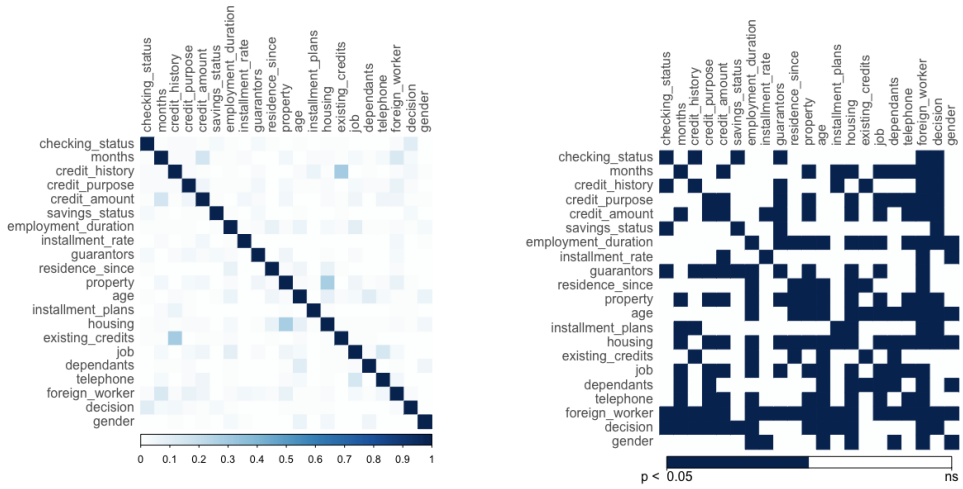
To get a better understanding of how the variables interrelate, bivariate analyses of each combination of variables was carried out. Mutual information was chosen as the measure of inter-variable dependence. Brown et al. [11] provide a short introduction to this metric. The mutual information between two variables, X_1 and X_2, tells us how much knowing the value of one variable, say X_1, informs us about the value the other, X_2, might take. It is calculated as

$$\text{MI}(X_1; X_2) = \sum_{\mathcal{X}_1, \mathcal{X}_2} p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)}, \tag{9}$$

where $\mathcal{X}_1$ and $\mathcal{X}_2$ each represent the values taken by the X variables. Note that if the variables are independent, $p(x_1, x_2) = p(x_1)p(x_2)$ so that the MI value will be zero. In reality, it will be close to, but not exactly, zero if the variables are independent.

The maximum possible value of MI is achieved if the two variables are completely dependent, so that knowing one tells you what value the other will take. However, MI is not just a function of dependence, but also depends of the number of different values taken by each variable and the marginal distribution of each, so that the maximum possible value varies between variable combinations. It is common to normalize MI when making comparisons between diverse variables. By dividing the MI by $\min(\text{entropy}(X_1), \text{entropy}(X_2))$ a normalized MI is achieved. It takes a minimum value of zero if the variables are independent, and a maximum value of one if they are fully dependent. It should be noted that there is more than one way of normalizing [12].

MI can only be applied to discrete variables. We discretised the continuous variables by binning each into a maximum of five bins with an equal number of observations in each bin. Figure 2 displays the



(a) NMI values

(b) Significance testing

Figure 2: Normalized mutual information for each combination of variables in the German Credit Data in plot (a). In plot (b), dark blue squares indicate a statistically significant relationship between two variables at the 0.05 level.

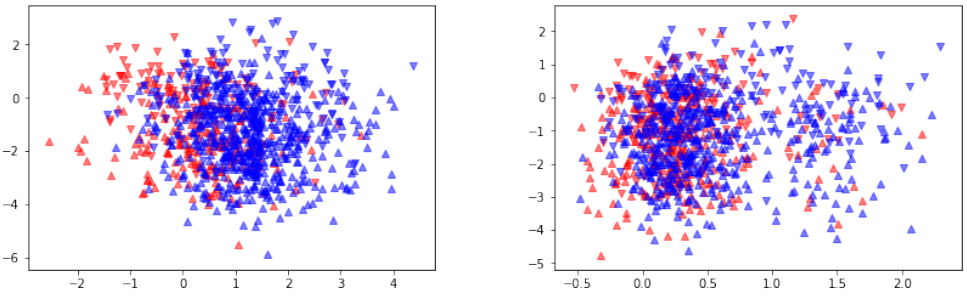
normalized MI values for each variable combinations. To test the significance of these values, calculate the G-statistic as $2 * N * MI(X_1; X_2)$. It follows the χ^2 distribution [13].

Further exploratory analysis could build on the MI work to look at the markov blanket around variables of interest, say, the protected variable gender, [14]. This would give an understanding of the variables that are most important for predicting gender.

5.2 Disparate impact

5.2.1 Evaluating disparate impact

Feldman et al. [3] propose a model based approach to identifying disparate impact by using the concept of balanced error rate (BER). If



(a) No correction (b) Corrected

Figure 3: The plots show the data set (a) before it is corrected for disparate impact, and (b) after the correction has been applied.

$X = (X_1, ..., X_n)$ are the non-sensitive predictors of the German Credit Data and G is gender, build a model, $f(X)$ to predict G from X. Then the BER is defined as:

$$\text{BER}(f(X), G) = \frac{P(f(X) = m|G = m) + P(f(X) = f|G = F)}{2} \tag{10}$$

Then G is said to be ϵ-predictable from X if $\text{BER}(f(X), G) \leq \epsilon$. And it is '$\epsilon$-fair' if the BER exceeds this threshold. See Feldman for a full description of the method. As this method is model dependent, it must be kept in mind that if the 'right' classifier is not applied, the BER will not be accurate. In this study, a hinge-loss SVM was applied, in line with the approach taken by Feldman et al. The paper [3] also uses the German Credit Data.

Where bias is not corrected for, or not completely corrected for, it is possible for this bias to become amplified in the model build process. Bias amplification identification was proposed by Zhao [15] to evaluate the change in disparity between two groups in terms of outcome. We did not focus on this method in the Data Study Group, but wish to highlight its existence. In summary, first calculate the maximum likelihood probability of default based on the observed outcome for each subgroup, $P(D = 1|G = m)$ and $P(D = 1|G = f)$. Then compare this to the classifications output from the model, $P(\hat{D} = 1|G = m)$ and $P(\hat{D} = 1|G = f)$. If there is a difference between the values, this suggests that bias amplification has occurred.

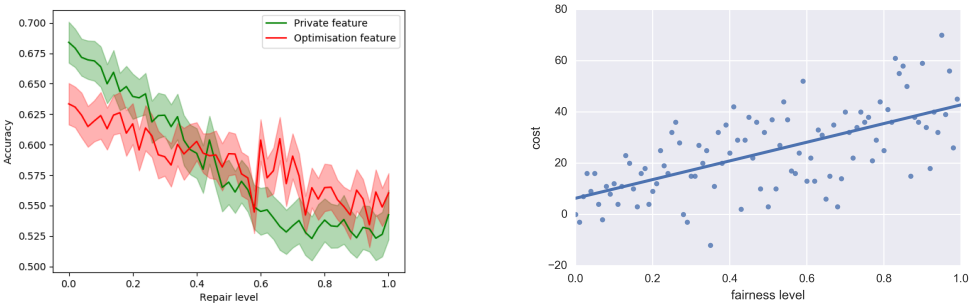
5.2.2 Correcting for disparate impact

Feldman et al. state that their method removes all information leakage that leads to disparate impact while preserving the rank. In our case, this is the rank of individuals in terms of credit-worthiness (we did not verify this in our experiments). They propose several approaches for partially 'repairing' the data so that the disparate impact is reduced. The *complete* removal of disparate impact is cautioned against because it can lead to a significant reduction in model accuracy.

Our experiments based on the credit data show the effect of repairing the data with respect to protected variable gender, Figure 3. This is a high-dimensional data set, so, to enable a meaningful visualisation of how the classification quality changes with adjustment for disparate impact, we projected the dataset into a two-dimensional embedding. The axes are chosen as the normal vectors of the separating planes of classifiers trained on reconstructing the protected attributes and predicting the credit rating, respectively. So the x-axis equals the decision value of a linear classifier for reconstructing the protected attribute, and the y-axis equals the decision value of a linear classifier for predicting the credit rating.

It is advisable to explore the fairness-accuracy trade-off of this method for various degrees of repair, ranging from none to complete, before finalising the extent of repair. Any compromise made on the accuracy of the model will impact a bank's risk profile. So, similarly, the fairness-cost trade-off should also be made clear. See Figure 4 for what this visualisation might look like.

If achieving the desired standard of fairness requires too great a compromise in terms of model accuracy, the data collection process should be scrutinised. A completely new data set could be the more appropriate solution.



(a) fairness-accuracy trade-off

(b) fairness-cost trade-off

Figure 4: Visualising the trade-offs that are involved in enforcing fairness constraints in disparate impact. Cost analysis is based on a cost of five units for a true positive and one unit for a false positive.

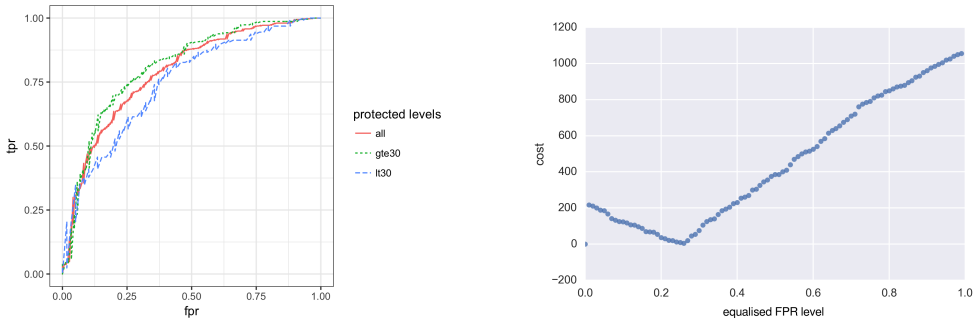
5.3 PPV, FPR, and TPR

5.3.1 Classification calibration

Several methods have been proposed ([16], [17]) to impose equal treatment amongst subgroups, say male/female gender, of a data set by adjusting the classification threshold on model output. These methods can easily be inserted into a workflow in so far as they are agnostic to the type of model that has been used to generate the output. The only requirement is that the model outputs a continuous prediction. In this study we focus on probability outputs, but it is straightforward to extend this approach to other continuous model outputs.

A disadvantage of this approach is that it can be naive. Consider the problem of information leakage described in Section 5.2. Correcting for differences with respect to a protected variable won't address the issue completely if predictor variables are related to it.

In this example we only explore the age variable. However, it is possible to explore subgroups based on more than one protected variable. For example, we could look at the four subgroups, women under 30, men under 30, women 30 or older, men 30 or older, too. Of course, the number of subgroups explored is limited by the subgroup sizes. We need



(a) ROC curve

(b) Cost analysis

Figure 5: (a) compares the ROC curves of subgroups before classification calibration is applied. (b) shows the fairness-cost trade-off for different levels of equalised FPR.

a reasonable number of observations in a subgroup in order to draw reliable conclusions from statistical analysis.

A logistic regression model with elastic net is applied to the credit data to obtain predicted probabilities of being given a loan. We then visualise any divergence in model performance between the categories of the protected variable using ROC curves. ROC curves permit a comparison of the TPR and FPR for every possible threshold on the model output $\hat{D}$. The ROC curve in the Figure 5 (a) shows that there is a difference between the two age groups, under 30 years and 30 or older, in terms of model accuracy. There are statistical tests for determining whether two ROC curves are different with statistical significance, eg. DeLong's test [18], which enables a rigorous statistical evaluation of ROC curve similarity (this is not provided for in the report).

It is possible to rethreshold the probabilities output by the predictive model in order to achieve equal FPR and/or TPR in both subgroups. However, Chouldechova [19] has shown that the TPR, FPR, and PPV cannot be equalised at the same time using classification calibration techniques:

$$\text{FPR} = \left(\frac{p}{1-p}\right)\left(\frac{1-\text{PPV}}{\text{PPV}}\right)(1-\text{FNR}) \tag{11}$$

Chouldechova notes that if the PPV is kept the same across subgroups

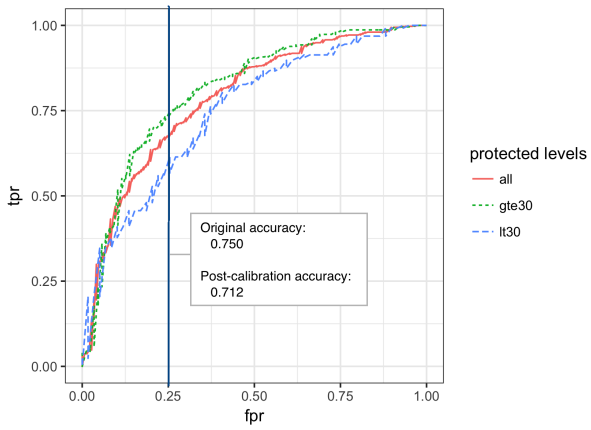


Figure 6: An interactive GUI for comparing the effects of different fairness constraints in classification calibration.

but the prevalence, p, differs between groups, equal FPR and TPR across subgroups cannot be achieved.

A cost analysis of the impact of different equalised FPR thresholds is provided in Figure 5 (b). Depending on what the primary fairness concerns are, an alternative plot can be generated to compare, say, the TPR-PPV trade-off instead of the TPR-FPR of Figure 5.

With more time, we would have made the ROC curve interactive, so that, for a given equalized TPR and/or FPR, the model accuracy, cost analysis, and PPV are recalculated. An example of what this might look like is provided in Figure 6. Here we visualise a vertical slider that can move across the plot. When moved to 0.25 on the FPR axis, fairness and accuracy metrics update to reflect the consequences of equalising the FPR at that threshold in all subgroups. Due to time constraints we did not implement this idea but it can be achieved with RShiny.

Apart from an interactive visualisation tool that the analyst can use at the point of deciding upon a classification calibration (Figure 6), we also suggest a final visualisation which shows the various fairness and accuracy metrics for an uncalibrated, traditional model and the 'fairer' model. See Figure 7 for an example of what this would look like.

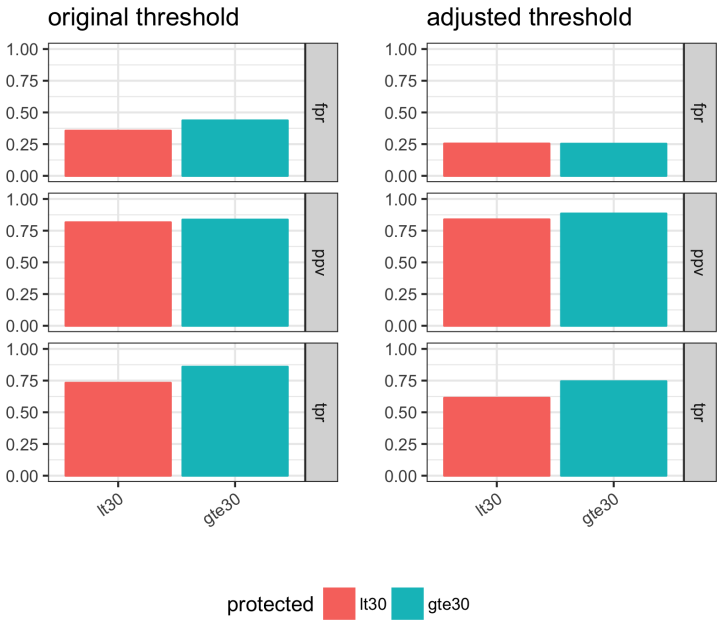


Figure 7: A comparison of the uncalibrated and calibrated, 'fairer' classification calibrations, split out by the protected variable: under 30 years, and 30 years or older. This visualisation could be enhanced with additional performance metrics.

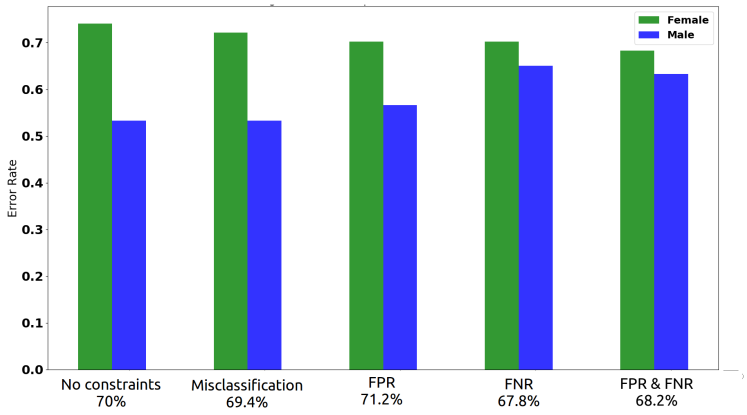


Figure 8: A comparison of overall accuracies (listed beneath each pair of bars) for the unconstrained approach and model calibrations which enforce parity amongst subgroups of misclassification rate, FPR, FNR, and FPR & FNR.

5.3.2 Model calibration

Control of the TPR and FPR can also take place at the stage of the modelling process in limited cases, as proposed by Zafar et al. [17]. Although it is not as convenient to implement as the classification calibrations of Section 5.3.1 are, we mention it because it can apply to logistic regression modelling, a common tool in the banking industry. Additionally, our experiments indicate that, for the same *fairness* constraints, higher model *accuracy* is achieved when we calibrate the model as opposed to the model output.

Model calibration per Zafar et al. is implemented in the loss function. They outline a way to maximise accuracy subject to fairness constraints, and a way to maximise fairness subject to accuracy constraints. Our experiments explored both of these options, see Figure 8. These constraints can be imposed strictly or to a specified extent.

A final visualisation equivalent to Figure 7 for classification is also recommended where model calibration is applied. We worked with code provided by Zafar et al., which we edited and applied to the credit data.

6 Future work

6.1 A fairer data collection strategy

It is important to develop a data set that is representative of the population. In the context of granting loans, this would mean giving a loan to a random sample of the population without first requiring them to meet requirements such as having a credit score above a certain threshold.

6.2 Concepts need further testing

Many of the ideas in this work are drawn from papers in the arXiv repository, which is not peer-reviewed. We advise that further research and testing be conducted around these concepts before their incorporation into processes that will affect people.

6.3 Sensitivity analysis

Friedler et al. [20] have pointed out that fairness-enhancing interventions can be sensitive to fluctuations in dataset composition, leading to instability in the predictions. Further work should involve evaluation of this sensitivity.

6.4 Disparate impact

There may be variables which it is undesirable to transform for the purpose of reducing data leakage, yet which should factor into the repair process. One example is an income variable, which is related to age and gender. It may be useful to banks if this variable could be used in the repair process, yet excluded from the transformation.

Additionally, it would be interesting to explore cases where the disparate impact repair algorithm had to flip the *outcome* label in order to reduce

bias in the data. This could provide insight into the types of decision contexts that are biased with respect to the protected variable. We suggest a nearest-neighbour search for the data points closest to the 'flipped' observation. This was cursorily applied to the credit data but yielded no useful insight.

6.5 Monitoring specific subgroups

Future work could involve quantifying probability of bias in a predefined set of subgroups. For example, customers that come from a group that is known to be underrepresented or discriminated against. These fairness focus groups could be the subject of particular scrutiny in the model build, and also at future model validations. An application belonging to an individual from this group could be flagged for review by a human decision-maker alongside the automated decision algorithm.

6.6 Catching up with *qualitative* fairness

We have focused on quantifying fairness and relating these concepts to qualitative notions of fairness. However, quantitative fairness is relatively new and has not yet caught up with "the jurisprudential, theoretical, and activist conversations which continue" [21] in this area. Quantitative fairness is limited in that is is based on data and data is often not neutral with respect to societal prejudices. Our work should be understood only as an entry point to considering the vast challenges of ensuring fairness in algorithmic decision-making.

6.7 Algorithms and accountability

We did not have time to explore issues around allocating responsibility for algorithmic discrimination. We point the interested reader to the work of Data&Society [22] for further information and resources.

7 Team members

- **Paul-Marie Carfantan:** Paul-Marie is a Postgraduate student in Data & Society at LSE. He investigates the transformative effects of personalisation algorithms, specifically user decisional autonomy, as well as the possibilities for understanding automated decision-making. He contributed to this report by providing meaningful theoretical background on fairness in algorithmic decision-making and conceptualising an integrated fairness tool.
- **Omar Costilla-Reyes:** Omar is currently a postdoctoral fellow at the Brain and Cognitive Sciences department at MIT. Omars research project aims to develop machine learning techniques to better understand neural data. Omar obtained his PhD with a thesis entitled spatio-temporal pattern recognition for security and healthcare from the University of Manchester, UK (2018) and his Msc from the University of North Texas, USA (2014). Omar contributed to the work by helping to map out the workflow, and by reviewing the counterfactual fairness literature.
- **Delia Fuhrmann:** Delia is a Postdoc at the MRC Cognition and Brain Sciences Unit at the University of Cambridge. She studies how brain and cognition develop over the lifespan. She helped scope the project and explore the data using mutual information.
- **Jonas Glesaaen:** Jonas is a Postdoc in the Physics Department of Swansea University working on lattice gauge theories. Besides working on code optimisation and HPC application he also dabbles in the potential of utilising ML techniques on problem in elementary particle physics. He contributed to this project through the disparate impact analysis. He assessed and visualised disparate impact on the dataset, and applied the fix as suggested in the literature.
- **Qi He (Katherine He):** Qi is a first year PhD candidate of UCL. She studies Bayesian variable selection and MCMC as well as implementing machine learning methods in Econometric Studies. She contributed to this report as the team facilitator. She helped the group work collectively and also worked on the disparate impact section.

- **Andreas Kirsch:** Andreas is a fellow at Newspeak House and will start a PhD in Oxford in September. He is engaging with AI safety and ethics as part of his fellowship. He contributed to the scoping of the project, and the experimentation and visualisation around disparate impact.
- **Julie Lee:** Julie is a PhD student in Neuroscience at UCL. She performs experiments (and, in the past, has built models) to understand perceptual and reward-based decisions in the brain. She contributed to this report by conceptualising a potential process for integrating fairness evaluations and solutions into an existing data processing pipeline.
- **Mohammad Malekzadeh:** Mohammad has started his PhD in Computer Science at Queen Mary University of London in January 2017 as a member of Centre for Intelligent Sensing. He is also a member of DataboxProject, an open-source project on privacy-aware personal data platform. His research focuses on machine learning for privacy-preserving data analytics, particularly for time-series data generated by the smart devices. He contributed to this report by implementing and evaluating existing methods of model calibration.
- **Esben Sørig:** Esben is a first year PhD student at Goldsmiths, University of London. His research is in Interactive Machine Learning, a field at the intersection of Machine Learning and Human Computer Interaction. He contributed to the project scoping and to the identification and evaluation of classification calibration metrics.
- **Caryn Tan:** Caryn is an Accenture Analytics strategist operating at the intersection of applied analytics and law/ethics. She advises senior decision-makers on analytics strategy, target operating model and analytics business case and manages technical teams to operationalise and realise these strategies. She also manages Accentures Responsible AI practice in the UK where she helps clients confidently deploy responsible AI models with technical, organisational, governance and brand considerations. She contributed to this report by bringing the context of the challenge and how Accenture works with clients on Responsible AI to the team.

- **Emily turner:** Emily is a Research Associate in the School of Law at the University of Manchester. She applies Machine Learning to understand recidivism in domestic violence. She was involved in the scoping of this project and helped collate the various methods already available for classification calibration. She worked on visualisations including subgroup-specific ROC curves, Figure 6.
- **Dang Quang Vinh:** Vinh is a Postdoc of Inria Nancy Grand-Est, France. His research interests include Applied Machine Learning in network security. He contributed to this report by studying different fairness measurements and repairing algorithms, and also evaluated methods experimentally.

References

[1] D. Dua and E. Karra Taniskidou, "UCI Machine Learning Repository."

[2] A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman, "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries," *SSRN Electronic Journal*, pp. 1–44, 2016.

[3] M. Feldman, J. Moeller, and C. Scheidegger, "Certifying and removing disparate impact ," *arXiv1412.3756*, pp. 1–28, 2015.

[4] W. Dieterich, C. Mendoza, and T. Brennan, "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity," 2016.

[5] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science Advances*, vol. 4, pp. 1–5, 2018.

[6] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," *arXiv1701.08230*, 2017.

[7] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, "Counterfactual Fairness," in *NIPS*, 2017.

[8] N. Bantilan, "De-biasing Classifiers with Themis-ml," *https://www.fatconference.org/static/tutorials/bantilan_themis18.html*, 2018.

[9] C. Russell, M. J. Kusner, J. R. Loftus, and R. Silva, “When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness,” *Nips*, no. Nips, pp. 6417–6426, 2017.

[10] S. Chiappa and T. P. S. Gillam, “Path-Specific Counterfactual Fairness,” *arXiv:1802.08139*, 2018.

[11] G. Brown, A. Pocock, M.-J. Zhao, and M. Lujan, “Conditional Likelihood Maximisation: A Unifying Framework for Mutual Information Feature Selection,” *Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.

[12] N. X. Vinh and J. Bailey, “Information Theoretic Measures for Clusterings Comparison: Variants , Properties , Normalization and Correction for Chance,” *JMLR*, vol. 11, pp. 2837–2854, 2010.

[13] B. Woolf, “The log likelihood ratio test (the G-test),” *Annals of Human Genetics*, vol. 21, no. 4, pp. 397–409, 1957.

[14] I. Tsamardinos and C. F. Aliferis, “Towards principled feature selection: relevancy, filters, and wrappers,” *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, vol. 4684, no. 10, pp. 986–992, 2003.

[15] J. Zhao, T. Wang, and M. Yatskar, “Men Also Like Shopping : Reducing Gender Bias Amplification using Corpus-level Constraints,” in *EMNLP*, 2017.

[16] M. Hardt, E. Price, and N. Srebro, “Equality of Opportunity in Supervised Learning,” *Computing Research Repository*, 2016.

[17] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, “Fairness Constraints : Mechanisms for Fair Classification,” *AISTATS*, vol. 54, 2017.

[18] E. R. Delong and N. Carolina, “Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves : A Nonparametric Approach,” *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[19] A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” pp. 1–17, 2017.

[20] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, “A comparative study of fairness-enhancing interventions in machine learning,” pp. 1–22, 2018.

[21] S. Mitchell and J. Shalden, “Mirror mirror: Reflections on quantitative fairness,” *https://speak-statistics-to-power.github.io/fairness/*, 2018.

[22] “Data & Society,” *https://datasociety.net*.

turing.ac.uk
@turinginst